Fast speech can sound slow

Effects of contextual speech rate on word recognition

Cover design: Leon Leeflang, "Ambiguity resolution"

ISBN: 978-94-92910-05-9

Printed and bound by Ipskamp Drukkers b.v.

Fast speech can sound slow

Effects of contextual speech rate

on word recognition

Proefschrift

ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,

volgens besluit van het college van decanen

in het openbaar te verdedigen op donderdag 12 december 2019

om 11.00 uur precies

door

Merel Maslowski

geboren op 3 maart 1989

te Enschede

Promotor:

Prof. dr. A.S. Meyer

Copromotor:

Dr. H.R. Bosker

Manuscriptcommissie:

Prof. dr. J.M. McQueen

Prof. dr. H. Quené (Universiteit Utrecht)

Dr. E. Reinisch (Ludwig-Maximilians-Universität München, Duitsland)

This research was supported by the Max Planck Society for the Advancement of Science, Munich, Germany.

Voor Ralf Maslowski

Contents

List of Figures

List of Tables

1 | General introduction

On May 11th, 2019, the Dutch National Unit (Landelijke Eenheid) posted on their twitter account that more than 2000 motor vehicles had been caught by a speed camera on the A2 motorway in the Netherlands that day. Because of road works, the speed limit had temporarily been changed from 130 km/h to 70 km/h, but some motorists had been driving faster than 100 km/h. Here, driving 100 km/h was clearly considered much too fast, but under normal conditions, 100 km/h on the same motorway would have been experienced as slow.

This example illustrates that something that is slow can be experienced as fast under different circumstances. Why does driving on a main road seem slow after coming from the motorway, but fast after having driven on a city road? What happens to one's perception of speed when driving on a bumpy road, where one continuously has to slow down and speed up, to slow down again? Which factors determine what feels fast and what feels slow? And what are the mechanisms responsible for this?

These are the type of questions the research presented in this doctoral thesis addresses with regard to perception of the rate of speech. In this thesis, the motorway is the metaphorical fast talker, the city road the slow talker, and the main road the talker with a medium speech rate. The bumpy road is the talker with a highly variable speech rate. This thesis measures the influence of coming from these different types of "roads" on speed perception when driving on the "main road", the medium speed. Just as one's perception of current driving speed is dependent on one's previous driving speed, listeners also perceive speech rate differently, depending on the speech rate just heard. That is, differences in the rates at which talkers speak can have strong effects on how listeners interpret individual speech sounds and words in the speech that follows. Understanding how listeners process speech rate in different contexts and use the speech rate cues derived from the speech signal is a crucial piece in the puzzle that is human cognition of language. In the remainder of this chapter, I first explain why speech rate is important for word recognition. Then, I briefly summarize different types of speech rate contexts, defined by their proximity to a stretch of speech

under study, and how these contexts influence speech perception. Third, I highlight open questions in the literature regarding speech rate processing. Finally, I present the research question of this thesis, as well as providing an outline of its chapters.

1.1 Perception of acoustic duration

A question one might ask, is why speech rate matters to listeners at all. In most cases, listeners seem to be able to understand their interlocutor, regardless of their interlocutor's speech rate. People converse with each other with ease, and speech rate rarely seems to cause misunderstandings. However, speech rate is surprisingly variable. Previous research has shown that speech rate differs across gender, age, and dialect (Jacewicz, Fox, & Wei, 2010; Quené, 2008). Not only does speech rate differ across talkers, substantial speech rate variation is also found within individual talkers (Miller, Grosjean, & Lomanto, 1984), depending, for instance, on the interlocutor and how noisy the environment is. To complicate matters, acoustic duration is a critical cue to many phonemic contrasts, such as the English consonants short /b/ and long /w/ (Miller & Liberman, 1979; Miller & Baer, 1983), the Maltese singleton-geminate contrast short /t/ and long /t:/ (Mitterer, 2018), the Danish vowels short /ɔ/ and long /ɔ:/ (Schachtenhaufen, 2010), and the Dutch vowels short /ɑ/ and long /a:/ (Reinisch & Sjerps, 2013), to name a few. As such, the produced length of a /b-w/ sound cues whether your fitness coach instructs you to stand with your head against "the ball" or against "the wall" during a new exercise. However, given the large amount of variation in speech rate, there is no one-to-one correspondence between acoustic cues to duration (short vs. long segment) and temporal phonological contrasts (e.g., short /b/ vs. long /w/). That is, the same acoustic material may signal different phonemic segments, depending on the surrounding speech rate: When speaking fast, individual speech sounds are shorter in duration, compared to when speaking slowly. This poses a challenge for the listener, who has to map the variable speech signal onto the intended phonemes. To accomplish this challenge, listeners need to take speech rate into account during speech processing. That is, they need to normalize for contextual speech rate. Thus, listeners perceive speech relative to the surrounding speech rate context.

1.2 Speech rate effects

The questions of how listeners normalize for contextual speech rate and the effects that speech rate contexts may have on perception have challenged language scientists for decades. As early as 1960, Pickett and Decker (1960) probed participants' perception of the test sentence *He was the [topic/top pick] of the year* spoken at different rates, to test the influence of speech rate on word segmentation. In the decades that followed, influences of speech rate on the identification of both consonants (e.g., Summerfield, 1981; Kidd, 1989) and vowels (e.g., Nooteboom, 1981) were investigated, and effects were found even when the context was unintelligible speech (Kluender, 1984), or consisted of non-speech (Gordon, 1988; Diehl & Walsh, 1989; Wade & Holt, 2005). The typical speech rate effect found in these studies is that a slow contextual rate (speech or non-speech) leads listeners to perceive a subsequent ambiguous stretch of speech as relatively short, whereas a fast context leads them to perceive it as relatively long. In other words, while listening to a slow-talking fitness coach would lead you to place your head against the ball, listening to a fast-talking one would have led you to place your head against the wall.

Many years of research have provided a wealth of evidence for the effects of rate normalization. To date, the topic continues to keep language and cognitive scientists occupied, highlighting the complexity of the question of how listeners use speech rate contexts in perception. Contextual speech rate has been found to affect the perception of subsequent and preceding phonemes (e.g., Reinisch, Jesse, & McQueen, 2011; Miller & Baer, 1983), morphemes (Brown, Dilley, & Tanenhaus, 2012), and words, influencing perception of both function words (Dilley & Pitt, 2010; Morrill, Dilley, McAuley, & Pitt, 2014; Pitt, Szostak, & Dilley, 2016) and content words (Dilley, Morrill, & Banzina, 2013; Baese-Berk, Dilley, Henry, Vinke, & Banzina, 2019). On the phoneme level, speech rate effects have been reported, for instance, by Diehl and Walsh (1989), who showed that a temporally ambiguous /b–p/ sound is more often perceived as short /p/ when followed by a short vowel, but as long /b/ when followed by a longer vowel. On the word level, Dilley and Pitt (2010) found that heavily co-articulated 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 (relative to the same function word in a faster context). Moreover, they found that function words not actually spoken could be perceived when they were embedded in fast speech. This effect, found for English, also generalizes to Russian (Dilley

et al., 2013), Mandarin (Lai & Dilley, 2016), and Finnish (O'Dell & Nieminen, 2018). These findings indicate that contextual speech rate affects perception of time-dependent contrasts in a contrastive way: A fast context makes a temporally ambiguous sound or word sound long, whereas a slow context makes that the same sound or word sound short.

1.3 Proximity of speech rate context

Speech rate effects are induced by different types of contexts, differentiated by their proximity to an ambiguous target sound or word. We can divide the speech rate contexts that listeners rely on into three types: the proximal context, the distal context, and the global context. These different types of contexts have all been shown to influence the perception of ambiguous speech sounds.

The proximal speech rate context denotes the short adjacent context rate within a distance of approximately 250 ms around an ambiguous target in both directions (i.e., approximately one preceding and one following syllable) (Newman & Sawusch, 1996; Reinisch et al., 2011; Sawusch & Newman, 2000; Summerfield, 1981). The proximal context directly precedes and follows an ambiguous stretch of speech. Evidence that this short context influences speech perception comes from Newman and Sawusch (1996) and Sawusch and Newman (2000), who explored to which extent the distance of speech rate cues influenced perception of temporally ambiguous speech sounds. Sawusch and Newman manipulated single vowels directly following a target and the consonants directly following those vowels. They tested the influences of both the vowel adjacent to a preceding /b-p/ target sound (e.g., /ʊ/ in *bush*) and the non-adjacent consonant following that vowel (e.g., /ʃ/ in *bush*) in ambiguous target words ranging from *bush* to *push*. They observed that phonemes that were temporally close to a target always affected perception of the target, compared to phonemes farther away (more than 250 ms) from the target. These results indicate that proximal duration cues can influence phonetic boundaries, independently of the rest of the surrounding rate context.

Distal speech context is at a relatively longer distance from a target sound or word. The crucial distinction from proximal context is that distal context typically constitutes the entire sentence (i.e., multiple syllables), whereas the proximal context only comprises a single sound or syllable. Moreover, distal speech rate usually concerns the non-adjacent context rate. Like proximal speech rate, distal speech rate can affect phonetic category boundaries and word segmenta-

tion, with listeners hearing shorter or longer phonemes, such as short /ɑ/ (in slow speech) and long /a:/ (in fast speech) in Dutch (Reinisch & Sjerps, 2013), and more (in a fast context) or fewer (in a slow context) morphophonological units, such as *our* in *The accountants are (our) wise advisors* in English (Dilley & Pitt, 2010). Both Reinisch and Sjerps and Dilley and Pitt manipulated distal rate, while keeping proximal speech rate constant (but note that Reinisch and Sjerps did this by inserting a silent gap before the target, whereas Dilley and Pitt used fixed-rate proximal speech). That is, the adjacent contexts (approximately one syllable either side of the target) were not rate-manipulated, in order to control for the effects of proximal rate. These studies found that distal speech rate induced speech rate effects on phonetic category boundaries and function word perception. Interestingly, these distal rate effects emerged even when the proximal context was controlled, compromising direct comparison of the target duration to adjacent context durations.

The global context rate is the speech rate beyond the sentence context in which an ambiguous target occurs, coming from previous sentences and potentially from other talkers. At present, there are only two studies that show evidence of global speech rate effects, namely Baese-Berk et al. (2014) and Reinisch (2016b). Baese-Berk et al. compared speech perception in different global rate contexts, in up to an hour of speech. Their stimuli consisted of co-articulated function words (e.g., *after her* and *are our*) that were embedded in context sentences (e.g., *Susan said those are (our) black socks*) that were expanded to different extents. They hypothesized that the slower the context speech rate was, the fewer function words participants would hear, in line with Dilley and Pitt (2010). However, instead of looking at within-sentence distal effects of speech rate like Dilley and Pitt, Baese-Berk et al. investigated the effect of the average speech rate, drawn from various speech rates, on the perception of function words.

To test influences of global speech rate, Baese-Berk et al. formed three listener groups that differed only in the average speech rate they were exposed to across the different context sentences. Group 1 listened to unmodified speech, slow speech (multiplier 1.2), and slower speech (multiplier 1.4). Group 2 listened to speech multiplied by factors of 1.2, 1.4, and 1.6, thus listening to speech that was on average slower than the speech rate in Group 1. Group 3 listened to the slowest speech, multiplied by factors 1.4, 1.6, and 1.8. The authors demonstrated a difference between groups in the perception of the targets, depending on the average rate they listened to. For example, Group 1, which was presented with unmodified speech and slower speech (1.2), heard more function words in

the slower speech (1.2) than Group 2, which was presented with the same slow speech (1.2) and even slower speech (1.4). Thus, the faster the average speech rate participants listened to, the more function words they reported in the overlapping speech rate conditions. These results indicate that participants relied on the overall speech rate of sentences across the rate conditions, with the relative difference between rates being underestimated. That is, slow speech sounded less slow in the presence of faster speech in the global speech context. This study shows that listeners keep track of the speech rate calculated over a longer period of time to interpret ambiguous stretches of speech.

Reinisch (2016b) investigated how a talker's global speech rate influenced subsequent perceptual processing of that talker's speech. Reinisch ran an experiment in which participants first listened to two talkers, a slow talker and a fast talker, followed by a test in which participants categorized isolated temporally ambiguous words spoken by these two talkers. She found that target words from the fast talker were more often categorized as long words than target words from the slow talker. Note that this effect of global speech rate is contrastive, contrary to the global rate effect in Baese-Berk et al. (2014), which is an assimilative effect. I will return to this disparity in the next section.

1.4 Towards a psychological model of speech rate effects

As described above, there is accumulating evidence that proximal and distal speech rate can induce speech rate effects, and there is also research indicating that wider speech contexts may have an influence on word recognition. When encountering a temporally ambiguous word in a context, listeners need to process the acoustic information in the word and also relate these acoustic cues to the rate cues in the context. In order to disambiguate the word as one or the other, the information from both the word itself and from the context needs to be merged for the most probable outcome. It is not yet well understood precisely when in time the different types of cues are merged and to which constraints they may be subject, particularly in the case of global speech rate.

Proximal and distal speech rate effects have been argued to involve early and automatic processes. For instance, speech rate effects are induced even when the context and the target are produced by different talkers (Newman & Sawusch, 2009), with a speech context from Talker A influencing subsequent perception of a target by Talker B. Proximal and distal rate effects also occur after self-

produced speech (Bosker, 2017b), unintelligible speech (Kluender, 1984), and even pulse trains (Wade & Holt, 2005; Bosker, 2017a). Additionally, they are unaffected by cognitive load (Bosker, Reinisch, & Sjerps, 2017). Proximal and distal speech rate effects have furthermore been shown to emerge between 300 to 400 ms after target onset in eye-tracking studies (Reinisch & Sjerps, 2013; Toscano & McMurray, 2015; Kaufeld, Ravenschlag, Meyer, Martin, & Bosker, 2019), which also supports the involvement of early perceptual processes.

As noted, only two studies have shown effects of global speech rate (Baese-Berk et al., 2014; Reinisch, 2016b), and these effects occurred in different directions (i.e., averaging effect vs. contrastive effect of global rate). It is unclear whether global speech rates from multiple talkers affect perception of phonetic category boundaries and how listeners keep track of speech rates from longer contexts. Moreover, there has been no in-depth research on the constraints on global speech rate tracking and its time course. In order to construct a psychological model that unifies proximal, distal, and global speech rate effects, more detail about the influence of wider speech rate contexts is needed. Therefore, this thesis focuses on the global speech rate effect and compares it to the distal speech rate effect.

1.5 Research question and outline

This doctoral thesis aims to illuminate how listeners use global speech rate in word recognition. It presents four studies investigating the effects of global, more distant speech rate contexts, and their separability from distal, within-sentence context effects. More specifically, I studied to what extent these effects are automatic and perceptual as compared to involving higher-level processes, by looking at their constraints and time courses. *Chapter 2* examined the effect of distal speech rate. *Chapters 3*, *4*, and *5* examined and compared the effects of distal and global speech rates. The chapters of this thesis together aimed to answer the broader research questions:

> *How do listeners use distal and global speech rate in word recognition and what are the underlying mechanisms at play during speech rate processing?*

Chapter 2 tested the hypothesis that distal speech rate normalization influences lexical access without an explicit recognition task. In the literature, studies on speech rate effects have only used explicit categorization or identification tasks to test rate normalization, making it difficult to investigate perceptual pro-

cesses without contributions from a decision-making level. By using an implicit measure of speech rate processing, contributions of operations at a decision making level could be excluded in order to measure the extent to which speech rate normalization is automatic.

The task employed was a two-alternative forced choice (2AFC) task, involving lexical decision in a cross-modal repetition priming paradigm. Lexical decision tasks are typically used to measure the speed of lexical access. A cross-modal priming paradigm involves a prime in one modality (e.g., auditory), and a target in another modality (e.g., visual) and it is used to measure the influence of a prime on the processing of a target. A lexical decision task is an explicit task, but the task concerns the target, not the prime. As such, this task enabled us to test an implicit effect of the prime on the target.

In this study, participants listened to rate-manipulated sentences with an auditory prime word with the Dutch /ɑ-a:/ distinction, after which they had to indicate whether a given written word on the screen was a word or a non-word. Participants have been shown to respond faster to a target if the prime and the target are identical, compared to if they are different. Therefore, I assumed that if participants responded fast to a target word on the screen, they were more likely to also have heard that word in the prime sentence. I predicted that participants would respond faster to short /ɑ/-vowel words on the screen if the spoken prime sentence was slow, and slower to the same target words if the prime sentence was fast.

Chapter 3 probed the hypothesis that the global speech rate of one talker can affect subsequent perception of the speech rate of another talker. Previous research has investigated the effects of global speech rate on perception of function words (Baese-Berk et al., 2014) and word categorization (Reinisch, 2016b), but did not explicitly test the role of talker voice and within-talker speech rate variability. Therefore, three behavioral experiments manipulated inter-talker and intra-talker speech rate variation to study whether the global speech rate effect is talker-specific or talker-independent. I used a 2AFC task. This 2AFC task involved explicit categorization of an ambiguous sound as belonging to one of two response options. In the 2AFC tasks in this study, participants categorized temporally ambiguous Dutch words (e.g., *stad/staat* "city/state") with the temporally distinctive vowels, short /ɑ/ and long /a:/. These words were embedded in rate-manipulated precursor sentences.

In the experiments, participants were assigned to one of two groups. The high-rate group was presented with neutral rate from Talker A and fast speech

from Talker B, and the low-rate group was presented with the same neutral rate from Talker A and slow speech from Talker B. Participants' responses informed us as to which target word, an /ɑ/-vowel word or an /a:/-vowel word, they had heard in the precursor sentences spoken at different speech rates by Talker A and B. The distal speech rate effect was tested by comparing responses in the within-groups rate conditions and the global speech rate effect was tested by comparing responses in the between-groups neutral rate conditions.

If inter-talker rate variability induces a global speech rate effect, but intra-talker variability does not, the global speech rate effect is talker-specific, with the speech rate of one talker affecting perception of another talker. However, if both inter-talker and intra-talker speech rate manipulation induces global speech rate effects, the effect is talker-independent, with the average speech rate (high vs. low) being the source of the effect.

Chapter 4 tested the hypothesis that self-produced speech rate induces global speech rate effects on other-produced speech rates. Effects of distal speech rate have been shown to be induced by self-produced speech, whether it is perceived during production or passively (i.e., listening to playback of one's own speech) (Bosker, 2017b). Bosker investigated effects of self-production on hearing an ambiguous target word immediately after having produced a sentence oneself at a fast or slow rate. He observed a difference in the perception of target words between the condition in which participants were instructed to speak fast compared to the condition in which they had to speak slowly. Given that, in natural conversation, one's own rate is often context for the speech of others, the distal speech rate effect found in Bosker (2017b) raises the question whether listeners also include their own speech in their estimate of a global rate. However, global and distal effects seem to involve distinct mechanisms: Whereas distal speech rate seems to involve exclusively perceptual normalization, *Chapter 3* suggested that global speech rate involves additional, higher-level cognitive adjustments. Therefore, self-produced speech in the global context may not affect perception of another talker's speech.

The question of whether a listener's own speech rate induces a global speech rate effect was addressed using a paradigm similar to that in *Chapter 3*. The high-rate group produced speech at a fast rate and the low-rate group produced speech at a slow rate. The groups were compared on their perception of ambiguous Dutch /ɑ, a:/ words in neutral rate speech. If listeners perceive the global speech rate of another talker relative to their own speech rate, the high-rate group should report hearing more long /a:/ words than the low-rate group.

Three 2AFC experiments tested the effects of self-produced speech, playback of self-produced speech, and other-produced speech rate on word categorization.

Chapter 5 investigated the hypothesis that the distal rate effect emerges earlier in time than the global rate effect. The previous chapters found that global and distal speech rate effects have different prerequisites: In order to track global speech rate, listeners take into account talker and rate consistency, whereas distal speech rate tracking generalizes across talkers. Bosker et al. (2017) speculated that the processes underlying the two effects may be different, with the distal rate effect being automatic and perceptual and the global effect involving higher-level cognitive adaptations. This model of acoustic context effects predicts that global and distal speech rate are processed at distinct time points during speech perception. To track the time courses of these two context effects, *Chapter 5* applied an eye-tracking paradigm to measure online auditory language processing (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) in addition to a 2AFC task. In the experiment, a high-rate and a low-rate group were compared on their button-press responses and eye-movements, listening to talkers speaking at different speech rates (fast and neutral vs. slow and neutral). If the between-groups global effect arises later than the within-groups distal effects, one would conclude that the global rate tracking needs more processing time than distal rate tracking. Alternatively, distal and global processing may overlap in time.

Finally, *Chapter 6* summarizes and discusses the four experimental chapters. In these chapters, influences of distal and global speech rates on speech perception were found to differ, with the global speech rate effect being constrained by factors not found for the distal rate effect. For instance, too much within-talker variability causes global rate tracking to fail. Relating this to the perception of driving speed, driving on a main road seems fast after coming from a city road, because driving on the main road is relatively fast compared to what one has just experienced. However, if the main road is bumpy, causing one to decelerate and accelerate over and over again, this variation in driving speed impairs tracking of the overall speed. Finally, *Chapter 6* relates the findings to mechanisms that may underlie speech rate effects and discusses how these may be implemented in a model of speech rate processing.

1.6 Reading guide

Chapters 2 to *5* were written as independent journal articles and therefore overlap to some extent in their literature reviews and methods sections. Furthermore, note that *Chapters 3* and *4* divide speech rate contexts defined by their proximity to a target speech sound into only two types: local and global. In these chapters, local speech rate is used as an umbrella term for within-sentence contexts, unifying proximal and distal speech rate contexts. The empirical chapters in this thesis only examined distal and global speech rate effects.

2 | Listeners normalize speech for contextual speech rate even without an explicit recognition task[1]

Abstract

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. While 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.

[1] Adapted from *Maslowski, M., Meyer, A. S. & Bosker, H. R. (2019). Listeners normalize speech for contextual speech rate even without an explicit recognition task. The Journal of the Acoustical Society of America, 146(1), 179–188.*

2.1 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, Souther, & Convis, 1980; Miller, 1981).[2] 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 & 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 & Baer, 1983), voicing contrasts (e.g., shift between /b/ and /p/; Gordon, 1988; Summerfield, 1981), singleton-geminate contrasts (Mitterer, 2018), word segmentation (Reinisch et al., 2011; Pickett & Decker, 1960), and reduced word forms (Baese-Berk et al., 2014; Dilley & 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.

[2] 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.

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., 2019). Kaufeld et al. compared effects of knowledge-based (morphosyntactic gender marking) and signal-based (speech rate) cues in a two-alternative forced choice (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 supports the view 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 & 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 (Newman & Sawusch, 2009; Bosker, 2017b; Maslowski, Meyer, & Bosker, 2018, 2019a). 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 & Reinisch, 2017), talker identity (habitually fast talkers induce more long responses; Bosker & Reinisch, 2015; Maslowski et al., 2018, 2019a; Reinisch, 2016b), 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 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 & 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, Newman, Dilley, & Idsardi, 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 (Reinisch & Sjerps, 2013; Toscano & McMurray, 2015; Kaufeld et al., 2019), 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 nor-

malized 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 some 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 /mɑt/*), 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 2.1). Lexical decision tasks require lexical access to the orthographic string (Monsell, Patterson, Graham, Hughes, & Milroy, 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 & 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

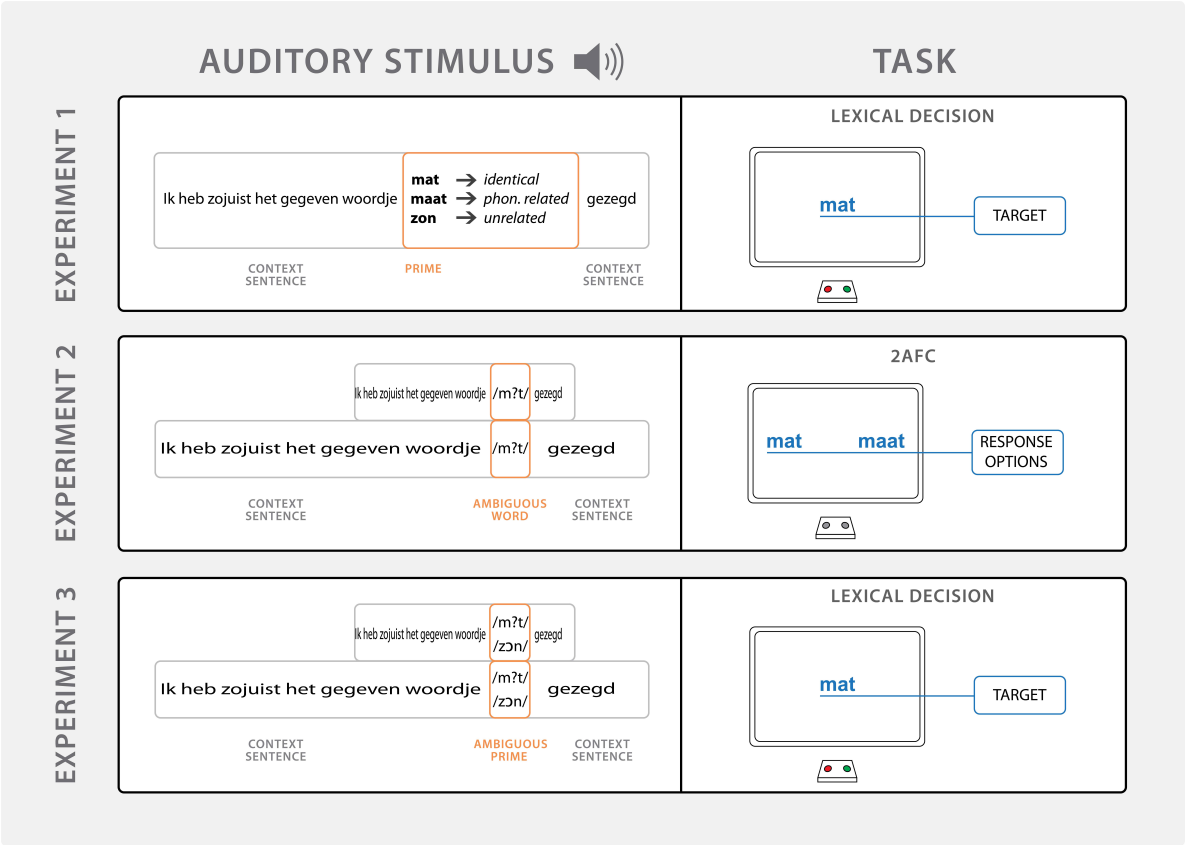

Figure 2.1: **Experimental design of Experiments 1–3.** Experiment 1 involved a cross-modal repetition priming paradigm with a lexical decision task. Auditory primes were either identical, phonologically related, or unrelated to the following orthographic target words. Experiment 2 tested rate normalization in a two-alternative forced choice (2AFC) task. Auditory stimuli consisted of spectrally ambiguous Dutch /ɑ, a:/ vowels embedded in fast and slow context sentences. Experiment 3 combined the methods of Experiments 1 and 2, testing rate normalization of ambiguous primes with a lexical decision task.

to their primes than for non-identical primes (Forbach, Stanners, & Hochhaus, 1974; Forster & Davis, 1984; Scarborough, Cortese, & Scarborough, 1977). This hypothesis was confirmed. Experiment 2 then validated our stimulus set, this time using ambiguous /ɑ, a:/ 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–a:/ vowel as long (i.e., /a:/), 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 type of target word on the screen.

2.2 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. First, Experiment 1 aimed at validating the constructed stimuli for finding differences in reaction times in phonologically related pairs. Second, 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.

2.2.1 Methods

Participants. Twelve native Dutch participants (female = 9, M_{age} = 22 yr) 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).

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, Brysbaert, & New, 2010). In total, there were 90 /ɑ, a:/ minimal pairs. Each member of each pair was matched with an unrelated prime with the properties described above (see Appendix). 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).

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 s 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, a:/ 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 min.

2.2.2 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. Figure 2.2 summarizes the reaction times (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.

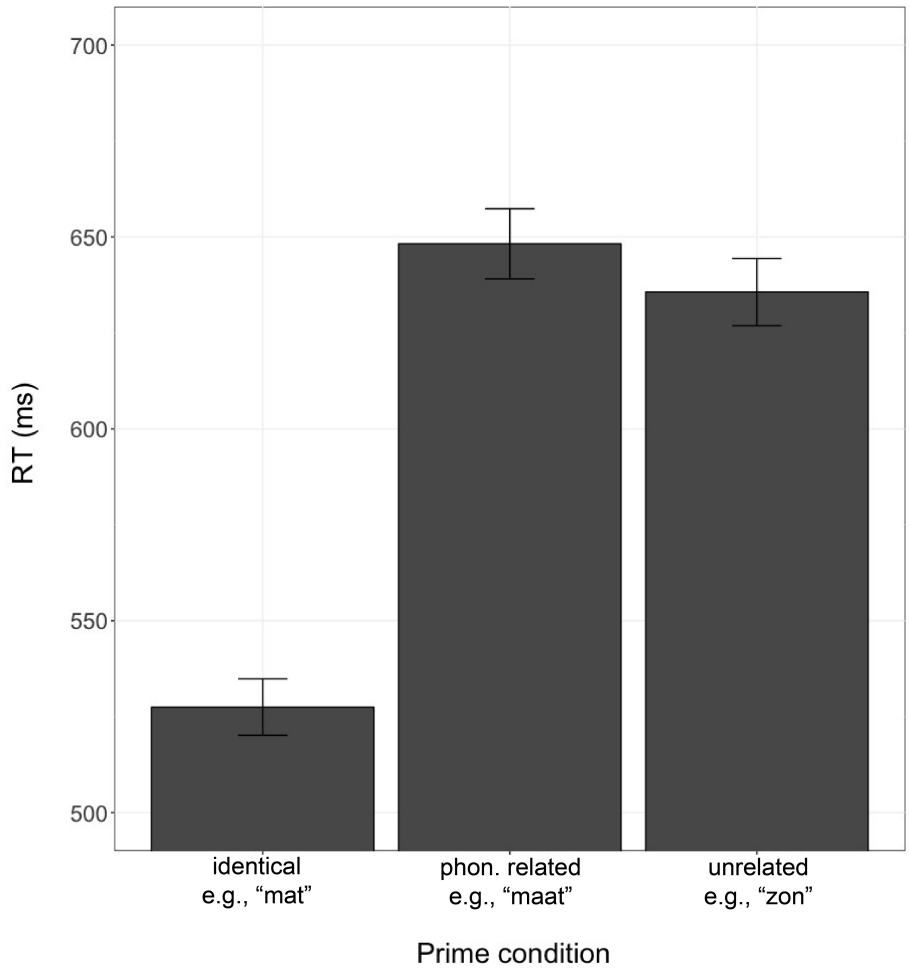

Figure 2.2: **Mean reaction times of Experiment 1 (cross-modal repetition priming) for correct responses in three Prime Conditions (identical, phonologically related, and unrelated).** Error bars indicate the standard error of the mean.

The RTs of accurate experimental trials (10.19% incorrect experimental trials excluded) were tested using a Linear Mixed Model (LMM) from the `lme4` package (Bates, Mächler, Bolker, & Walker, 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 by Barr, Levy, Scheepers, and Tily (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 with the lowest estimated variance were gradually removed 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 ($\beta = -106.068, t = -4.337, p = 0.001$)[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.

2.3 Experiment 2: Rate normalization in 2AFC task

Experiment 2 assessed rate normalization in a 2AFC task with the same /ɑ, a:/ words as in Experiment 1. Specifically, only the auditory primes from Experiment 1 were used. This time, however, the precursor sentences surrounding the /ɑ, a:/ words were rate-manipulated (fast vs. slow), and participants categorized temporally and spectrally ambiguous /ɑ, a:/ 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 2.1). 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–a:/ vowel towards a long vowel /a:/, whereas a slow context biases perception of the same vowel towards hearing /ɑ/.

[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`.

2.3.1 Methods

Participants. Fourteen native Dutch participants (female = 12; M_{age} = 24 yr) 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.

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 Appendix) 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, Van Hout, & Smits, 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 & 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 =) 0.11 of the /ɑ/-vowel, and (8/9 =) 0.89 of the /a:/-vowel, the second step (2/9 =) 0.22 and (7/9 =) 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 2.3). 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 Appendix) were excluded, as a consequence of not being perceived as suf-

ficiently 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 & 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 75 minimal pairs, one of the two sentence recordings of a pair was used as 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.

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 min.

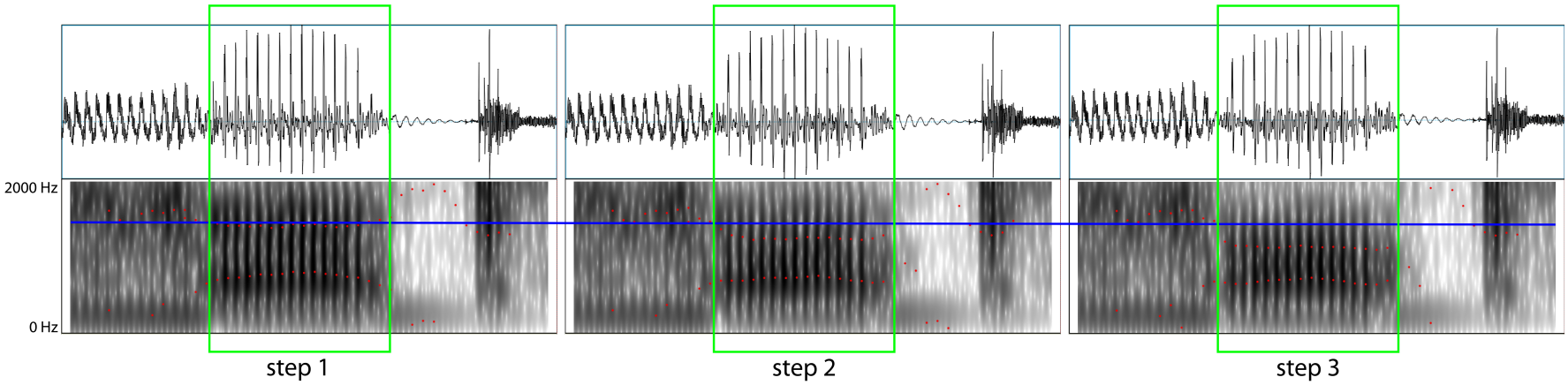

Figure 2.3: **Spectrograms (0 – 2000 Hz) of the three steps of the same minimal pair "hak/haak".** Step 1 is most /a:/-like (relatively high F2) and step 3 is most /ɑ/-like (relatively low F2). The green rectangles show the vowel portions. The red dots show the formant trajectories. The blue line is drawn to more easily see that F2 decreases from the left panel to the right.

2.3.2 Results and discussion

The categorization data of Experiment 2 are represented in Figure 2.4. 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.

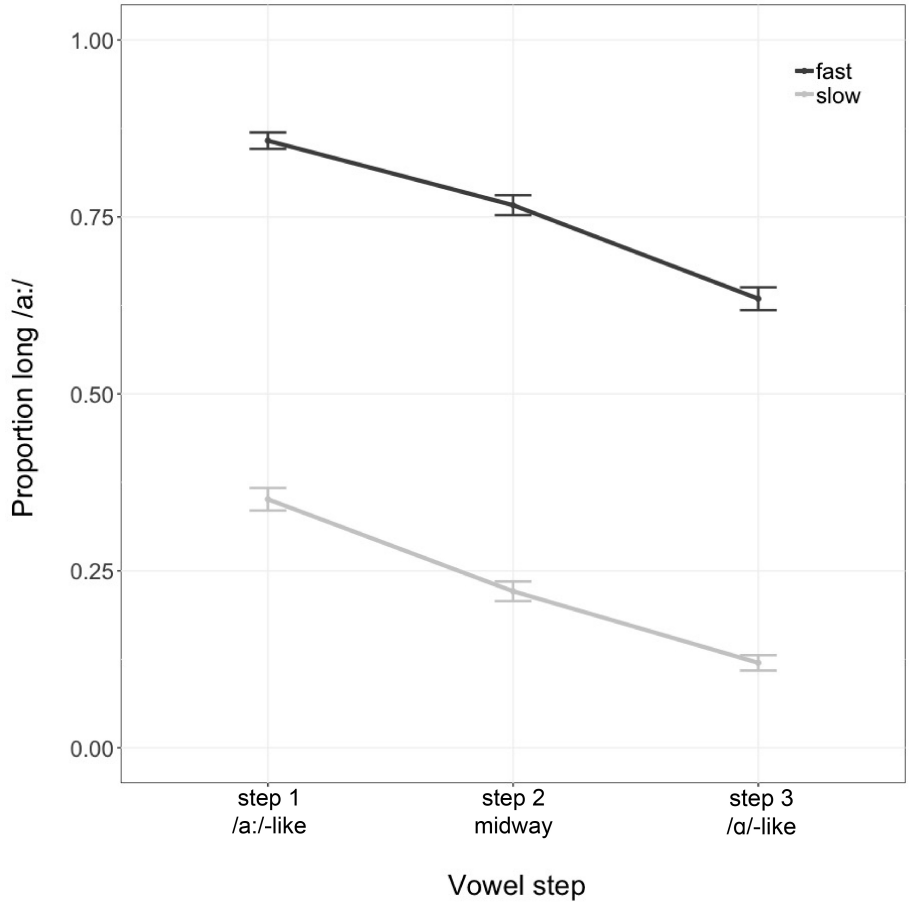

Figure 2.4: **Average categorization data of Experiment 2 (rate normalization in 2AFC task).** The x-axis indicates Vowel Step (1: /a:/-like; 3: /ɑ/-like). Colours indicate Rate Condition, with the fast condition shown in dark grey and the slow condition shown in light grey. Error bars indicate the standard error of the mean.

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 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; Reinisch & Sjerps, 2013; Kaufeld et al., 2019).

2.4 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 panel of Figure 2.1). This experiment tested whether speech rate effects are induced even when no explicit attention is drawn to the spectrally ambiguous word.

2.4.1 Methods

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

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).

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 min.

2.4.2 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 2.5 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 such as "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 ($\beta = 26.459, t = 2.356, p = 0.020$)[4], 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, Bartolotti, Chabal, & Shook, 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 those to 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 am-

[4] 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`.

biguous word in the prime.

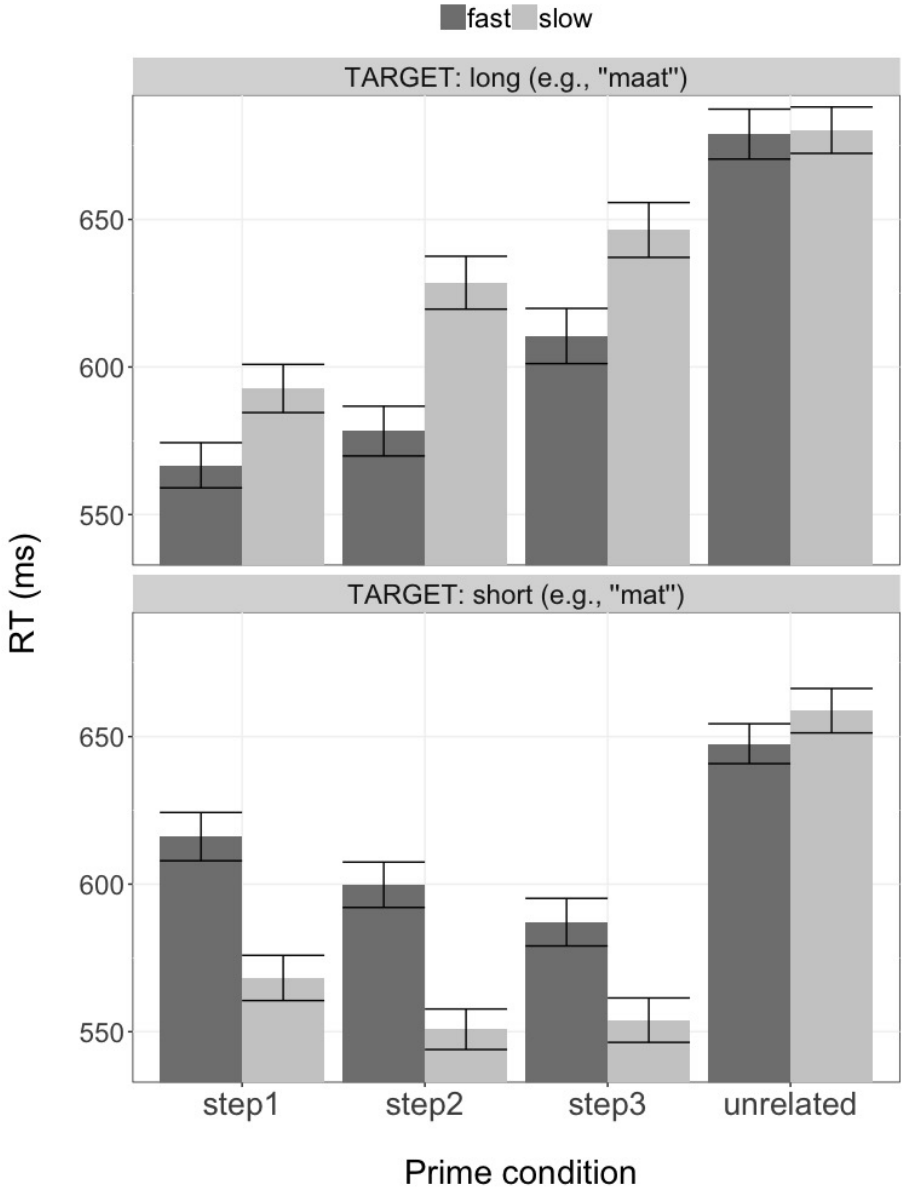

Figure 2.5: **Mean reaction times of Experiment 3 (rate normalization in repetition priming) for correct responses in four Prime Conditions.** These conditions consisted of Vowel Step 1 (most /a:/-like), 2 (midway between /a:/ and /ɑ/), and 3 (most /ɑ/-like), as well as an unrelated control condition. 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.

2.5 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., /a:/ rather than /ɑ/; Reinisch & 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 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 rate effects have been shown to be insensitive to talker voice changes (Newman & Sawusch, 2009; Maslowski et al., 2018, 2019a) 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 & 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 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 (Pitt et al., 2016; Bosker & Reinisch, 2017; Maslowski et al., 2018, 2019a). 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.

Appendix

Stimulus characteristics of Dutch minimal /ɑ, a:/ pairs in Experiment 1 to 3. The first column shows the pair numbers and the second column the member words and their English translations. The third and fourth column show the lexical frequency per million (as measured by Subtlex; Keuleers et al., 2010) and the dominant part-of-speech (Keuleers et al., 2010) of the minimal pair members respectively. The last column shows the unrelated prime words matched on word frequency and part-of-speech with a minimal pair member, followed by their English translations.

	Dutch minimal pair	*Word freq.*	*POS*	*Unrelated prime*
1	al (already)	2344.26	ADV	ook (also)
	aal (eel)	0.8	N	stift (marker)
2	arts (doctor)	32.79	N	schot (shot)
	aards (earthly)	1.14	ADJ	holst (dead [of night])
3	as (ash/axis)	15.39	N	ruil (exchange)
	aas (ace/bait)	9.06	N	pret (fun)
4	bal (ball)	80.63	N	rug (back)
	baal (am fed up with)	3.96	V	pleit (plead)
5	ban (ban)	3.16	N	klif (cliff)
	baan (job)	158.45	N	recht (right)
6	bar* (bar)	53.83	N	dorp (village)
	baar* (give birth)	0.89	N	slurf (trunk [of elephant])
7	bars* (bars)	2.7	N	milt (spleen)
	baars* (bass)	1.07	N	juk (yoke)
8	bas (bass)	1.92	N	gen (gene)
	baas (boss)	167.21	N	mond (mouth)
9	bad (bath)	42.58	N	feit (fact)
	baat (benefit)	2.47	N	erf (yard)
10	blad* (leaf)	11.14	N	muis (mouse)
	blaat* (bellow)	0.09	V	gom (erase)
11	brak (broke)	21.08	V	kleed (dress)
	braak (breaking)	0.91	N	zool (sole)
12	dat (that)	22077.22	PRO	het (the)
	daad (deed)	16.65	N	grot (cave)

13	dag* (day)	848.56	N	huis (house)
	daag* (dawn)	14.22	V	smeer (smear)
14	dal** (valley)	3.89	N	som (sum)
	daal** (descent)	1.01	V	wreef (rubbed)
15	das* (tie)	13.38	N	vloot (fleet)
	daas* (scatterbrained)	0.11	ADJ	ferm (firm)
16	draf (trot)	0.32	N	vijl (file)
	draaf (trot)	0.78	V	stuit (am held up)
17	gaf (gave)	191.52	V	zoek (search)
	gaaf (intact)	39.22	ADJ	flink (robust/firm)
18	gap** (buddy)	0.32	N	slib (silt)
	gaap** (yawn)	0.39	N	duin (dune)
19	gas (gas)	26.41	N	koers (course)
	gaas (netting)	0.82	N	kuip (tub)
20	gat (hole)	49.97	N	neef (cousin)
	gaat (goes)	2265.77	V	kom (come)
21	graf (grave)	34.53	N	brood (bread)
	graaf (count)	19.55	N	tekst (text)
22	hak (chop)	8.19	V	schuif (push)
	haak (hook)	12.65	N	link (link)
23	hal (hall)	20.38	N	wolf (wolf)
	haal (fetch)	302.29	V	stop (stop)
24	halt (halt)	17.49	N	brein (brain)
	haalt (fetches)	56.62	V	gooi (throw)
25	hard (hard)	159.46	ADJ	kwijt (lost)
	haard (fireplace)	6.43	N	romp (trunk [of body])
26	had (had)	2106.2	V	zien (see)
	haat (hate)	151.32	V	leg (lay)
27	jacht (hunt)	23.99	N	bril (glasses)
	jaagt (hunts)	10.61	V	stort (fall/deposit)
28	kak (shit)	0.8	N	lier (lyre)
	kaak (jaw)	5.12	N	toets (test/key)
29	kap (hood)	12.28	V	eis (requirement)
	kaap (cape)	1.1	N	lel (lobe)
30	kas (greenhouse)	4.37	N	rits (zipper)
	kaas (cheese)	22.85	N	nicht (cousin)

31	knak (kink)	0.09	N	fust (cask)
	knaak (guilder)	0.07	N	bies (piping)
32	knap (good-looking/clever)	68.05	ADJ	links (left)
	knaap (boy)	8.76	N	doek (cloth)
33	krak (crack)	0.16	N	zerk (tombstone)
	kraak (robbery)	1.99	N	snob (snob)
34	kwal (jellyfish)	1.78	N	drum (drum)
	kwaal (ailment)	1.6	N	dop (cap)
35	lag (lay [from lie])	87.04	V	rot (rot)
	laag (low)	31.1	ADJ	wit (white)
36	lat (slat)	2.01	N	toer (trip)
	laat (omit/leave)	2032.73	V	moet (must/should)
37	mag** (may)	1058.17	V	kijk (look)
	maag** (stomach)	23.55	N	hof (court)
38	macht (power)	83.99	N	eind (end)
	maagd (virgin)	22.43	N	fiets (bicycle)
39	mak (tame)	1.37	ADJ	stug (stiff)
	maak (make)	648.19	V	blijf (stay)
40	mal (mould)	8.28	N	vest (cardigan)
	maal (meal)	17.29	N	golf (wave/golf)
41	man (man)	1403.81	N	tijd (time)
	maan (moon)	42.1	N	beurt (turn)
42	mand (basket)	4.3	N	wrok (resentment)
	maand (month)	93.64	N	geest (spirit)
43	mat (mat)	3.59	N	sul (softy/dope)
	maat (size)	69.18	N	zon (sun)
44	nat (wet)	30.57	ADJ	eng (scary)
	naad (seam)	1.78	N	strijk ((doing the) ironing)
45	nam (took)	116.08	V	stuur (steer/send)
	naam (name)	470.6	N	uur (hour)
46	nar** (fool)	2.33	N	ui (onion)
	naar** (to(wards))	4447.55	PREP	voor (to/for)
47	part (piece)	4.07	N	heup (hip)
	paard (horse)	83.63	N	rij (queue)
48	plat (flat)	18.98	ADJ	vies (dirty)
	plaat (plate/record)	13.68	N	ton (cask/ton)

49	rad (wheel)	2.31	N	wicht (child)
	raad (council)	83.01	N	ziel (soul)
50	ram (Aries)	8.37	N	bocht (bend)
	raam (window)	70.84	N	volk (people)
51	rap (quick)	5.15	ADJ	zuid (south)
	raap (turnip)	9.86	N	ploeg (crew/plow)
52	ras (race)	16.46	N	knop (button)
	raas (rage)	0.14	V	ent (graft)
53	sap* (juice)	7.09	N	hint (hint)
	saab* (Saab)	0.69	N	grind (gravel)
54	schaf* (procure)	0.32	V	lest (quenches)
	schaaf* (plane/graze)	0.21	V	gruist (pulverizes)
55	schap (shelf)	0.37	N	dooi (thaw)
	schaap (sheep)	6.54	N	pomp (pump)
56	schrap* (cross off)	6.54	V	sleep (drag)
	schraap* (scrapings)	0.27	N	blos (bloom)
57	slag (warp)	63.94	N	helft (half)
	slaag (beating)	4.6	N	juf (teacher)
58	slak (snail)	2.38	N	dunk (opinion)
	slaak (heave [utterance])	0.11	N	mees (tit)
59	slap (slack)	7.11	ADJ	wijd (wide)
	slaap (sleep)	112.9	N	grond (ground)
60	smak (smack)	1.92	N	troef (trump)
	smaak (flavour)	29.18	N	huur (rent)
61	span (span)	2.17	N	bok (male goat)
	spaan (oar)	0.27	N	grut (trash/toddlers)
62	spar (spruce)	0.14	N	teil (washtub)
	spaar (save (up))	10.02	V	klets (chat)
63	sprak (spoke)	53.56	V	leer (learn)
	spraak (speech)	1.05	N	drift (anger/passion)
64	staf (scepter)	11.62	N	trui (sweater)
	staaf (bar)	1.67	N	dij (thigh)
65	stak (stabbed)	17.68	V	lust (like)
	staak (strike)	3.41	V	poets (clean)
66	stal (cowshed)	22.07	N	pest ((bubonic) plague)
	staal (steel)	10.18	N	non (nun)

67	stand (posture/state)	15.71	N	tuig (gear/scum)
	staand (standing)	1.51	V	print (print)
68	star** (frozen/rigid)	8.03	ADJ	bros (brittle)
	staar** (stare)	4.48	V	flirt (flirt)
69	start (start)	37.11	N	blik (look/can)
	staart (tail)	17.95	N	joch (lad)
70	stad (city)	272.61	N	pijn (pain)
	staat (stands)	652.88	V	spijt (regret)
71	tak (branch)	8.48	N	steeg (alley)
	taak (task)	42.63	N	lunch (lunch)
72	tal (number)	0.21	N	mie (Chinese noodles)
	taal (language)	36.29	N	pers (press)
73	vacht* (fur)	3.32	N	klos (chock)
	vaagt* (blurs)	0.09	V	riek (smell)
74	vak (section/subject)	22.02	N	shirt (shirt)
	vaak (often)	180.02	ADV	neer (down)
75	val (fall)	115.46	N	oom (uncle)
	vaal (faded)	0.23	ADJ	pril (early)
76	vat (barrel)	19.05	N	loon (pay)
	vaat (dishes)	1.42	N	korst (crust)
77	vlag (flag)	17.79	N	soep (soup)
	vlaag (gust)	1.12	N	mok (mug)
78	vracht (freight)	5.24	N	duif (pigeon)
	vraagt (asks)	103.2	V	hoef (need)
79	wacht (wait)	834.29	V	geef (give)
	waagt (risks)	3.84	V	bof (am lucky)
80	wak (hole)	0.41	N	pul (tankard)
	waak (watch)	1.37	N	clou (point)
81	want* (because/as)	419.08	CONJ	toen (when/then)
	waant* (imagine)	0.64	V	gist (ferment)
82	war (tangle)	3.43	N	beek (brook)
	waar (where)	3198.67	PRO	hem (him)
83	was (was)	5303.09	V	zijn (be)
	waas (haze)	0.91	N	lor (rag)
84	wrak (wreck)	9.33	N	pil (pill)
	wraak (revenge)	44.02	N	ster (star)

85	zat (sat)	339.98	V	heet (am called)
	zaad (seed)	6.33	N	bui (squall)
86	zag (saw [from see])	461.29	V	krijg (receive)
	zaag (saw)	3.54	N	rund (bovine)
87	zacht (soft)	22.85	ADJ	puur (pure)
	zaagt (saws)	0.39	V	stoof (stew)
88	zak (pocket)	96.87	N	vuur (fire)
	zaak (business)	239.34	N	hulp (help)
89	zal (shall)	2198.25	V	doen (do)
	zaal (hall)	15.41	N	nood (distress)
90	zwart (black)	57.31	ADJ	rechts (right)
	zwaard (sword)	37.48	N	pond (pound)

* Excluded in Experiments 2 and 3 as a consequence of low accuracy (< 50%) for at least one member of the pair.
** Excluded in Experiments 2 and 3 as a consequence of the pair not being perceived as ambiguous between the two members.

3 | How the tracking of habitual rate influences speech perception[1]

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. A 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 Talkers A and B, while maintaining the average rate differing between groups. In Experiment 3, 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.

[1]Adapted from *Maslowski, M., Meyer, A. S., & 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.*

3.1 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). In the present study, we 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 way. 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. We refer to this observation 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 et al., 2011; Summerfield, 1981; Wade & Holt, 2005). For instance, perception of the Dutch vowel contrast between short /ɑ/ and long /a:/ is biased toward long /a:/ in a fast, compared with a slower, speech context (Reinisch & Sjerps, 2013). A fast context makes an ambiguous vowel sound 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 a talker other than that of 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 normaliza-

tion processes that arise early in perception (Bosker, 2017a; Bosker et al., 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 (2016b) 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-min dialogue in which one woman spoke fast and another woman spoke slowly. After this exposure phase, participants categorized isolated words (i.e., words presented without a speech context) with temporally ambiguous vowels (midway between German /a/–/a:/) that had been 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 instance, for the (slow or fast) speech rate at 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' (2016b) 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 reapplied 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 (2016b) 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, such that estimates have been found to depend on the properties of the stimuli judged previously (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 PBSs (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:/ vowel sounds 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 produced by a man and a woman who had 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-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-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. First, 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 (i.e., habitual rate of Talker B influenced perception of Talker A) and talker-independent (i.e., 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.

3.2 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. In addition, the aim of Experiment 1 was to test the magnitude of these local contextual effects in our stimuli to compare them with possibly diverging patterns resulting from differences in habitual speech rate in the subsequent experiments.

3.2.1 Method

Participants. Participant were 16 native Dutch women ($M_{age} = 23$) with no hearing, visual, or reading deficits who were recruited from the Max Planck Institute participant pool. Only a sample of women was obtained, because women 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, because for these participants, the stimuli would have been 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. Talkers were a native Dutch man and woman who were recorded producing multiple tokens of two sets of four sentences (see Table 2.1), with 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*; see Table 2.1).

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

Table 3.1: **Stimulus set of eight Dutch stimulus sentences (English paraphrase below).** Two talkers were recorded producing a set of eight Dutch stimulus sentences. 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].

	Sentences and translations
1	[Peter fluisterde Ilse iets verkeerd in en toen hoorde] **Ilse het** tak-/taak**je** [gezegd worden]. *Peter whispered something in Ilse's 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]. *While 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.*

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 regardless 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 et al., 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 80–120 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 varying in 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 12 participants had to classify a set of vowels for each of the two talkers (five F2 values × five vowel durations × two 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. Note that these vowel manipulations did not result in audible changes in sound quality of the target vowel. For the resulting temporally and spectrally manipulated vowels, the intensity and pitch contours were controlled. The consonantal frames for the vowels were 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 (eight context phrases × three rates × five vowel durations × two talkers).

Procedure. Stimulus presentation was controlled by Presentation software (v16.5; Neurobehavioral Systems, Albany, CA). 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. This screen was then replaced by another screen with two response options (e.g., *takje* and *taakje*), after which participants had 4 s 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 s, a missing response was recorded. The 240 stimuli were presented to each participant once in a randomized order. One session lasted approximately 25 min.

3.2.2 Results and discussion

Figure 3.1 summarizes the categorization data (proportion of /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 with 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 et al., 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, which amounts to 16 ms), and their interaction. In addition, 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 toward long /a:/ and slow phrases inducing a perceptual bias toward 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.

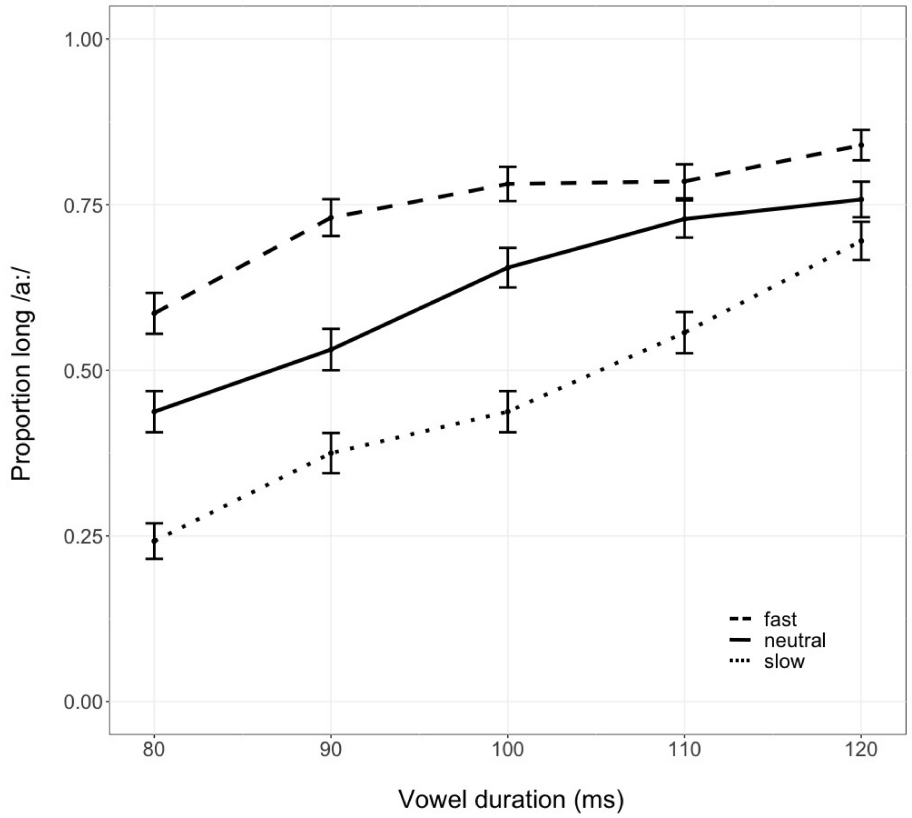

Figure 3.1: **Average categorization data of Experiment 1 (local rate effects).** The x-axis indicates Vowel Duration (80–120 ms). The fast context rate is indicated by the dashed line, neutral by the solid line, and slow by the dotted line. Error bars represent the standard error of the mean.

3.3 Experiment 2: Inter-talker variation

In Experiment 2, we aimed to evaluate whether talkers' long-term habitual speech rates were perceived absolutely or relatively (to other talkers). This was done by comparing listeners' categorization responses to vowels midway 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 at a slow rate. If the perception of the neutral habitual speech rate of Talker A was influenced by the habitual rate of Talker B, we would expect differential perception of Talker A's speech in the two groups.

3.3.1 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 six participants were excluded, because their responses were outside the set performance range described in Experiment 1, resulting in two pseudorandom groups, each comprising 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 (i.e., the average speech rate was low; see Figure 3.2). Rate assignment to talker was counterbalanced across participants, such that Talker A was the woman half of the time. Therefore, each participant listened to 80 of the 240 unique stimuli (eight context phrases × five vowel durations × two 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 s to indicate which word they had heard at the position of the question mark. One session lasted for a duration of approximately 40 min in the high-rate group and 50 min in the low-rate group.

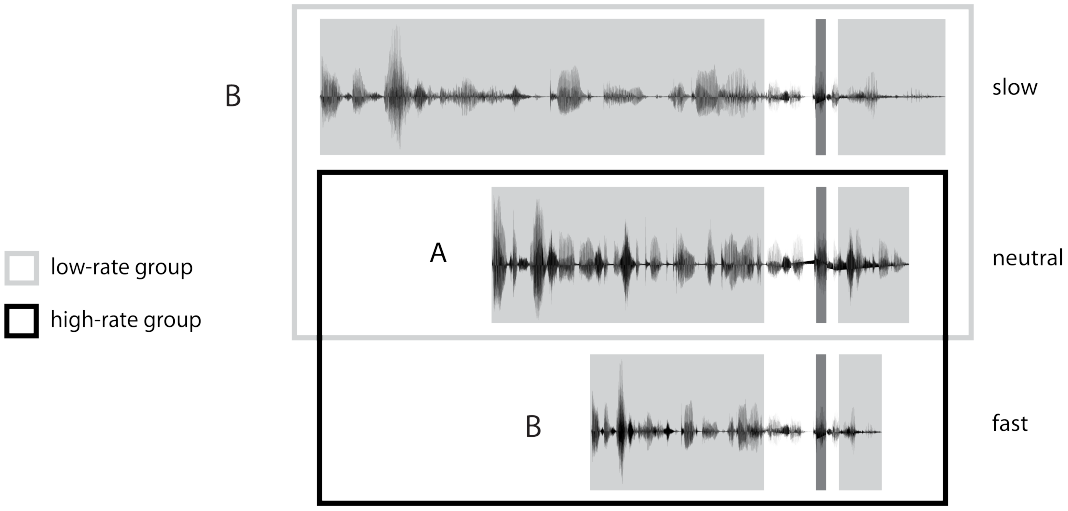

Figure 3.2: **Experimental design of Experiment 2 (inter-talker variation).** Each stimulus sentence consisted of a rate-manipulated (fast, neutral, slow) context phrase (light grey background), buffers on either side of the target (fixed duration; white background) and the target vowel itself (dark grey background). The 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 Talker A at a neutral rate and Talker B at a fast rate (black box).

3.3.2 Results and discussion

Figure 3.3 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-groups condition (high/low average rate) with the within-group con-

dition (fast/neutral/slow trial). Rate Condition consisted of four contiguous levels of rate, corresponding to the four lines represented in Figure 3.3, namely high_fast, high_neutral, low_neutral, and low_slow (where the between-groups 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 important, performance in low_neutral versus high_neutral contexts was also significantly different (i.e., a between-groups 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 3.1 and Figure 3.3 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 with 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.

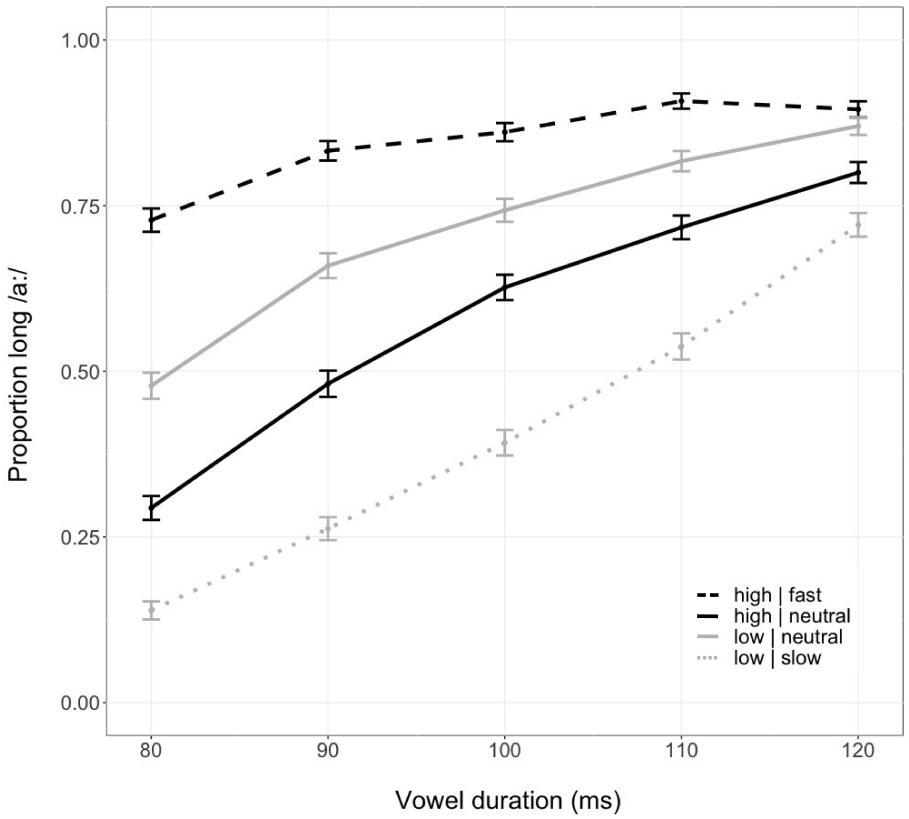

Figure 3.3: **Average categorization data of Experiment 2 (inter-talker variation).** The x-axis indicates Vowel Duration (80–120 ms). The fast rate condition is indicated by the dashed line, neutral by the solid line, and slow by the dotted line. Colors indicate Group, with the high-rate group shown in dark grey and the low-rate group shown in light grey. Error bars represent the standard error of the mean.

3.4 Experiment 3: Intra-talker variation

In Experiment 2, a discrepancy was found between groups in the perception of Talker A. This could either be a result of listeners tracking talker-specific habitual rates (e.g., 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. Whereas 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, respectively, 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.

3.4.1 Method

Participants. Native Dutch women ($N = 40, M_{age} = 21$) 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 pseudorandom 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 (eight context phrases × five vowel durations × two rates × two talkers). These stimuli were presented in a randomized order in each of three blocks. One session lasted for a duration of approximately 50 min in the high-rate group and 60 min in the low-rate group.

3.4.2 Results and discussion

Figure 3.4 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 perfor-

mance, 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.

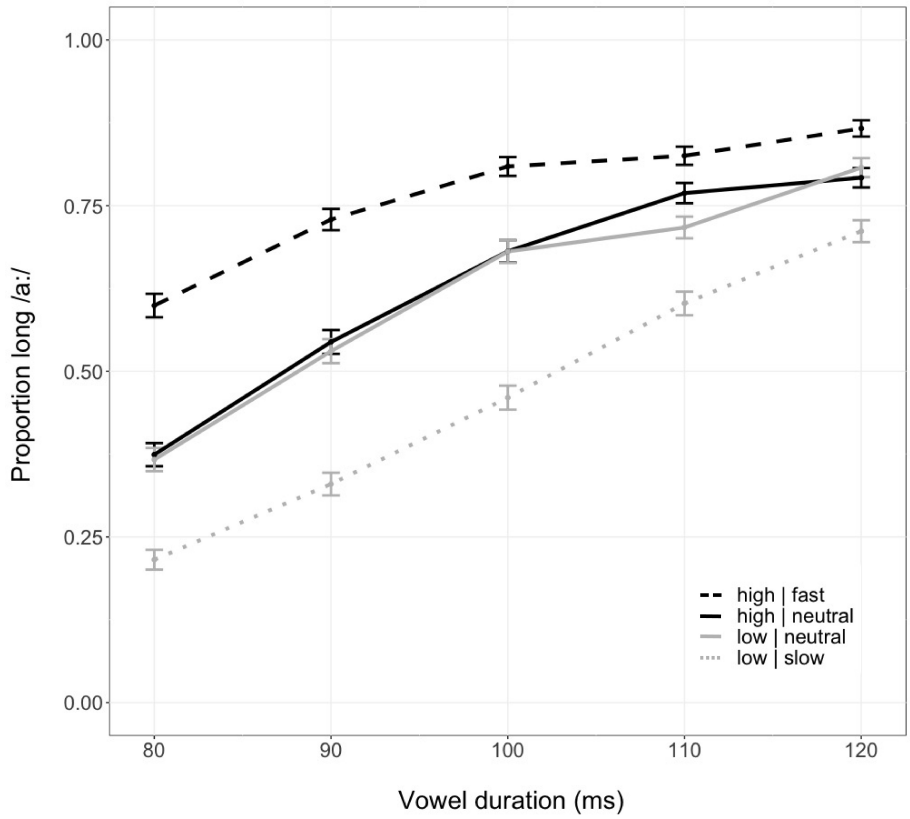

Figure 3.4: **Average categorization data of Experiment 3 (intra-talker variation).** The x-axis indicates Vowel Duration (80–120 ms). The fast rate condition is indicated by the dashed line, neutral by the solid line, and slow by the dotted line. Colors indicate Group, with the high-rate group shown in dark grey and the low-rate group shown in light grey. Error bars represent the standard error of the mean.

3.5 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 PBS (e.g., Reinisch & Sjerps, 2013). Results indicated that listeners were more likely to categorize an ambiguous vowel midway 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 because of listeners tracking individual talkers' habitual speech rates (i.e., talker-specific), 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 (2016b), who investigated whether listeners tracked talkers' habitual rates in a conversation. After listening to a 2-min dialogue between two women 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's 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's (2016b) second experiment, the test phase involved categorization of ambiguous words embedded in rate-manipulated con-

text 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's 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. In addition, 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, whereas Baese-Berk et al. (2014) investigated a lexical effect,

the perception of function words. Pitt et al. (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 nonspeech 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 (2016b), 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 et al., 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), that may exist 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, downweighing 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 (see our Experiment 2; Reinisch, 2016b). 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.

4 | Listening to yourself is special: Evidence from global speech rate tracking[1]

Abstract

Listeners are known to use adjacent contextual speech rate in processing temporally ambiguous speech sounds. For instance, an ambiguous vowel between short /ɑ/ and long /a:/ in Dutch sounds relatively long (i.e., as /a:/) embedded in a fast precursor sentence, but short in a slow sentence. Besides the local speech rate, listeners also track talker-specific global speech rates. However, it is yet unclear whether other talkers' global rates are encoded with reference to a listener's self-produced rate. Three experiments addressed this question. In Experiment 1, one group of participants was instructed to speak fast, whereas another group had to speak slowly. The groups were compared on their perception of ambiguous /ɑ/-/a:/ vowels embedded in neutral rate speech from another talker. In Experiment 2, the same participants listened to playback of their own speech and again evaluated target vowels in neutral rate speech. Neither of these experiments provided support for the involvement of self-produced speech in perception of another talker's speech rate. Experiment 3 repeated Experiment 2 but with a new participant sample that was unfamiliar with the participants from Experiment 2. This experiment revealed fewer /a:/ responses in neutral speech in the group also listening to a fast rate, suggesting that neutral speech sounds slow in the presence of a fast talker and vice versa. Taken together, the findings show that self-produced speech is processed differently from speech produced by others. They carry implications for our understanding of rate-dependent speech perception in dialogue settings, suggesting that both perceptual and cognitive mechanisms are involved.

[1]Adapted from *Maslowski, M., Meyer, A. S., & Bosker, H. R. (2018). Listening to yourself is special: Evidence from global speech rate tracking. PLoS ONE, 13(9), e0203571.*

4.1 Introduction

Self takes a special role in the processing of cognitive and perceptual information. For instance, one's own face is recognized faster and more accurately than other familiar and unfamiliar faces (Keyes & Brady, 2010). Also, self-relevant stimuli, such as self-owned or self-associated items, attract more attention compared to stimuli associated with others (Cunningham, Turk, Macdonald, & Macrae, 2008; Turk et al., 2011; Sui, He, & Humphreys, 2012; Truong & Todd, 2017; Truong, Roberts, & Todd, 2017). Not only self-relevant, but also self-produced items aid processing: Stroke recognition in hand-writing is facilitated when strokes are self-produced (Knoblich & Flach, 2001; Knoblich, Seigerschmidt, Flach, & Prinz, 2002). Additionally, words were remembered better when participants read them aloud themselves during encoding than when they heard them read by others (Forrin & MacLeod, 2018). This advantage of self-produced words remains even when participants' own voices are recorded earlier and played back to them at test.

However, whether and how self-produced items influence perception of other-produced items is less well studied. The most common situation in which humans constantly switch between experiencing self-produced and other-produced input is dialogue. In dialogue, interlocutors easily alternate between speaking and listening, with turn gaps being remarkably short (~200 ms) (Stivers et al., 2009). Given that one's own speech often constitutes the context for an interlocutor's utterance, self-produced speech may affect the perception of the speech from another talker. The present study investigated how our own speech production, and specifically self-produced speech rate, affects how we process temporal cues in speech from another talker. The study replicates prior work on speech rate effects in global speech contexts and provides new empirical evidence that self-produced speech is processed differently from speech produced by others. As such, it contributes to our understanding of the representation of one's own voice in dialogue.

Temporal features of speech vary considerably with speech context. One reason for this is that acoustic cues map differently onto phonemic categories at different speech rates. Listeners must therefore normalize for contextual speech rate in order to interpret temporally ambiguous speech sounds (Bosker, 2017a; Diehl et al., 1980; Reinisch et al., 2011; Miller, 1981; Summerfield, 1981; Wade & Holt, 2005; Dilley & Pitt, 2010; Baese-Berk et al., 2014). That is, temporal cues in the ongoing acoustic signal are perceived relative to the surrounding

speech rate, such that the signal can be identified as a meaningful linguistic object (a segment, syllable, or word). Therefore, perception of speech sounds that mainly differ temporally, such as short and long vowels (e.g., German /a/ and /a:/; Reinisch, 2016b) or consonants (e.g., English /b/ and /w/; Miller & Baer, 1983) can shift from one phoneme to another based on contextual speech rate. For instance, an ambiguous vowel midway between Dutch /ɑ/ and /a:/ is biased towards short /ɑ/ in a slow speech context (Reinisch & Sjerps, 2013), as the adjacent speech rate makes the vowel sound relatively short (i.e., as /ɑ/ in *stad* [stɑt] "city"). Similarly, the same ambiguous vowel is biased towards long /a:/ in fast speech contexts, where it sounds relatively long (i.e., as /a:/ in *staat* [sta:t] "state"). This phenomenon is referred to as rate normalization and it is the process that we investigate here in relation to self-produced speech.

Temporal cues can be affected both by the local surrounding sentence context and more global speech contexts. Most studies so far have focused on local context effects; that is, effects of an adjacent sentence. Such local rate-dependent context effects have been argued to involve general auditory mechanisms. For instance, they have been shown to occur independently of talker voice changes (Bosker, 2017b; Newman & Sawusch, 2009), with a fast speech context from Talker A influencing subsequent perception of a target by Talker B. Moreover, the speech-like nature of the context seems to be inconsequential; both speech and non-speech induce local rate effects (Bosker, 2017a; Diehl & Walsh, 1989; Gordon, 1988). Context effects have furthermore been shown to hold even for 2–4 months old infants (Eimas & Miller, 1980) and non-human species (Dent, Brittan-Powell, Dooling, & Pierce, 1997). Lastly, effects of adjacent rate contexts are unaffected by attentional modulation, which supports the involvement of early perceptual processes (Bosker et al., 2017).

However, language users are also sensitive to talker-specific variation (Creel & Bregman, 2011; Eisner & McQueen, 2005). More global effects of speech rate (induced by cues from non-adjacent larger speech contexts and multiple talkers) seem to be sensitive to such higher-level influences such as talker voice. Maslowski et al. (2019a, see *Chapter 3*) investigated whether one Talker A's global rate is perceived relative to another Talker B's speech rate. In their experiments, two groups of participants listened to sentences spoken by a male and a female talker. In one experiment, examining effects of talker-specific global speech rate, a high-rate group listened to one Talker A speaking at a high speech rate and another Talker B speaking at a 'neutral' speech rate. A second group, the low-rate group, listened to the same neutral speech rate (Talker B), but to

Talker A speaking at a low speech rate. On each trial, participants categorized a word with a temporally ambiguous vowel between Dutch /ɑ/ and /a:/ (e.g., /tɑkjə, ta:kjə/, "twig"/"task") that was embedded in a trial sentence (e.g., *Toen Luuk mompelend iets tegen Lotte vertelde, hoorde Lotte "het takje/taakje" gezegd worden*, "When Luuk muttered something to Lotte, Lotte heard "the twig/task" being said"). The two participant groups were then compared on their perception of these vowels in sentences from Talker B speaking at a neutral rate. That is, whilst the local rate cues in Talker B's speech were identical for both groups, the global context (fast/slow speech from other talker) in which Talker B was heard differed between groups. The authors observed an effect of global speech rate; in the high-rate group listening to a fast Talker A, they observed significantly fewer /a:/ responses in neutral Talker B's speech, compared to the low-rate group with a slow Talker A (reproduced in Figure 4.1). This suggests that B sounded slow when A was faster, but fast when A was slower.

Interestingly, the effect disappeared in another experiment, where each talker spoke at two speech rates in separate trials. That is, the high-rate group heard fast and neutral rate speech from Talker A as well as from Talker B. Conversely, the low-rate group heard slow and neutral speech from Talker A and B. As a consequence, there was large intra-talker variability in speech rate, in contrast to the previous experiment where the intra-talker speech rates were held constant. However, the average rate across the two talkers in the two groups was identical in both experiments. The results of this experiment showed no difference between the two groups in the proportion of /a:/ responses in neutral speech. The authors interpreted their results from the two experiments together as listeners tracking talker-specific global rates rather than a talker-independent average rate. That is, with sufficient intra-talker regularity, cues to global speech rate are used in perception of another talker, with the global rates from different talkers being perceived relative to each other.

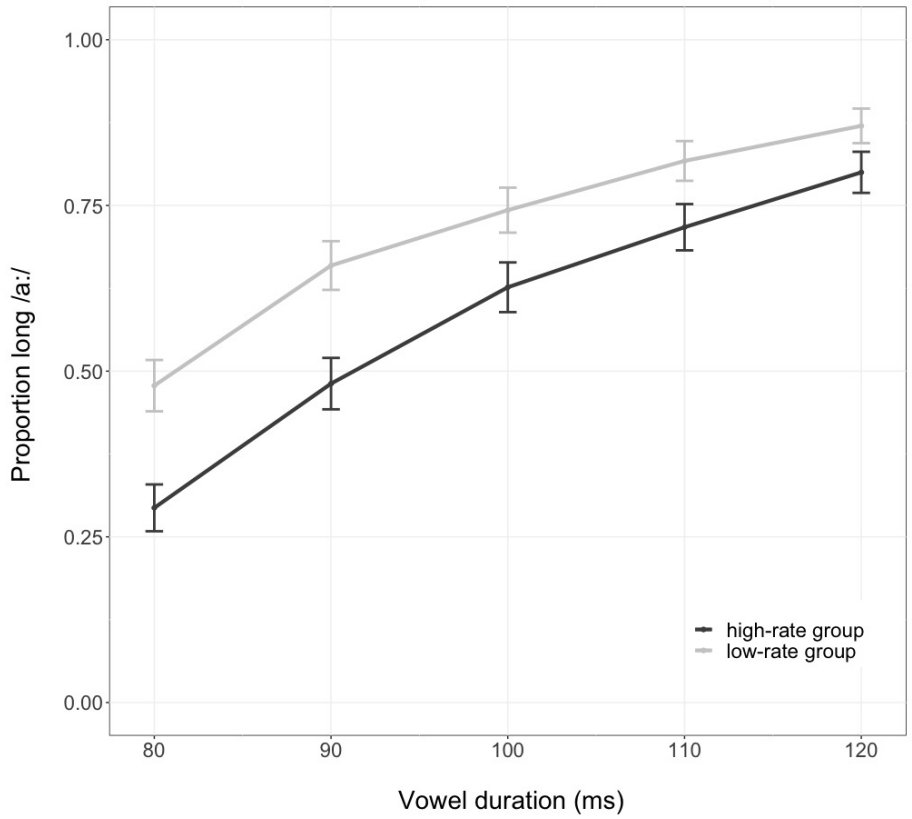

Figure 4.1: **Adapted reprint of only the neutral rate average categorization data of Maslowski et al.'s (2019) Experiment 2.** The X-axis indicates Vowel Duration (80–120 ms). Colors indicate Group, with the high-rate group shown in dark grey and the low-rate group shown in light grey. Error bars represent the 95% confidence intervals. Adapted from "How the tracking of habitual rate influences speech perception," by M. Maslowski, A. S. Meyer, and H. R. Bosker, 2019, Journal of Experimental Psychology: Learning, Memory, and Cognition, 45(1), 128–138.

A question arising from these experiments on local and global rate effects is how talkers' self-produced speech rates are represented in comparison to other talkers' rates. When producing speech, talkers also hear themselves speaking, with their own speech typically comprising the context for their interlocutor's speech. Therefore, a listener's self-produced speech rate may modulate perception of the global rate of another talker. Alternatively, global rate tracking may involve operations that factor out self-produced rate, as self-produced speech is not necessarily informative in disambiguation of speech from other talkers.

We already know that in local contexts, speech rate effects can be induced by self-produced speech. In a recent study by Bosker (2017b), participants were instructed to read out sentences at two pre-specified rates (fast and slow). Im-

mediately after participants had produced a sentence themselves, they heard an ambiguous target word (/ɑ/ vs. /a:/) produced by someone else. Bosker observed a difference in perception of target words between the condition in which participants were instructed to speak fast and the condition in which they had to speak slowly; more long /a:/ responses were observed when participants spoke fast. This suggests that self-produced speech can affect perception of speech produced by others. No speech rate effect was observed in another experiment testing effects of fast and slow covert speech (produced silently without articulation). That is, the effect of overt self-produced speech seems to be a consequence of self-monitoring of the external signal.

What is noteworthy is that Bosker's (2017b) study also included an experiment in which the recordings from the previous experiment on self-production were played back to the same participants (passive listening to one's own voice). Here, the difference between the fast and the slow condition was somewhat larger than in the first experiment. Bosker speculated that the smaller effect of self-production in the first experiment (relative to passive listening in the other experiment) may be due to speaking-induced suppression (SIS). SIS involves a reduction in the neural response to self-produced speech in auditory cortex (as compared to the neural response to perception without production, i.e., playback of one's own voice; Ventura, Nagarajan, & Houde, 2009; Houde, Nagarajan, Sekihara, & Merzenich, 2002). It may be that SIS attenuates the processing of self-produced speech rate during production compared with passively listening to self-produced speech.

Building on Bosker (2017b), the present study investigates how self-produced speech may affect *global* speech rate perception, using the design and materials from Maslowski et al. (2019a). Experiment 1 tested whether self-produced speech rate plays a part in perception of another talker's global rate. The experiment featured equal proportions of production and perception trials. One group of participants (high-rate group) was instructed to produce sentences at a fast speech rate (production trials) and to categorize words in 'neutral' rate sentences from another talker (perception trials). Another group (low-rate group) spoke at a slow rate (production trials) and listened to the same neutral speech from the other talker (perception trials). The two groups differed only in the rate at which they produced sentences in production trials. The production trials were mixed with perception trials from the other talker, which contained an ambiguous Dutch /ɑ/-/a:/ vowel embedded in minimal pairs that were only distinguishable by this vowel.

If listeners perceive the global speech rate of another talker relative to their self-produced speech rate, categorization responses should differ between the high-rate group and the low-rate group. The high-rate group should then report hearing fewer long /a:/ vowels than the low-rate group, because the neutral talker sounds relatively slow. Such a finding would mirror that of the global rate experiment in Maslowski et al. (2019a). However, if listeners base their perception of global speech rate only on other talkers' speech and do not rely on their own productions, no group difference should be observed in /ɑ/–/a:/ word categorization in neutral speech in this experiment.

To preview the findings, the results of Experiment 1 showed no group difference. This result could be a consequence of the production task itself, corresponding to the attenuated self-produced rate effects (relative to passively listening to oneself) in Bosker (2017b). Bosker speculated that the smaller effect of self-produced speech rate could be a consequence of SIS. If the null result of our Experiment 1 is indeed due to SIS, an effect of self-produced rate may emerge when no production task is involved. That is, listening to playback of self-produced speech may modulate perception of another talker.

Alternatively, self-awareness may lead listeners to factor out their own speech, whether they are listening to themselves during production or listening to themselves passively. Listeners would recognize their own voice when listening to playback of their own speech. Because self-produced speech rate is typically uninformative for the perception of others, listeners may ignore their own productions when interpreting speech from another talker. This account would predict no effect of passively listening to self-produced speech rates (i.e., no group effect).

To distinguish between these two accounts, in Experiment 2, the participants from the previous experiment were invited back to listen to their own speech recorded in Experiment 1. The experiment was identical to the first experiment, except now production trials were replaced with playback of the recordings of the same self-produced trials. If listening to oneself passively is different from listening to oneself whilst speaking (for instance as a result of SIS), the results of this experiment should deviate from those of Experiment 1 (i.e., a group difference, with the high-rate group reporting fewer long /a:/ responses). However, if it is not the input itself, but rather the fact that the input was self-produced that led to the lack of an effect in the previous experiment (because of self-awareness), the results of this experiment should parallel those of Experiment 1. To preview findings once more, no effect of passively listening to self-produced

speech was found in Experiment 2 (i.e., no difference between the two groups). This may suggest that the null result in Experiment 1 was not a consequence of SIS, but rather self-awareness.

An alternative interpretation of the null results found in Experiment 1 and 2 is that the results stem from variability within the fast and slow rates produced by the previous participants. Maslowski et al. (2019a) found that when intra-talker variability is increased, global rate effects disappear. Similarly, because the speech produced in fast and slow production trials in Experiment 1 naturally included some intra-talker variability (within limits), this may have eliminated the global rate effects of the previous experiments (i.e., compared to the highly controlled and artificially compressed and expanded fast and slow materials in Maslowski et al.'s Experiment 2).

If the null results observed in Experiments 1 and 2 were due to intra-talker variability within the fast and slow rates, no global rate effect should emerge when Experiment 2 is repeated with a different participant sample, who are unfamiliar with the voices from the participants from before. However, if the results in the preceding experiments were indeed due to self-awareness, this would predict an effect of global speech rate (as found in Maslowski et al., 2019a, with the same neutral rate materials) when presented to different participants. Therefore, in Experiment 3, Experiment 2 was repeated with a new participant sample, who did not know the participants from before. As such, each participant heard one talker speaking at a neutral rate in perception trials and passively listened to (fast or slow) production trials from one of the participants from Experiment 1. After each neutral rate trial, participants again evaluated an /ɑ/-/a:/ vowel in a target word. We predicted that the results of the experiment would replicate the findings in Maslowski et al.: Listening to a fast Talker A and a neutral Talker B, should make Talker B sound relatively slow, whereas listening to a slow Talker A should make Talker B sound relatively fast.

4.2 Experiment 1: Self-production

Experiment 1 addressed the question whether self-produced speech rate affects perception of other talkers in global speech contexts. On the one hand, this may not be the case, as tracking self-produced rate may not necessarily be useful for comprehension of other talkers. On the other hand, self-produced speech may affect perception of others in global speech contexts in the same way as the global speech rate of one talker, Talker A, influences perception of the speech

rate of another talker, Talker B (Maslowski et al., 2019a). Moreover, effects of self-produced speech have previously been found in local contexts (Bosker, 2017b), with one's own voice affecting subsequent perception of an immediately following target produced by another talker.

4.2.1 Methods

Participants. A sample of native Dutch female participants ($N = 41, M_{age} =$ 23, range = 19–33) with no hearing, visual, or reading deficits were recruited from the Max Planck Institute participant pool. All gave their informed consent to participation, 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, applying the same criterion as in our prior study (Maslowski et al., 2019a), from which the stimulus set was adopted. Data from 9 participants were excluded, either because they performed outside the aforementioned range ($n = 7$) or because of non-compliance (e.g., frequently talking to themselves during perception trials; $n = 2$).

Design and materials. Experimental materials consisted of the 'neutral rate' materials used in Maslowski et al. (2019a). These materials comprised eight 24-syllable sentences with one of two Dutch /ɑ/-/a:/ minimal pairs: *takje*/*taakje* (/tɑkjə, ta:kjə/, "twig"/"task") and *stad*/*staat* (/stɑt, sta:t/, "city"/"state") (e.g., *Terwijl Niels rustig zijn tijdschrift stond te lezen, hebben de heren eens "stad/staat" tegen hem gebruld*, "Whilst Niels was peacefully reading his magazine, the gentlemen roared "city/state" to him once"). None of these sentences contained other instances of the vowels /ɑ, a:/, nor did they bias either member of a minimal pair semantically. The sentences were recorded by a native Dutch male and a native Dutch female talker. All speech up to a target vowel was set to the mean duration of that interval across the two talkers, using the PSOLA algorithm as implemented in Praat (Boersma & Weenink, 2015). Similarly, all speech after vowel offset was matched across the two talkers.

In Dutch, the /ɑ, a:/ vowel contrast is acoustically differentiated both temporally and spectrally, with /ɑ/ being short with a relatively low F2 and /a:/ being long with a high F2 (Adank et al., 2004). To construct vowel duration continua, one clear long vowel /a:/ from each talker was extracted. The vowel duration continua were created by linear compression using PSOLA and ranged from 80 to 120 ms (in five steps of 10 ms). To make the vowels spectrally ambiguous, the

F1s and F2s from both talkers were computed and set to a fixed ambiguous value using Burg's LPC algorithm in Praat (male talker: F1 of 764 Hz and F2 of 1261 Hz; female talker: F1 of 728 Hz and F2 of 1327 Hz). For each sentence, target vowels were then concatenated with the intervals before and after the target vowel, resulting in a stimulus set of 80 unique stimuli (eight context phrases × five vowel durations × two talkers). For more details on stimulus construction, see Maslowski et al. (2019a).

Procedure. The experimental procedure consisted of production trials and perception trials. Participants were randomly divided into two groups (both $n = 16$), who were both presented with an equal number of perception trials and production trials. The perception trials were identical across groups. A perception trial involved listening to 'neutral rate' speech from one of the two talkers (male or female), after which a button press response was required to indicate which member of a minimal pair the participant had heard in the sentence. Talkers were counterbalanced across participants.

Each perception trial started with a fixation cross (for 330 ms) that was always replaced by a stimulus sentence shown on the screen in black on a white background at auditory onset. The target word in the stimulus sentence was replaced by a question mark (e.g., *Terwijl Niels rustig zijn tijdschrift stond te lezen, hebben de heren eens [?] tegen hem gebruld*). At sentence offset, this screen was replaced by a screen showing two response options (e.g., *stad* and *staat*). For the word shown on the left side of the screen, participants pressed "1" and for the word on the right they pressed "0", with the position of the response options on the screen being counterbalanced across participants. Participants had 4 seconds to respond by button press, before a missing response was recorded.

Crucially, the two groups differed on production trials, which were randomly intermixed with perception trials. A production trial involved reading out a sentence at a pre-specified speech rate. Participants in the high-rate group had to produce speech at a fast speech rate, whereas participants in the low-rate group produced speech at a slow articulation rate. These pre-specified speech rates were based on the durations of fast trials ($1/1.6 = 0.625$ × the durations of neutral trials) and slow trials (1.6 × the durations of neutral rate trials) in the experiments in Maslowski et al. (2019a). Participants were explicitly instructed to speak without pausing between words. The sentences participants were instructed to read out in the production trials were the same as those in the perception trials, except that the target words *takje*/*taakje* and *stad*/*staat* in

the production items were substituted by *tukje* (/tʏkjə/, “nap") and *stoet* (/stut/, “procession"), to prevent participants' own /ɑ, a:/ vowels from affecting the perception of /ɑ, a:/ in the perception trials.

Production trials were cued by showing the sentences in red. After 1200 ms, the sentence turned green, to prompt the participant to start speaking. During production trials, recordings were made of the participant's speech. The experimenter could hear the participant through headphones throughout the experiment. After 0.625 (high-rate group) or 1.6 (low-rate group) times the durations of the neutral rate stimuli, a beep was played to the experimenter (inaudible to the participant). When the participant had finished producing the sentence, the experimenter pressed a button to give feedback on the participant's rate as indicated by the beep (“Please try to speak somewhat faster"/“Please try to speak somewhat slower"/“Well done!").

Prior to the experiment, participants completed two separate practice blocks, one for each modality. In the listening practice, each of the eight different sentences with various instances of the vowel continua endpoints were presented once. In the production practice, all eight sentences (with the substituted target words) were presented once in a practice block, but this block was repeated until the participant successfully produced them at the pre-specified rate.

Stimulus presentation was controlled by Presentation software (v16.5; Neurobehavioral Systems, Albany, CA, USA). Stimuli were presented in five blocks of 80 trials. Each block consisted of a random mix of all unique auditory stimuli of one talker (perception trials: $n = 40$) and five instances of each individual production item (production trials: $n = 40$), resulting in 200 perception trials and 200 production trials in total. One session lasted for a duration of approximately 55 minutes in the high-rate group, and 70 minutes in the low-rate group. After the experiment, participants indicated that they could clearly hear their own voice, despite wearing headphones.

4.2.2 Results and discussion

Production. The participants' sentence durations on production trials were analyzed as a proxy for speech rate. Production trials were disregarded from analysis when they contained word errors, coughs, or pauses of $>$ 500 ms in the low-rate group and $>$ 200 ms in the high-rate group. In total, 18.6% of production trials were excluded in the low-rate group and 14.6% of trials in the high-rate group, mainly due to pauses and word errors in the low-rate group and the high-rate group, respectively. Figure 4.2 illustrates the mean duration of

production trials for each participant in the high-rate and low-rate groups. This figure shows a relatively small difference in speech rate within the two groups and a clear separation between groups, as confirmed by a paired-samples t-test ($t(30) = 18.503, p < 0.001, d = 12.53$) comparing the mean durations of production trials between the high-rate group ($M = 3122$ ms, $SD = 347$ ms) and the low-rate group ($M = 6344$ ms, $SD = 604$ ms). This verifies that participants complied with our instructions.

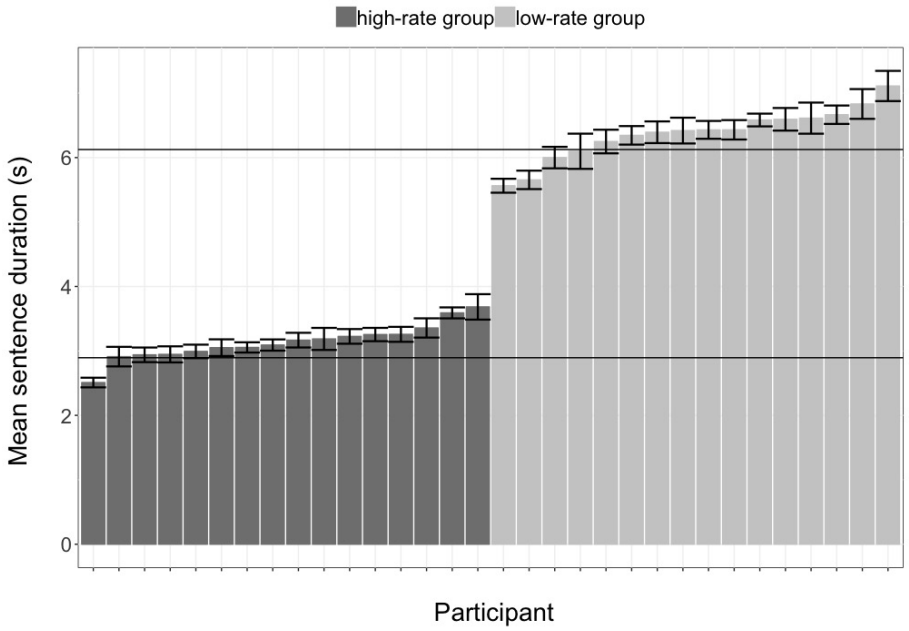

Figure 4.2: **Mean sentence durations of speech production trials of Experiment 1 (self-production).** On the X-axis, sentence durations are given for each participant in ordered sequence. Colors indicate Group, with the high-rate group shown in dark grey and the low-rate group in light grey. The horizontal lines indicate the intended sentence duration for the high-rate group (bottom) and the low-rate group (top). Error bars represent the 95% confidence intervals.

Perception. Figure 4.3 shows the categorization data on perception trials (proportion /a:/ responses) in Experiment 1. Participants reported a lower proportion of /a:/ when target vowels were at the shorter end of the duration continuum and a higher proportion when they were at the longer end of the continuum. Although Figure 4.3 seems to suggest that the two groups may differ slightly in their perception of vowels embedded in neutral speech in the opposite direction of our prediction, the following statistical analysis showed otherwise.

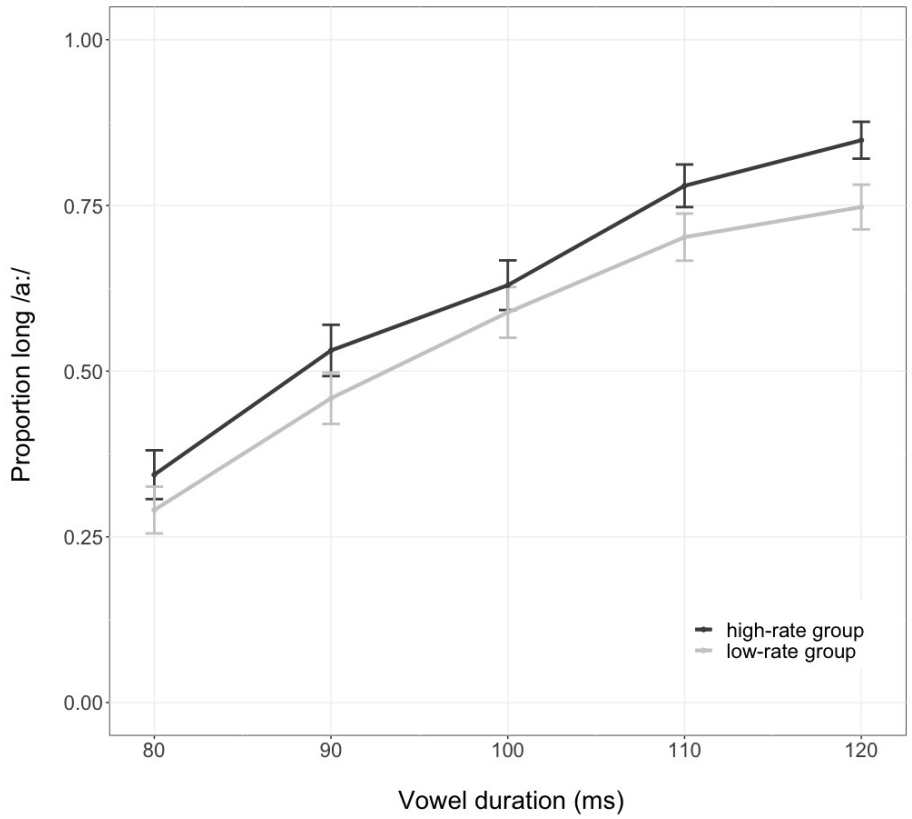

Figure 4.3: **Average categorization data of Experiment 1 (self-production).** The X-axis indicates Vowel Duration (80–120 ms). Colors indicate Group, with the high-rate group shown in dark grey and the low-rate group shown in light grey. Error bars represent the 95% confidence intervals.

The binomial responses to the perception trials (0.03% missing responses excluded) were quantified statistically with a Generalized Linear Mixed Model (GLMM) with a logistic linking function from the `lme4` package (Bates et al., 2015) in R (R Core Team, 2014). This GLMM tested whether there was a difference between the two groups in perception of ambiguous vowels embedded in neutral speech. Unless otherwise stated, the same full model was used in all three experiments and included the predictors Group (categorical; intercept is high-rate), 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). The included interactions between fixed factors were between Group and Vowel Duration, Group and Block, and Vowel Duration and Block. Random intercepts were included for Participants and Items with random slopes for all predictors by both random effects, except for the control variable Talker. Because the full model failed to reach conver-

gence for the current experiment, the random slopes for Vowel Duration and Block were dropped for both random effects.

Vowel Duration significantly affected the proportion of /a:/ responses ($\beta = 0.974, z = 20.649, p < 0.001$), with vowels of longer durations more often being reported as /a:/ than shorter vowels. The predictor Group did not reach significance ($\beta = -0.374, z = -1.316, p = 0.188$), providing no evidence for an effect of self-produced speech rate on the perception of another talker. The dependent variable proportion of /a:/ responses was also significantly affected by Block ($\beta = -0.156, z = -3.614, p < 0.001$), indicating that participants perceived decreasingly fewer /a:/ vowels as the experiment went on. The control variable Talker reached significance ($\beta = 0.889, z = 2.795, p = 0.005$), with an overall significantly higher proportion of /a:/ responses for the female talker. There were no significant interactions between fixed factors (Groups and Vowel Duration: $\beta = -0.036, z = -0.542, p = 0.586$; Groups and Block: $\beta = -0.023, z = -0.369, p = 0.712$; Vowel Duration and Block: $\beta = 0.006, z = 0.194, p = 0.846$).

The fact that no effect of self-production was observed (i.e., no group effect) could be because effects of one's self-produced speech rate are more short-lived than effects of others' global rates. Therefore, a more fine-grained analysis was performed on a subset of the data, consisting of only the perception trials directly following a production trial ($n = 3258$, 51.0%). However, no qualitative differences were observed compared to the results of the data in the full set.

The results of this experiment provide no evidence that target word perception in the perception trials was sensitive to participants' self-produced speech rates in production trials, suggesting that self-produced speech does not affect perception of another talker's global speech rate. This is in contrast to the results in Maslowski et al. (2019a). In their Experiment 2, participants evaluated the same neutral rate trials, but (instead of the present production trials) listened to another talker producing speech at a consistently fast/slow speech rate. They found an effect of global speech rate on the perception of another talker. However, replacing their fast and slow perception trials with the fast and slow production trials here seemed to remove the effect. This suggests that listening to oneself whilst talking has a different effect on perception than passively listening to another talker.

Additionally, the results of Experiment 1 differ from Bosker (2017b), who compared participants' vowel categorization immediately after having produced either fast or slow speech. In this very local self-produced speech context, Bosker

observed a difference between participants' perception of target words. Together, these studies suggest that self-produced speech induces a bias in local (adjacent) speech contexts, but not in global (distant) speech contexts.

4.3 Experiment 2: Playback self-production

Experiment 2 tested whether the lack of an effect of self-produced speech in Experiment 1 could be related to the production task itself. Auditory input from self-produced speech has been found to lead to reduced responses in auditory cortex (i.e., speaking-induced suppression; SIS), which may in turn have reduced the magnitude of a potential effect of self-produced global speech rate. As such, SIS could be argued to account for the lack of a shift in phonetic categorization in Experiment 1. Therefore, Experiment 2 repeated the experiment without a speech production task, by having the same participants listen to playback of their own speech (produced in Experiment 1).

4.3.1 Methods

Participants. The same sample of participants as in Experiment 1 was invited back to participate in Experiment 2. Out of the 32 participants from Experiment 1 who were included in the analyses, 22 returned for Experiment 2 (high-rate group: $n = 10$; low-rate group: $n = 12$). Group sizes were mildly unbalanced.

Design and materials. The same materials were used as in Experiment 1. This included all production trials (including word errors, coughs, and pauses) and all perception trials.

Procedure. The procedure of Experiment 2 was identical to Experiment 1, except now participants listened to playback of their own 200 sentence recordings from the previous experiment, instead of producing speech. As before, participants were only prompted to respond after neutral rate trials from the other talker. After playback of a self-produced speech trial, the next trial was presented directly. Participants were aware that half of the stimuli were self-produced. They listened to experimental stimuli and self-produced stimuli in the exact same order as presented and recorded in Experiment 1.

4.3.2 Results and discussion

Figure 4.4 summarizes the categorization data of Experiment 2. The figure indicates that participants reported hearing a higher proportion of /a:/ for targets with longer vowel durations. The overlap of the two lines suggests that there is no difference between the categorization data of the two groups.

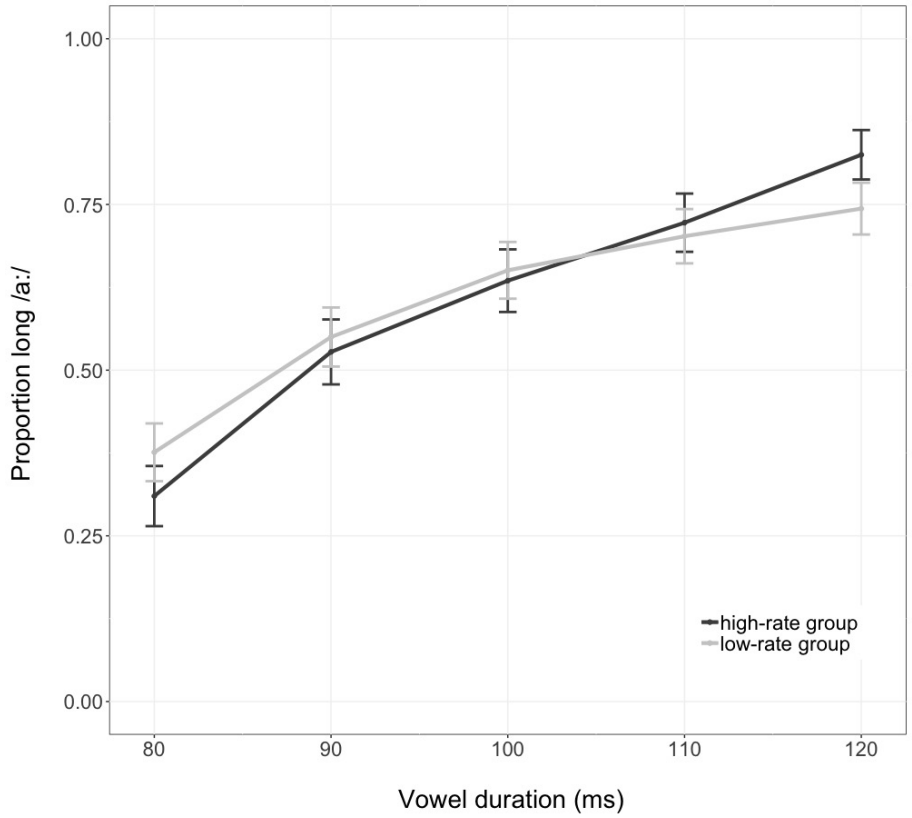

Figure 4.4: **Average categorization data of Experiment 2 (playback self-production).** The X-axis indicates Vowel Duration (80–120 ms). Colors indicate Group, with the high-rate group shown in dark grey and the low-rate group shown in light grey. Error bars represent the 95% confidence intervals.

The categorization data of Experiment 2 were tested with the full GLMM as described in Experiment 1, except that the random slopes for Block were dropped due to convergence issues. Vowel Duration significantly affected the proportion of /a:/ responses ($\beta = 1.261, z = 4.699, p < 0.001$), with participants more often reporting hearing /a:/ for longer vowel durations. Group had no significant influence on /a:/ categorization ($\beta = 0.096, z = 0.199, p = 0.843$), suggesting that the likelihood of hearing /a:/ in neutral speech was the same for both

groups. Block did not reach significance ($\beta = -0.043, z = -0.745, p = 0.456$), showing that performance did not change over time. However, the interaction between Vowel Duration and Block was significant ($\beta = 0.104, z = 2.511, p = 0.012$), with the difference in /a:/ categorization between the two endpoints of the duration continuum being larger in later blocks than earlier ones. Talker also significantly affected categorization ($\beta = 1.229, z = 2.340, p = 0.019$), with a higher proportion of /a:/ for the female talker. Finally, the interactions between Group and Vowel Duration ($\beta = -0.175, z = -0.555, p = 0.579$) and Group and Block ($\beta = 0.060, z = 0.755, p = 0.450$) were not significant.

Experiment 2 tested whether listening to playback of self-produced fast or slow speech induces variation in speech perception of another talker (speaking at a neutral speech rate). No effect of listening to one's own speech was found on the perception of another talker. This suggests that the lack of an effect of self-produced speech in Experiment 1 was not due to speaking and listening at the same time. Moreover, the result contrasts with Maslowski et al. (2019a), who found effects of talker-specific global speech rate. Therefore, this finding suggests that listening to oneself is intrinsically different from listening to other talkers.

4.4 Experiment 3: Unfamiliar listeners

Experiment 3 aimed to evaluate whether the null results from the previous experiments were related to participants hearing themselves (rather than another talker). Therefore, Experiment 2 was repeated but with a new sample of participants, who listened to the fast or slow sentence recordings made in Experiment 1 and evaluated neutral rate sentences.

4.4.1 Methods

Participants. Native Dutch female participants ($N = 40, M_{age} = 22$, range = 19–27) were recruited and divided into a high-rate group and a low-rate group. All gave their consent to participation. Data from eight participants were excluded, because their responses were outside the performance criterion described in Experiment 1, resulting in two pseudo-random groups of 16 participants each.

Design and materials. The same materials were used as in Experiment 2, including the self-produced trials from the participants from Experiment 1.

Procedure. The procedure was identical to that of Experiment 2, with the only difference between Experiment 2 and 3 being that, in Experiment 3, participants listened to speech from (to them) unfamiliar talkers.

4.4.2 Results and discussion

Figure 4.5 presents the categorization data of Experiment 3. The figure shows that participants reported higher proportions of long /a:/ for target vowels with longer durations. The difference between the two lines suggests that participants in the high-rate group reported hearing fewer long vowels than the low-rate group.

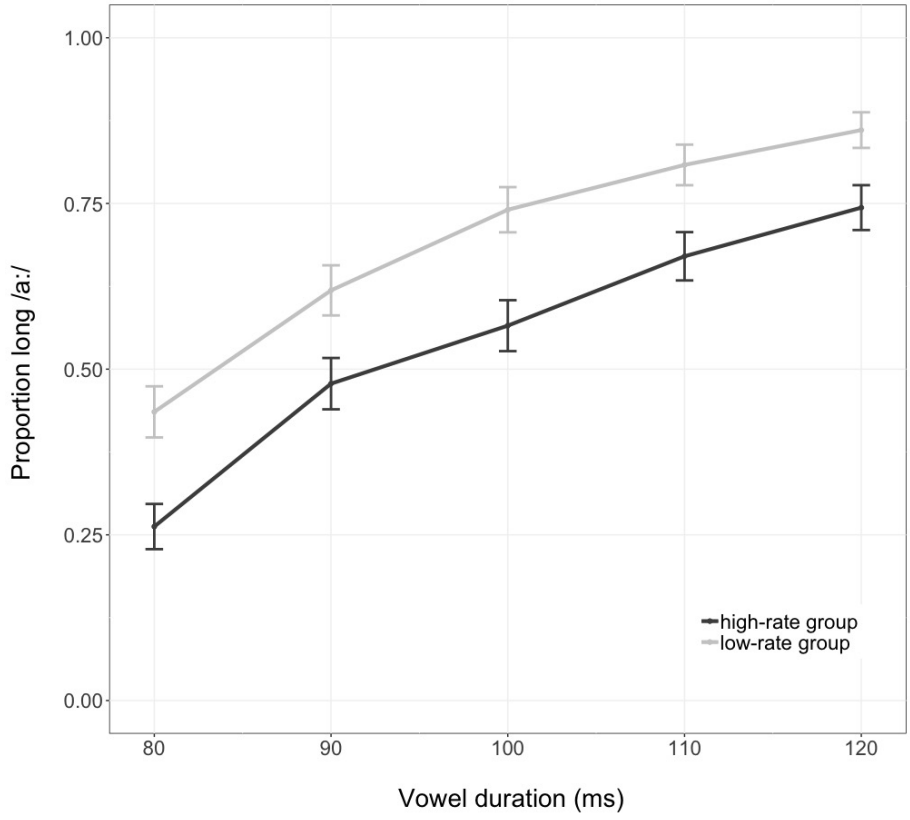

Figure 4.5: **Average categorization data of Experiment 3 (unfamiliar listeners).** The X-axis indicates Vowel Duration (80–120 ms). Colors indicate Group, with the high-rate group shown in dark grey and the low-rate group shown in light grey. Error bars represent the 95% confidence intervals.

The full GLMM as described in Experiment 1 tested the categorization data of Experiment 3. The model revealed a significant effect of Vowel Duration ($\beta = 1.195, z = 5.298, p < 0.001$), with the proportion of /a:/ responses increasing for longer vowel durations. Moreover, a significant effect of Group was observed between the high-rate group and the low-rate group ($\beta = 1.064, z = 2.895, p = 0.004$), with the high-rate group (who listened to fast and neutral speech) reporting a lower proportion of /a:/ than the low-rate group (who listened to slow and neutral speech). The model revealed no significant effects of Block ($\beta = -0.163, z = -1.469, p = 0.142$) and the control variable Talker ($\beta = 0.196, z = 0.641, p = 0.522$) on vowel categorization. None of the interactions between predictors reached significance (Groups and Vowel Duration: $\beta = 0.135, z = 0.469, p = 0.639$; Groups and Block: $\beta = -0.173, z = -1.234, p = 0.217$; Vowel Duration and Block: $\beta = 0.049, z = 1.268, p = 0.205$).

We performed an omnibus analysis on the combined data from all three experiments, to test whether the group effects in each experiment were significantly different from each other. A GLMM was run, comprising Group (sum-to-zero coded: slow coded as 0.5, fast as −0.5), Experiment (dummy coded, with Experiment 3 mapped onto the intercept), Vowel Duration, Block, and Talker, as well as the interaction between Group and Experiment. This model revealed two significant interactions. First, the interaction between Group and the contrast between Experiments 1 and 3 was significant ($\beta = -1.262, z = -14.72, p < 0.001$), demonstrating that the contrast between Groups in Experiment 1 was significantly different from the contrast in Experiment 3. Similarly, the interaction between Group and the contrast between Experiments 2 and 3 was significant ($\beta = -1.014, z = -10.19, p < 0.001$).

The results of Experiment 3 show that the global speech rate of an unfamiliar talker affects perception of temporally ambiguous vowels in another talker's speech; neutral rate speech sounds slow in the presence of a faster talker and vice versa. The results replicate the finding in Maslowski et al. (2019a) that the global rate of one talker is perceived relative to the global rate of another talker. Furthermore, the results of Experiment 3 indicate that the results obtained in Experiment 1 and 2, in which no differences between groups were found, were due to recognition of one's own voice. Finally, the results indicate that global rate effects are resilient to small variations in speech rate. In contrast to the (artificially compressed/expanded) fast and slow speech in Maslowski et al., the speech recorded in Experiment 1 was natural speech, exhibiting slight variabil-

ity in speech rate both within and between sentences. Therefore, Experiment 3 demonstrates that listeners can rely on roughly stable global speech rates.

4.5 General discussion

This study investigated the involvement of self-produced speech in perception of another talker's global speech rate. In each of the three experiments, two groups of participants listened to and evaluated neutral speech rate trials from another talker. In Experiment 1, these perception trials were interspersed with production trials in which a high-rate group was instructed to produce speech at a (pre-specified) fast rate, whereas a low-rate group was instructed to produce speech at a slow speech rate. We measured the difference in perception of the Dutch /ɑ, a:/ vowel contrast in the neutral rate speech from the other talker between the two groups. The results indicated that self-produced speech did not influence rate perception of the other talker (i.e., there was no difference between groups).

Because Experiment 1 could not exclude the possibility that speaking-induced suppression (SIS: reduced auditory response to self-produced speech) veiled a potential effect of self-produced speech, we performed another experiment to test this account. Experiment 2 was identical to Experiment 1, but this time, the participants from the first experiment listened to playback of their own speech, whilst evaluating target vowels in neutral rate speech as before. Again, no group difference was found on perception of neutral rate speech from another talker, indicating that the absence of a group effect in Experiment 1 was not a by-product of SIS.

Experiment 3 was conducted to confirm that the absence of global rate effects in the preceding two experiments was due to participants listening to themselves. In Experiment 3, a new participant sample performed the task of the second experiment. As such, the participants listened to two unfamiliar talkers, one of which was a prior participant. Here, the global speech rate effect previously observed in Maslowski et al. (2019a) was replicated; neutral rate sounded slow in the context of a faster talker (as evidenced by a lower proportion of long /a:/ responses), but fast in the context of a slower talker (higher proportion /a:/). Moreover, this global rate effect emerged in naturally produced speech contexts, showing for the first time that this effect is robust against small within-talker rate variability.

The results of the experiments provide valuable clues to which aspects of our own productions play a role in perception. In the literature, talkers have been suggested to be aware of sub-phonemic details in their own speech. For instance, talkers make online corrections when receiving altered auditory feedback during a speech production task (Houde & Nagarajan, 2011; Niziolek, Nagarajan, & Houde, 2013). Moreover, perception can be facilitated when stimuli are self-produced in L1 speech (Schuerman, Meyer, & McQueen, 2015), L2 speech (Eger & Reinisch, 2019; Sheldon & Strange, 1982), and in action perception (Knoblich & Flach, 2001; Knoblich et al., 2002). Studies of the neurobiological correlates of speech perception have argued that processing advantages for self-produced speech are due to a self-awareness network involving mirror-like systems (Jardri et al., 2007; Treille, Vilain, Kandel, Schwartz, & Sato, 2015). Interestingly, self-benefit seems to occur only when listeners recognize their own voice (Schuerman et al., 2015; Schuerman, 2017). These findings suggest that listeners must be very sensitive to their own voice, which is processed differently from other talkers' voices (Guenther, 2006).

Our experiments support the idea that one's own voice must somehow be marked in comparison to other talkers' voices. The lack of effects of self-produced speech in Experiment 1 and 2 may consequently be a result of participants strategically ignoring their own productions, regardless of whether they were listening to themselves whilst speaking or during passive listening. This may be due to reduced attention when listening to one's own voice (Graux et al., 2013). The findings are in line with a study on explicit judgments of speech rate by Koreman (2006). Koreman found no systematic differences in speech rate perception of others as a function of a listener's own habitual articulation rate or (clear vs. sloppy) speaking style (but see Schwab, 2011), suggesting that talkers disregard feedback from their own speech rate when listening to others.

Another possible interpretation of the absence of context effects in Experiments 1 and 2 is that the global self-produced speech rates produced by the participants were not their habitual rates. The participants listening to self-produced speech had an enormous amount of prior experience with their own habitual rates, and, consequently, the artificial and imposed rates at which they had to speak in the first experiment may have had little impact on the perception of another talker's rate in both experiments.

Both interpretations of the results support the involvement of self-awareness in perception of self-produced speech. From the data presented here, we cannot distinguish specifically between awareness of hearing oneself and awareness of

the self-produced speech rate in the lab not being representative of one's habitual rate. Therefore, manipulating awareness could be an interesting avenue for future research to enhance our understanding of the involvement of self-awareness in perception of self-produced speech. However, given the other facets of cognition in which self representations seem to be different from representations of others, we argue that it is more plausible that the effects were due to self rather than to familiarity with one's own habitual rate as a consequence of greater exposure.

Both of these accounts are consistent with episodic models of speech perception. In episodic models, word recognition is shaped by distributional properties coming from detailed representations (i.e., exemplars) of every instance of a word in the input (Goldinger, 1998; Bybee, 2006). If our results stem from extensive prior exposure to one's own production, this experience could have led to richer and more robust representations of their own voices (Xu, Homae, Hashimoto, & Hagiwara, 2013). This would restrict another talker's speech to be perceived relative to one's own new tokens produced at an extraordinary and artificially imposed speech rate. However, exemplars are also assumed to be labeled for various indexical features, such as talker voice (Pierrehumbert, 2001). If our findings are a result of self-produced speech being encoded as such (i.e., with a talker-specific label for 'self'), one's own voice may consequently be ignored in perception of others.

The lack of a context effect in Experiment 1 is particularly interesting in relation to findings by Bosker (2017b). Bosker compared participants' vowel categorization immediately after having produced either fast or slow speech. As soon as participants had produced a sentence, they would hear an ambiguous target word from another talker. In this very local self-produced speech context, Bosker observed a difference in participants' perception of target words, suggesting that a talker's own speech can modulate perception of another talker when immediately preceding an ambiguous word. Experiment 1 shows, however, that in larger contexts, incoming speech is not necessarily encoded with reference to representations reflective of listeners' own productions.

Bosker (2017b) also found an enhanced effect of self-produced speech when participants listened to playback of their own speech relative to a production experiment, which he speculated may have been a consequence of SIS in the production experiment. The difference between the local rate-dependent effects in Bosker (2017b) and the null effect in our Experiment 2 suggests that local and global speech rate normalization involve different mechanisms.

Local and global rate-dependent context effects may be interpreted with reference to Bosker et al.'s (2017) two-stage model of normalization processes in speech perception. This two-stage model includes a first stage that is related to early perceptual adjustments, involving online low-level processing of temporal and spectral information in the signal. This first stage includes effects of local surrounding contexts, which are obligatory (Bosker & Ghitza, 2018), happen prelexically, are independent of talker changes (Bosker, 2017b; Newman & Sawusch, 2009), not specific to speech contexts (Bosker, 2017a; Diehl & Walsh, 1989; Gordon, 1988; Sjerps, Mitterer, & McQueen, 2011; Wade & Holt, 2005), and continue to exist under cognitive load (Bosker et al., 2017). Because perceptual normalization is automatic, self-produced speech rate – either actively produced or passively heard (Bosker, 2017b) – in *local* contexts directly modulates perception of others.

The second stage involves domain-specific cognitive adjustments performed later in time (rather than perceptual normalization), after talker segregation. Here, word recognition may be modulated by comparing the speech input to an expected form considering a certain speech context or talker (Bosker & Reinisch, 2015, 2017; Reinisch, 2016b). Therefore, talker-specific global speech rate effects, such as the effect reported in Maslowski et al. (2019a), likely occur during the second stage. Importantly, feedback from one's own present speech rate is disregarded in global speech rate normalization; regardless of whether listeners hear themselves actively or passively, they ignore their own speech in perception of another talker. Whether talkers' own habitual rates play a part in the online processing of others' speech remains to be determined. Yet, such an influence is argued to be unlikely, since adjustments based on one's own speech would not facilitate perception of other talkers. That is, tracking self-produced speech rate presumably does not facilitate comprehension of others' speech, whereas tracking other talkers' speech rates may help perception in the long term.

The current study shows that listening to one's own voice is special: Self-produced speech is processed differently from speech produced by others (Experiment 1). This seems to be task independent, as playback of one's own speech also does not elicit an effect of self-produced rate on perception of others (Experiment 2). Furthermore, this study shows that global rate effects can be replicated with naturally produced speech (Experiment 3). Importantly, this indicates that some amount of within-talker variability in speech rate is allowed before global rate tracking fails. These findings shed further light on the complex mechanisms of speech perception in dialogue settings, highlighting the hierarchical processes

involved in rate normalization, as suggested by the two-stage model in Bosker (2017b). To further empirically test this model, future work may investigate the time course of global rate effects to explore timing differences between local and global rate normalization.

5 | The time course of speech rate normalization depends on the distance of the context[1]

Abstract

To comprehend speech sounds, listeners tune in to speech rate information in the proximal (immediately adjacent sounds), distal (non-adjacent sentence context), and global context (further removed preceding and following sentences). Global contextual speech rate cues have been shown to have constraints not found for proximal and distal speech rate. Therefore, listeners may process such cues at distinct time points during word recognition. We conducted a printed-word eye-tracking experiment to compare the time courses of distal and global context effects. Results indicated that the distal rate effect emerged immediately after target sound presentation, in line with a general-auditory account. The global rate effect, however, arose more than 250 ms later than the distal rate effect, indicating that distal and global context effects involve distinct processing mechanisms. Results are interpreted in a two-stage model of acoustic context effects. This model posits that distal context effects involve very early, possibly domain-general, perceptual processes, while global context effects arise at a later stage, involving cognitive adjustments conditioned by higher-level information.

[1] Adapted from *Maslowski, M., Meyer, A. S. & Bosker, H. R. (under review). The time course of speech rate normalization depends on the distance of the context.*

5.1 Introduction

When humans listen to speech, they pick up on many different acoustic cues that contribute to comprehension of the signal. Not only signal-intrinsic cues (e.g., pitch; vowel length) are utilized; listeners also pay attention to context-specific (signal-extrinsic) indexical properties in making sense of an acoustic signal. Those include, for instance, voice characteristics of the interlocutor (Creel & Bregman, 2011; Eisner & McQueen, 2005; Pufahl & Samuel, 2014) and environmental noise (Cooper et al., 2015; Creel et al., 2012; Pufahl & Samuel, 2014). Such context-specific contributors to comprehension might influence perception at different time points during perceptual processing. In this study, we tested when listeners use different contextual speech rate cues in word recognition.

Listeners track and adapt to temporal information in speech. That is, they use the speech rate context to tune their perception of temporally ambiguous stretches of speech, such as short and long vowels (Bosker, 2017a), consonants (Miller & Baer, 1983), and words (Baese-Berk et al., 2019; Dilley & Pitt, 2010). As such, listeners take into account the surrounding speech rate. Interestingly, listeners have been shown to be sensitive to speech rate in at least three types of context: proximal, distal, and global contexts. The proximal context is defined as the context directly preceding and following an ambiguous stretch of speech to a distance of approximately 250 to 300 ms (Newman & Sawusch, 1996; Reinisch et al., 2011; Sawusch & Newman, 2000; Summerfield, 1981). For instance, Diehl and Walsh (1989) showed that the phonetic category boundary between /b/ (short VOT) and /p/ (long VOT) can be shifted from one phoneme to another by altering the duration of the following vowel; reduction of the vowel in /ba/ led to a bias towards hearing /pa/.

The distal context is the sentence context beyond the proximal context in both directions, typically the surrounding sentence context (cf. Reinisch et al., 2011). That is, while proximal context is controlled, listeners are more likely to hear an ambiguous Dutch /ɑ–a:/ vowel as short /ɑ/ when the distal context is slow (Reinisch & Sjerps, 2013). Conversely, listeners tend to perceive the same ambiguous vowel as long /a:/ when the sentence is fast.

Global speech rate information comes from longer contexts (up to an hour of speech; Baese-Berk et al., 2014) and multiple talkers, where one talker affects perception of another talker (Maslowski et al., 2018, 2019a). Maslowski et al. (2019a, see *Chapter 3*) investigated inter-talker effects of global speech rate on perception of the Dutch vowel contrast between /ɑ/ and /a:/. Two participant

groups listened to sentences spoken by two different talkers. In the high-rate group, Talker A always spoke at a neutral speech rate, whereas Talker B had a fast speech rate. In the low-rate group, Talker A's speech rate was again neutral, but Talker B spoke at a slow speech rate. Maslowski et al. found a contrastive effect of global speech rate; if Talker B was fast, neutral Talker A sounded slow, but if Talker B was slow, neutral Talker A sounded fast. This was evidenced by more long /a:/ responses for neutral Talker A in the low-rate group than in the high-rate group. The experimental results were replicated in Maslowski et al. (2018, see *Chapter 4*) with naturally produced fast and slow speech with similar results.

In the current study, we focused on the time course of distal and global context effects of surrounding speech rate. Speech rate cues in the distal context have been suggested to affect the perception of target speech sounds immediately. For instance, Reinisch and Sjerps (2013) investigated the timing of the integration of contextual temporal and spectral cues in a printed-word visual world paradigm. Participants listened to context sentences that were manipulated either temporally (fast vs. slow) or spectrally (high F2 vs. low F2). In these sentences, the proximal context was controlled; that is, each sentence included a fixed 300 ms silent interval preceding and following the target. The authors measured participants' fixations to written words on a screen, to test how the context sentences would affect perception of a following target word with the Dutch /ɑ, a:/ vowel contrast. They found that both spectral and durational cues immediately influenced perception of the target vowel. The effect of fast versus slow distal speech contexts on participants' eye fixations could already be picked up around 300 to 400 ms after vowel onset. Reinisch and Sjerps argued that distal contextual influences happened at a very early stage of processing.

Toscano and McMurray (2015) also investigated how listeners coped with variability in speech rate, testing the influence of the sentence rate context on VOT, using a visual world paradigm with visual stimuli representing minimal word pairs such as *beach*/*peach*. They found that speech rate cues immediately modulated the uptake of VOT, as soon as the information was available. Toscano and McMurray's speech rate effect arose approximately between 300 and 400 ms after target word onset, in corroboration with the distal speech rate effect in Reinisch and Sjerps (2013). However, note that Toscano and McMurray manipulated both the proximal (adjacent) and the distal context (further removed sentential context) simultaneously and that they looked at target word onset rather than vowel onset.

Recently, a third eye-tracking study tested effects of speech rate, this time on morphosyntactic gender marking in Dutch (Kaufeld et al., 2019). Just as Reinisch and Sjerps (2013) and Toscano and McMurray (2015), Kaufeld et al. found effects of speech rate normalization in an early time window after target vowel offset (i.e., 250 ms; ca 350 ms after vowel onset).

The fact that distal rate effects arise very early in perception has been taken as evidence for the involvement of general auditory (i.e., domain general) mechanisms (Reinisch & Sjerps, 2013; Toscano & McMurray, 2015). More evidence for distal speech rate normalization involving general auditory mechanisms comes from findings that listeners use the context in perception of a target even when the context is produced by a different talker (Newman & Sawusch, 2009). Even the fast or slow speech rate of one's own voice can change how following other-produced speech sounds are perceived (Bosker, 2017b). Distal speech rate effects have also been found to be insensitive to cognitive load manipulations, thus being unaffected by attentional modulation (Bosker et al., 2017). Moreover, distal rate effects are also induced by non-speech auditory contexts (Bosker, 2017a; Gordon, 1988; Wade & Holt, 2005) and are not task-driven, taking place even without explicit attention being drawn to the ambiguous target word (Maslowski, Meyer, & Bosker, 2019b, see *Chapter 2*). These findings all argue for distal rate effects to involve domain-general mechanisms that happen very early on in speech perception.

Different prerequisites have been found for global speech rate effects. Global speech rate tracking is subject to constraints that have not been found for distal speech rate. Firstly, global rate tracking is talker-specific, whereas distal rate tracking is talker-independent. Maslowski et al. (2019a) conducted an experiment providing evidence for this. In this experiment, two groups of participants listened to two talkers (A and B) each speaking at two rates (fast and neutral in the high-rate group; slow and neutral in the low-rate group). With considerable rate variation within each talker's speech, no global rate effect was found on perception of the /ɑ, a:/ vowel contrast. The authors interpreted this as the global rate effect being driven by talker-consistent habitual speech rates. That is, global speech rate effects are observed only when talkers show distinct habitual speech rates and global rate tracking fails with a reasonable amount of intra-talker variation.

Another difference between the distal and global rate effect is that the global speech rate effect is easily overriden by local variation. Reinisch (2016b) conducted an experiment in which participants were exposed to speech from two

female talkers, one of whom spoke fast and the other slowly. At test, Reinisch observed an effect of habitual speech rate when participants categorized isolated words with temporally ambiguous vowels. That is, target words from the fast talker were more often categorized as long vowel words than target words from the slow talker. However, this global habitual rate effect disappeared in the subsequent experiment, in which the same manipulated vowels were embedded in fast and slow context sentences. Thus, listeners used habitual rate as a cue when no other rate information was available, but this effect was overriden by the fast and slow distal context rates.

A third argument for the global speech rate effect being different from the distal rate effect is that the global rate effect is not induced by one's own voice. Bosker (2017b) showed that one's own distal speech rate can affect perception of a following ambiguous target word spoken by another talker. However, one's own speech rate does not affect perception of another talker's speech rate in larger contexts. Maslowski et al. (2018) recruited two groups of participants, one of which was instructed to speak fast and the other to speak slowly. The two groups were compared on their perception of /ɑ, a:/ words embedded in a neutral Talker A's speech. They found no global rate effects for self-produced global contexts (Experiment 1). Even playback of one's own voice (i.e., passive listening) did not induce a global speech rate effect (Experiment 2). Only when participants listened to speech that was not self-produced (Experiment 3), an effect of global rate was found.

Therefore, global rate effects are constrained by higher-level information such as a talker's habitual rate. Listeners disregard their own speech rate and unreliable habitual rates of others when taking global context into account. Such constraints do not apply to effects of distal speech rate information. Therefore, distal and global speech rate processing may involve distinct processing mechanisms.

Bosker et al. (2017) proposed that speech rate normalization takes place at two hierarchical stages in a normalization framework of acoustic context effects, such as spectral and rate normalization. The first stage involves early and automatic perceptual auditory processes. Since distal speech rate effects are impervious to talker changes, attentional modulation, and the speech/non-speech nature of the sound context, distal rate normalization happens at this early and automatic stage. The second stage involves cognitive adjustments that take place later. These adjustments are conditioned by signal-extrinsic and indexical higher-level information, such as the identity of the talker (Maslowski

et al., 2018; Reinisch, 2016b), the habitual speech rate of the talker (Maslowski et al., 2019a; Reinisch, 2016b), the language that is spoken (Bosker & Reinisch, 2017), the speech register (Reinisch, 2016a), and situation-specific expectations (Bosker et al., 2017).

Considering that global rate effects are sensitive to talker identity and stable habitual rates, this entails, according to the two-stage model by Bosker et al. (2017), that global rate effects arise at the second stage, involving cognitive adjustments, while distal rate effects arise at the first stage, involving perceptual normalization. This might be evident in the time courses of both effects. The two-stage model predicts that distal and global speech rate effects happen in distinct time windows, with global rate influencing perception later in time than distal rate. The time course of the global rate effect has never been assessed directly, nor has it been compared to that of the distal rate effect. Here, we investigated the time courses of both the distal and global rate effects using an eye-tracking paradigm.

The present experiment mimicked Maslowski et al.'s (2019a) categorization experiment on inter-talker variation (i.e., Talker A speaking at one speech rate and Talker B at another). The experimental design and materials were adopted from Maslowski et al. (2019a). The current experiment goes beyond that experiment through the addition of measures of eye fixations, enabling us to investigate the time course of global and distal speech rate effects by analyzing when participants looked at an orthographic target word (cf. Reinisch & Sjerps, 2013). Specifically, a high-rate group listened to neutral speech rate sentences from Talker A and fast sentences from Talker B (i.e., the average rate across talkers was high), while a low-rate group listened to neutral Talker A, but to Talker B speaking at a slow rate (i.e., the average speech rate was low). Their task was to categorize /ɑ, a:/ words embedded in these rate-manipulated sentences (with fixed-rate proximal contexts). During sound presentation, the participant's eye movements and fixations on the two members of a minimal pair were recorded. If global rate effects arise later than distal rate effects, this should become apparent in the participant's eye-tracking data. Alternatively, distal and global processing arise simultaneously.

Concretely, we predicted that *within groups* the relatively faster rates would induce more long /a:/ responses than the relatively slower rates: In the high-rate group, fast speech should induce more long /a:/ responses than neutral rate speech, and in the low-rate group, neutral rate speech should induce more long /a:/ responses than slow speech. This within-groups distal rate effect should be

reflected in more looks to the word with long /a:/ in the relatively faster rates within the two groups. Moreover, based on Reinisch and Sjerps (2013), Toscano and McMurray (2015), and Kaufeld et al. (2019) we predicted that the distal rate effect should arise very rapidly after vowel offset, which is the earliest moment that participants have access to vowel duration. Additionally, we predicted that *across groups* a difference in looking patterns would arise in the neutral rate condition: Participants in the low-rate group should show more looks to long targets compared to participants in the high-rate group. This is a global rate effect. Following the two-stage model by Bosker et al. (2017), this global rate effect was predicted to arise only after the distal effect because it involves more higher-level cognitive adjustments.

5.2 Method

Participants. 42 native Dutch participants (female = 33, M_{age} = 23 years, range = 18–28 years) with normal hearing and vision were recruited from the Max Planck Institute participant pool. All participants gave informed consent to participation. Ethical approval of the study was provided by the Ethics Committee of the Social Sciences faculty of Radboud University (project code: ECSW2014-1003-196). Just as in previous experiments using the same stimuli (Maslowski et al., 2019a, 2018), it was decided a priori to exclude participants for whom the stimuli were insufficiently ambiguous, that is, when they categorized all vowels as being from the same category more than 90% of the time. Eight participants had to be excluded based on this criterion (high-rate group = 5; all eight participants showed > 90% long /a:/ responses). Two other participants were excluded because of technical difficulties. This resulted in two groups of 16 participants each: a high-rate group (female = 12, M_{age} = 23, range = 20–28), who heard fast and neutral speech rates, and a low-rate group (female = 13, M_{age} = 23, range = 20–28), who heard slow and neutral speech rates. With a sample size of 32 participants, we had a power of .95 to observe a global rate effect of > 80% of the size of the global rate effect obtained in Maslowski et al. (2019a) using the same stimuli (see Brehm & Goldrick, 2017, for simulating sample data to estimate power).

Design and materials. The spoken stimuli were taken from Maslowski et al. (2019a). The materials consisted of two minimal pairs differing only in their vowel (*stad/staat*, /stɑt, sta:t/, "city"/"state" and *takje/taakje*, /tɑkjə, ta:kjə/,

"twig"/"task"), each embedded in four Dutch context sentences (all containing 24 syllables) without any other instances of /ɑ/ and /a:/. Neither member of a word pair was favoured by the semantic context of the sentence (e.g., *Femke lette goed op of ze niet ging stotteren en toen heeft ze eens "stad/staat" tegen Roos gezegd*, "Femke took care not to stutter and then she said 'city/state' to Roos once"; see Appendix for all sentences and English paraphrases). All sentences were recorded by a native Dutch female and a native Dutch male talker, to increase the salience of talker voice differences. Recordings were divided into target words, buffers (i.e., three syllables before and one syllable after the target word to control for influences of proximal rate; $M_{pre-buffer} = 538$ ms, $M_{post-buffer} = 247$ ms), and context sentences (i.e., all speech up to the first buffer and all speech following the second buffer; see formatting in Appendix). Context sentences produced by the two talkers were set to the mean of their durations with PSOLA in Praat (Boersma & Weenink, 2015), such that they were matched in duration across the two talkers. Context sentences were then rate manipulated through linear expansion (factor of 1.6) and compression (factor of $1/1.6 = 0.625$) with PSOLA, resulting in three context speech rates: slow ($M_{pre-carrier} = 4106$ ms, $M_{post-carrier} = 1195$ ms), neutral (no further rate manipulation; $M_{pre-carrier} = 2566$ ms, $M_{post-carrier} = 747$ ms), and fast ($M_{pre-carrier} = 1604$ ms, $M_{post-carrier} = 467$ ms).

Spoken target words were excised and manipulated to create two duration continua, ranging from more /ɑ/-like to more /a:/-like perception. The Dutch vowel contrast /ɑ, a:/ is distinguished by both temporal and spectral cues, with /ɑ/ having a shorter duration and a lower F1/F2 than /a:/ (Adank et al., 2004). Therefore, five-step vowel duration continua ranging from 80 to 120 ms (in steps of 10 ms) with ambiguous spectral information (perceptually midway between /ɑ/ and /a:/) were created. First, one long vowel /a:/ was extracted for each talker and durations were manipulated using PSOLA. Then, the F1s and F2s from both talkers were computed and set to fixed ambiguous values with Burg's LPC algorithm as implemented in Praat. The male talker's F1 was 764 Hz and the F2 was 1261 Hz, and the female talker's F1 was 728 Hz and the F2 was 1327 Hz. Finally, the ambiguous vowels were spliced into their consonantal frames /st_t/ and /t_k/. The final set of auditory stimuli was created by concatenating the rate-manipulated context sentences, the original buffer intervals, and the manipulated target words. This resulted in 240 unique stimulus sentences, crossing eight context phrases with three rates, five vowel durations, and two talkers.

The visual targets on the screen were the two members of a minimal pair (e.g., *stad* and *staat*), presented orthographically in Arial, font size 16. Traditionally, the screen displays different objects or scenes (Allopenna, Magnuson, & Tanenhaus, 1998; Altmann & Kamide, 1999), but eye-tracking paradigms have also been used with orthographic words instead of pictures (McQueen & Viebahn, 2007; Huettig & McQueen, 2007; Huettig, Rommers, & Meyer, 2011).

Participants were allocated to either the high-rate or the low-rate group. The high-rate group was presented with 40 fast and 40 neutral auditory stimuli (eight sentences × two rates/talkers × five vowel duration targets) with the corresponding visual printed-word stimuli. The 80 auditory items were randomized within each of five blocks. The low-rate group heard 40 slow and 40 neutral items, which were also presented in randomized order in each of five blocks. As such, the high-rate group and the low-rate group listened to the same neutral speech from one talker, but to different rates from the other talker. Which talker (male/female) spoke at a neutral rate was counterbalanced between participants.

Procedure. In the experiment, participants were presented with an auditory stimulus while they looked at an 50.8 cm × 28.6 cm experimental screen with two written targets. The experiment was controlled using Presentation software (v16.5; Neurobehavioral Systems, Albany, CA, USA) combined with a tower-mounted EyeLink 1000 system (SR Research Ltd., Ottawa, Ontario, Canada) sampling at 1000 Hz. Participants were tested individually in a sound-attenuating booth, listening to the auditory stimuli over headphones. Before the start of the experiment, the eye-tracker was adjusted to a height that was comfortable for the participant, after which the system was calibrated. Eye-tracking data were obtained from participants' right eyes from stimulus onset until stimulus offset plus 1000 ms.

For both groups, the experiment started with instructions, followed by a practice round of eight trials that allowed participants to familiarize themselves with the experimental sentences and speech rates. In the high-rate group, four practice items had a fast speech rate and the other four had a neutral speech rate. In the low-rate group, four practice items had a slow speech rate and four a neutral speech rate. Target words in the practice items contained vowel tokens from the extremes of the duration continua (i.e., 80 and 120 ms) in order to emphasize the vowel contrast.

Each trial started with a fixation cross presented for 300 ms, followed by presentation of two written words as response options in black (either “takje” and “taakje” or “stad” and “staat”) at sound onset. The short /ɑ/-word was always shown on one side of the screen and the long /a:/-word was always shown on the other side of the screen. The position of response options (left/right) was counterbalanced between participants. The response options were shown during the whole trial until 1000 ms after sound offset. Participants were instructed to press either “1” on a regular keyboard for the word shown on the left of the screen or “0” for the word on the right side of the screen, thus categorizing the ambiguous target words. The response options were present on the screen from sound onset, but participants were instructed to respond only after they had heard the target word. At button press, the chosen response turned yellow until the end of the trial. If no response was given until 1000 ms after sound offset, a missing response was recorded. The session lasted approximately 50 minutes for the high-rate group and 70 minutes for the low-rate group.

5.3 Results

Categorization data. Figure 5.1 illustrates the categorization data in proportion of long vowel responses. The figure shows that participants more often indicated having heard a long vowel when the absolute durations of vowels were longer. Additionally, within each group, participants responded differently to the same vowels, depending on the distal speech rate context in which they were embedded, as depicted by the different line types. That is, within each group, the relatively faster rate (fast in high-rate group; neutral in low-rate group) induced more long /a:/ responses. Moreover, the figure suggests a difference in the perception of the neutral rate condition between groups, with a higher proportion of long /a:/ responses in the neutral rate for the low-rate group than for the high-rate group, as illustrated by the separation between the two lines in the middle.

We performed a logistic Generalized Linear Mixed Model (GLMM) from the `lme4` package (Bates et al., 2015) in `R` (R Core Team, 2014) on the categorization data (0.73% missing responses excluded). The fixed factor Rate Condition merged the between-participants global rate condition (high vs. low average rate) with the within-participants distal rate condition (fast/neutral/slow trial). As such, Rate Condition consisted of four levels of rate, namely high|fast and high|neutral in the high-rate group, and low|neutral and low|slow in the low-

rate group. High|neutral was mapped onto the intercept. We also included Vowel Duration as a continuous predictor, centered around the mean and divided by one standard deviation, and Block as a continuous predictor, centered around the mean and divided by one standard deviation, as fixed effects. Random intercepts were included for Participant and Item, with random slopes for all fixed effects by both random effects. Initially, the control condition Talker (categorical predictor; sum-to-zero coded) and the three two-way interactions between predictors Rate Condition, Vowel Duration, and Block were also included in the model. However, none of these predictors significantly improved model fit (as assessed by log-likelihood model fit using the `anova` function in `R`). They were therefore left out of the final model.

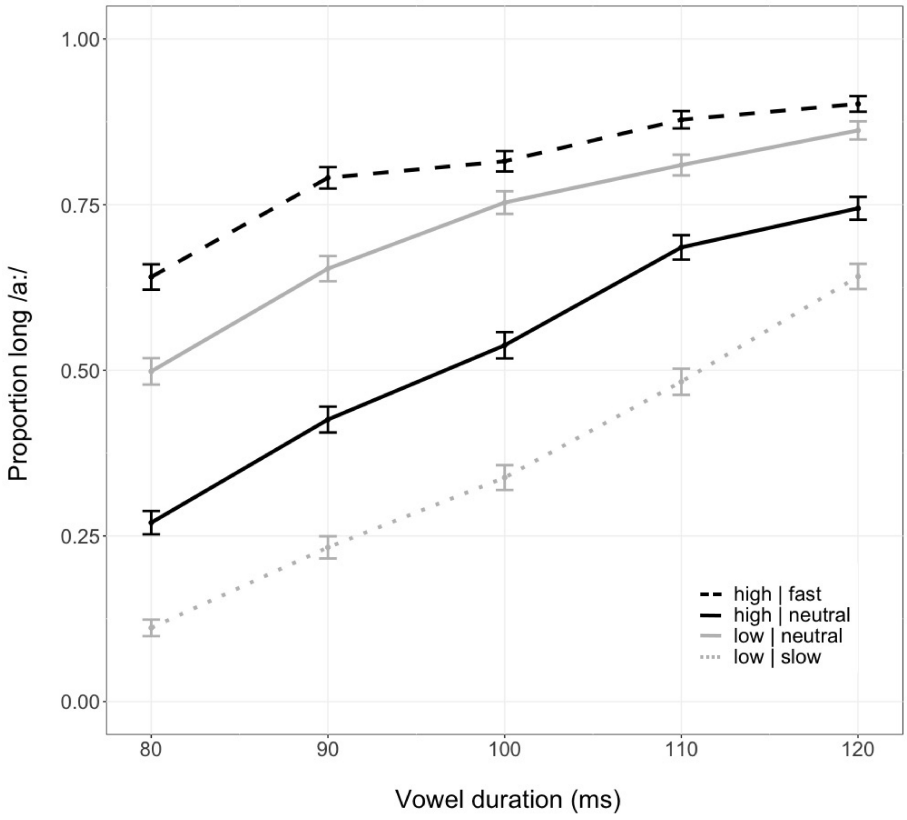

Figure 5.1: **Average categorization data in proportion of long /a:/ responses.** The X-axis indicates Vowel Duration (80–120 ms). Color indicates Group, with black representing the high-rate group and gray the low-rate group. Line type indicates Rate Condition, with a dashed line for fast contexts, solid for neutral contexts, and dotted for slow contexts. The critical comparison for the global rate effect is between the two solid lines, reflecting perception of the neutral speech rate condition in the two groups. Error bars represent the standard error of the mean.

The proportion of long /a:/ responses differed significantly with Vowel Duration; long vowel categorization increased for longer durations ($\beta = 1.254, z = 10.210, p < 0.001$). Rate Condition was significant within groups. High|fast received more long /a:/ responses than high|neutral ($\beta = 2.466, z = 4.981, p < 0.001$). A mathematically equivalent model mapping low|neutral onto the intercept revealed a similar pattern within the low-rate group: low|neutral received more long /a:/ responses than low|slow ($\beta = -2.760, z = -7.355, p < 0.001$). Moreover, between groups, the contrast between low|neutral and high|neutral was significant ($\beta = 1.896, z = 3.072, p = 0.002$), with more long vowel responses for neutral rate in the low-rate group (for whom neutral rate was relatively fast) compared to the high-rate group (for whom neutral rate was relatively slow). There was no significant main effect of Block ($\beta = -0.120, z = -1.168, p = 0.243$).

In sum, the results from the button-press categorization responses show within-groups effects of distal speech rate context; within groups, participants categorized target vowels more often as /a:/ when they were embedded in a relatively fast speech rate as compared to a relatively slow speech rate. Crucially, the results also demonstrate a between-groups global speech rate effect. The low-rate group categorized vowels in neutral rate speech more often as /a:/ than the high-rate group. That is, when neutral rate from one talker sounds relatively fast compared to the speech from another talker, perception of target words in neutral rate is biased towards hearing long vowel target words. Likewise, if a neutral rate seems slow, perception of target words in neutral rate is biased to hearing a shorter target word. This global rate effect in the categorization data replicates the results reported in Maslowski et al. (2019a, 2018). Thus, we proceed with our analysis of the eye-tracking data to assess the time courses of the distal and global rate effects.

Eye fixations. The raw eye-tracking data from each participant were downsampled from 1000 Hz to 250 Hz for simplicity. Samples with blinks and saccades were excluded from analysis. The areas of interest were set at 300 × 300 pixels around the center point of each target word. We only analyzed fixations to these interest areas. Hence, fixation proportions were calculated against the fixations to the two interest areas, not to the total number of fixations.

Figure 5.2 depicts the proportions of fixations to long /a:/ target words for each vowel duration (80–120 ms), collapsing across the rate conditions and

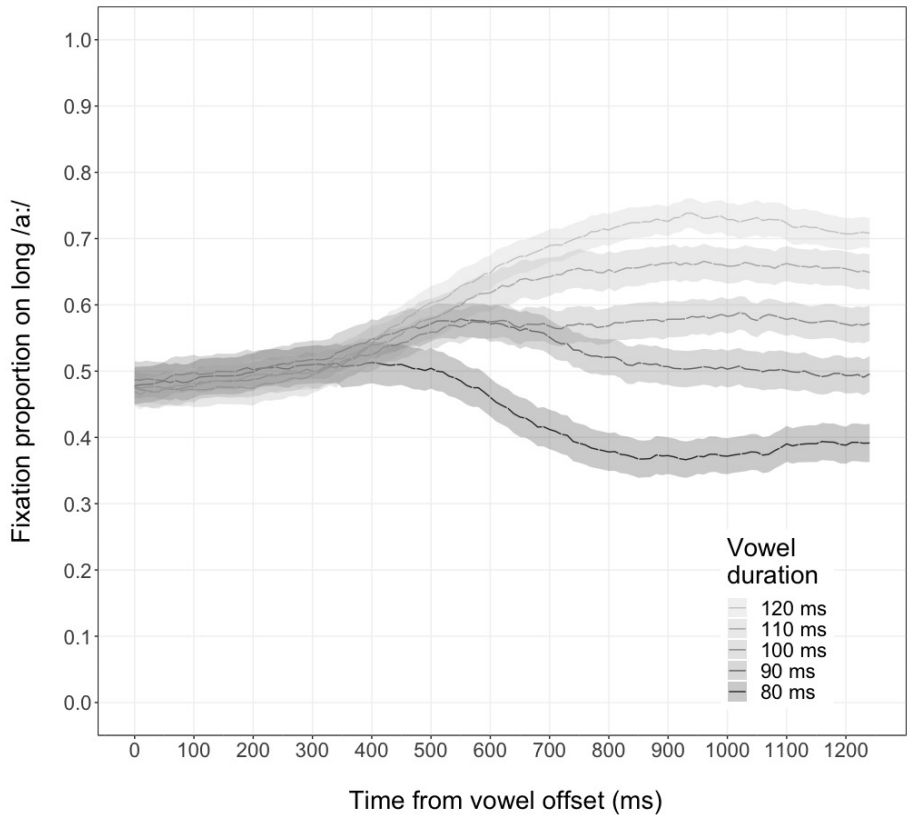

Figure 5.2: **Average fixation proportions to the long /a:/ target word as a function of target vowel duration (80–120 ms), collapsed across contextual speech rate conditions**. Increasingly longer vowel durations induced more looks to the long /a:/ target word. Time point 0 is the offset of the target vowel. The gray-shaded areas represent the standard error of the mean.

groups. This figure shows that the longer the duration of the vowel, the more participants fixated on the target with a long vowel.

Figure 5.3 shows the proportions of long /a:/ target word fixations as a function of the context speech rate in which the target word was embedded (collapsing across vowel durations), roughly reflecting the outcomes as illustrated in Figure 5.1. Figure 5.3 suggests that, within groups, participants were more likely to look at the long vowel words if the context speech rate was relatively fast (i.e., fast in the high-rate group; neutral in the low-rate group) than if the speech rate was relatively slow (i.e., neutral in the high-rate group; slow in the low-rate group). Also, it suggests that, across groups, participants' gaze patterns differed in the neutral rate conditions depending on the group: The low-rate group shows more fixations to long /a:/ targets in the neutral rate condition than the high-rate group. Moreover, Figure 5.3 suggests a difference between

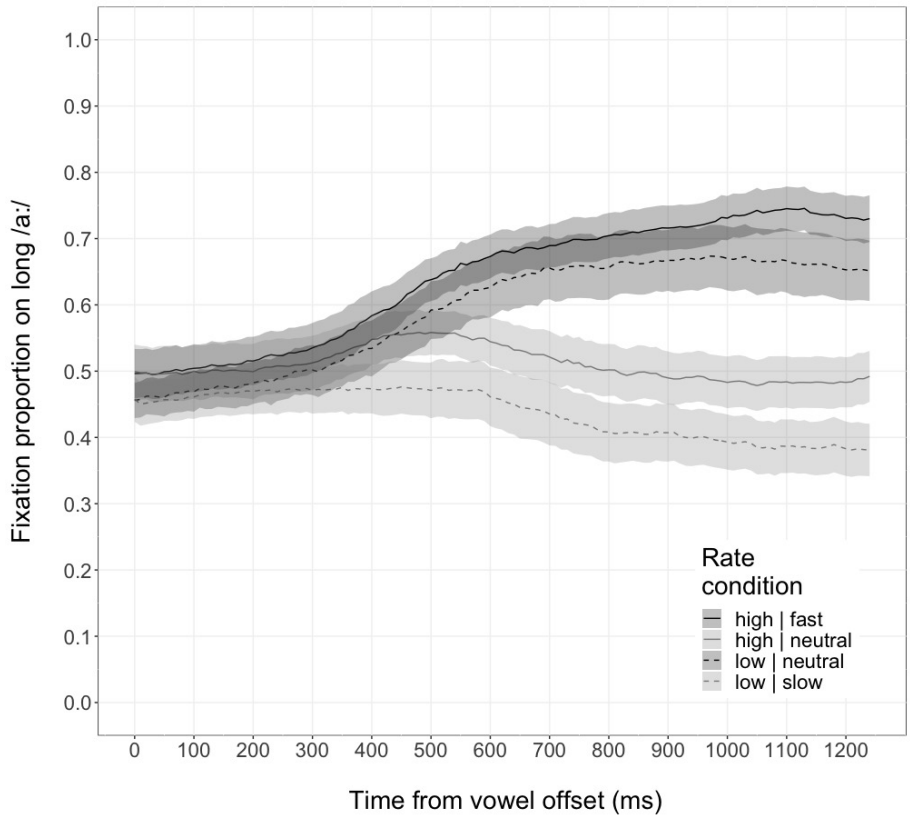

Figure 5.3: **Average fixation proportions to the long /a:/ target word as a function of contextual speech rate, collapsed across vowel durations**. Time point 0 is the offset of the target vowel. Line type indicates group. The black lines represent relatively fast speech rates within groups (fast in high-rate group; neutral in low-rate group) and gray lines represents relatively slow rates (neutral in high-rate group; slow in low-rate group). The gray-shaded areas represent the standard error of the mean.

the time courses of distal speech rate normalization (relatively high vs. relatively low speech rates) and global speech rate (neutral speech rate in high-rate group vs. low-rate group), with the two middle lines diverging at a later time point compared to the within-groups divergences.

Before the main analysis of the time courses of the distal and global speech rate effects, we determined whether the effects were present in the eye gaze data at all, reflecting the effects in the categorization responses. To statistically test the eye gaze data, we defined a time window of interest starting from 200 ms after target vowel offset. Target vowel offset is the earliest time point at which listeners can access the duration of the target vowel and 200 ms is the time it takes to program and launch a saccade (Altmann & Kamide, 1999). The

time window ended at 1250 ms after target vowel offset, since visual inspection showed stabilization of gaze patterns after around 1000 ms.

To test the influence of distal and global speech rates on the looks to long vowel targets, the logit-transformed proportions of eye fixations on the long /a:/ target word were quantified with a GLMM in R. Similar to the model that tested the categorization data, the fixed effects included in the model were Rate Condition (categorical predictor; high|neutral mapped onto the intercept), Vowel Duration (continuous predictor; centered and divided by one standard deviation), and Block (continuous predictor; centered and divided by one standard deviation). The model also comprised the interaction between Vowel Duration and Rate Condition. Other interactions between fixed effects were left out of the model, because their inclusion did not improve model fit. The random effects structure included random intercepts for Participant and Item as well as random slope terms for the three main effects by both Participant and Item.

The model showed a significant effect of Vowel Duration ($\beta = 0.825, t = 5.531, p < 0.001$), with more fixations to the long vowel target with increasing vowel durations. Within groups, Rate Condition was significant for the contrast between high|fast and high|neutral ($\beta = 1.560, t = 3.777, p = 0.002$), with more looks to the long target word in fast speech than in neutral rate speech. A mathematically equivalent model, this time mapping low|neutral onto the intercept, additionally showed a significantly lower proportion of looks to long /a:/ words in low|slow versus low|neutral ($\beta = -2.053, t = -7.170, p < 0.001$). Moreover, the between-groups contrast between low|neutral and high|neutral was significant ($\beta = 1.431, t = 2.843, p = 0.009$). Eye fixations did not differ significantly with Block ($\beta = 0.044, t = 0.543, p = 0.591$). There was a significant interaction between Vowel Duration and the contrast between Rate Conditions high|fast and high|neutral ($\beta = -0.434, t = -3.525, p < 0.001$), indicating that the difference between the context rates in the high-rate group was smaller for vowels with longer durations. The other interactions between Vowel Duration and Rate Conditions did not significantly affect the proportions of fixations to the long target words (low|neutral: $\beta = -0.133, t = -1.018, p = 0.313$; low|slow: $\beta = -0.087, t = -0.736, p = 0.464$).

The GLMM reported above showed differences in vowel perception within groups (i.e., distal rate effects). Moreover, there was a difference between the high-rate group and the low-rate group in the fixations to the long target word in the neutral speech condition (i.e., global rate effect), corroborating the global rate effect in the categorization data reported above.

To statistically test when, in the time window of interest, these effects arose, is not straightforward. Statistical methods typically used to measure time courses, such as growth curve analyses, cannot detect the onsets of effects in time series data like the present eye-tracking data. Eye-tracking data are typically a product of sampling from multiple random factors (e.g., participants and items; Baayen, Davidson, & Bates, 2008). Moreover, an auto-correlational structure underlies these densely sampled time series (Seedorff, Oleson, & McMurray, 2018; Cho, Brown-Schmidt, & Lee, 2018) and, hence, adjacent samples in time are generally the product of the same physiological events. At present, there is no single statistical tool that overcomes all these analytical factors (as comprehensively described in Seedorff et al., 2018). We selected the divergence metric implemented in the R package `eyetrackingR` (Dink & Ferguson, 2015).

We performed two comparisons, one on the within-groups distal rate effect and the other on the between-groups global rate effect. In order to measure the time course of the distal rate effect, the different rate conditions were coded with respect to the 'relative rate' within each group. That is, in the high-rate group, fast speech was coded as 'relatively high rate' and neutral speech was coded as 'relatively low rate'. Similarly, in the low-rate group, neutral speech was coded as 'relatively high rate' and slow speech was coded as 'relatively low rate'. As such, the distal rate effect was measured by a comparison of relatively high rates versus relatively low rates. Note that with this coding the neutral speech rate is 'relatively low' in the high-rate group but 'relatively high' in the low-rate group. The global rate effect was tested on the neutral rate conditions only (high|neutral vs. low|neutral).

For both comparisons, the eye gaze data were divided into time bins of 10 ms in the time window of interest. The data inside each bin were summarized by items ($n = 40$; five vowel duration steps combined with eight sentences) because of our between-participants design, which had no measurement of global rate within participants. This resulted in 40 fixation proportions on the long /a:/ target word for each condition inside each individual time bin, giving an indication of the proportion of participants that looked at the long /a:/ word at a given time. These were then subjected to a *t*-test, providing a significance estimate for the difference between two conditions in each time bin. In order to control the family-wise error rate of the large number of *t*-tests (one for each time bin), resulting *p*-values were Bonferroni corrected, providing a conservative statistical test. The first time bin that indicated a statistically significant difference between

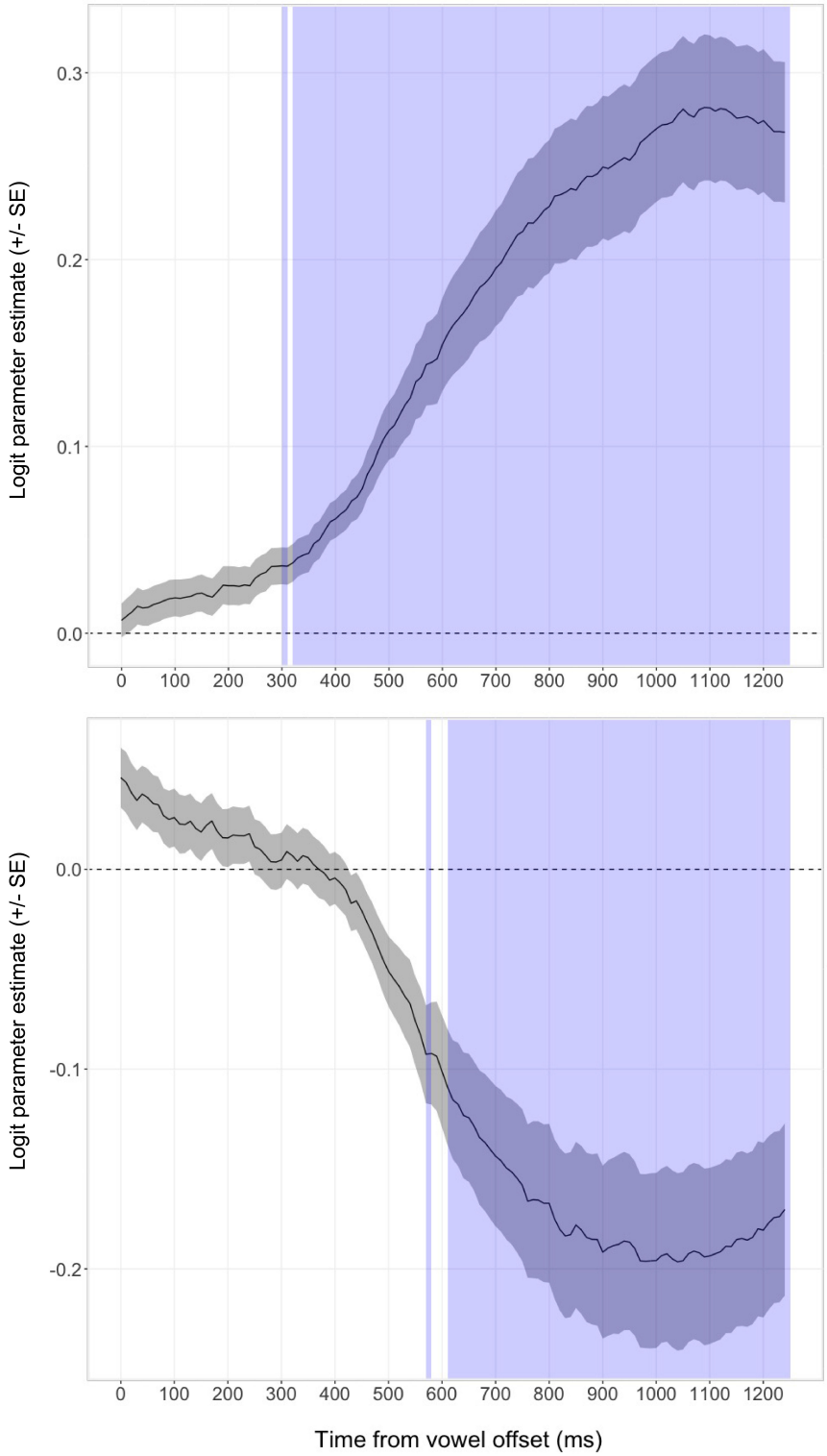

Figure 5.4: **Difference curves of distal and global rate effects.** Upper panel: Difference curve of distal speech rate effect (within groups) as established by the divergence metric. Lower panel: Difference curve of global speech rate effect (across groups) as established by the divergence metric. The blue shaded areas show where t-test divergences were significant (after Bonferroni correction). The grey areas represent plus or minus one standard error of the mean.

two conditions (i.e., $p < 0.05$ after Bonferroni correction) was taken as the first time point at which the effect of interest could be reliably detected.

Figure 5.4 shows the difference curves of the distal speech rate effect (within groups; upper panel) and the global speech rate effect (across groups; lower panel). The blue shaded areas show all the time bins where *t*-tests resulted in $p < 0.05$ after Bonferroni correction. Within groups, the difference in looks to the long /a:/ target words between relatively faster rates versus relatively slower rates was estimated to arise approximately 300–320 ms after target vowel offset. Between groups, the difference in looks to the long /a:/ target words between high|neutral and low|neutral arose approximately 570–610 ms after vowel offset. This indicates that the distal rate effect was detected very early after vowel offset, while the global rate effect arose later.

5.4 General discussion

This study aimed to determine and compare the time courses of the distal speech rate effect (effect of the sentence context speech rate) and the global speech rate effect (effect of the speech rate context beyond the sentence) on /ɑ, a:/ perception in Dutch. The experiment tested two groups of participants. One group listened to a neutral rate Talker A and a fast Talker B (i.e., the high-rate group). The other group listened to neutral Talker A, but to Talker B speaking at a slow speech rate (i.e., the low-rate group). Participants performed a two-alternative forced choice task, in which they had to indicate whether they had heard a word with an /ɑ/ or with an /a:/ (e.g., *stad/staat*, /stɑt, sta:t/, "city"/"state"). Additionally, their eye fixations were measured to investigate when they looked at a given written target word on the screen. The distal rate effect was measured by comparing categorization responses and eye fixations within groups (fast vs. neutral in high-rate group; neutral vs. slow in low-rate group), whereas the global speech rate effect was measured by comparing the two between-groups neutral rate conditions.

Regarding the categorization results, we observed a within-group distal speech rate effect in each of the two groups, with relatively faster speech rates (i.e., fast in high-rate group; neutral in low-rate group) receiving more long /a:/ responses compared to relatively slower speech rates (i.e., neutral in high-rate group; slow in low-rate group). Moreover, we found a between-groups effect of global speech rate, with neutral rate in the low-rate group receiving more long /a:/ responses than neutral rate speech in the high-rate group. These categorization responses

replicate earlier work on distal and global speech rate effects (Maslowski et al., 2018, 2019a).

With regard to the time courses of the distal and global speech rate effects, we hypothesized that the effects would arise at different times, rather than manifesting themselves simultaneously. Specifically, we expected the global speech rate effect to emerge later than the distal rate effect. This hypothesis was based on predictions made by Bosker et al.'s (2017) two-stage hierarchical model for acoustic context effects. This model states that acoustic context effects take place at two different stages. The first stage involves perceptual processing of auditory input, which is domain-general, automatic, and obligatory, whereas the second stage takes into account higher-level factors such as talker identity. Because distal speech rate directly affects perceptual processing of temporally ambiguous sounds (e.g., Reinisch et al., 2011), distal rate information is argued to be used at the first stage. The global speech rate effect, however, has been suggested arise at the subsequent stage (Bosker & Ghitza, 2018; Maslowski et al., 2018, 2019a), given that global rate tracking is sensitive to talker-identity (Maslowski et al., 2018) and can be overriden by local speech rate variation (Reinisch, 2016b). The two-stage model therefore predicts that the global rate effect should be observed in a later time window than distal rate normalization.

The results from the eye fixations were consistent with the predictions from the two-stage model; the global rate effect arose considerably later (earliest significant time point was 570 ms after vowel offset) than the distal rate effect (earliest significant time point was 300 ms). The time course of the distal rate effect is comparable to the time courses of the speech rate effects found by Reinisch and Sjerps (2013), Toscano and McMurray (2015), and Kaufeld et al. (2019). Kaufeld et al. found rate normalization effects of distal speech rate after 250 ms after vowel offset. Reinisch and Sjerps and Toscano and McMurray estimated their speech rate effects to start between 300 and 400 ms after target vowel onset (Reinisch and Sjerps) and target word onset respectively (Toscano and McMurray). Given that time point 0 in the current study was target vowel offset, our time line should be shifted approximately 100 ms (i.e., the average vowel duration in the current study) to the right for an accurate comparison. That is, our distal rate effect arose about 400 ms after vowel onset, whereas the global effect arose after approximately 670 ms. The timing of our distal rate effect is very similar to the effects reported in Reinisch and Sjerps, Toscano and McMurray, and Kaufeld et al. However, since the estimated starting points are dependent on, for instance, the width of the time bins chosen, these specific time points

should be considered approximations, rather than precise estimates on a millisecond timescale. Furthermore, note that Toscano and McMurray manipulated proximal (immediately adjacent) and distal (non-adjacent) contexts simultaneously, entailing that their effect could be proximal, distal, or a combination of the two.

The current study established the time courses and separability of the distal and the global speech rate effects. Accounts of neural entrainment to speech have attempted to explain distal speech rate effects as a result of neural oscillations in the theta range phase-locking to the (slow and fast) amplitude modulations in the context (Ghitza, 2012; Giraud & Poeppel, 2012; Peelle & Davis, 2012). For instance, Kösem et al. (2018) tested whether neural oscillations can directly shape perception of the following speech signal, using magnetoencephalography (MEG). In their study, participants listened to fast and slow sentences (i.e., distal rate manipulation), followed by a Dutch ambiguous /ɑ-a:/ target word. They found that the brain tracked the speech rhythm of the fast and slow context sentences. Additionally, these fast and slow neural rhythms were observed to persist even after the context sentence had ceased: In the same target time window, evidence for a fast neural rhythm was present when preceded by a fast context sentence, but a slow neural rhythm was present when preceded by a slow context sentence. Hence, the speech-brain entrainment induced by the fast and slow context sentences carried on for a number of cycles after the rhythm it was entrained to changed. Moreover, the extent to which individuals showed this sustained neural entrainment in the target window was predictive of the behavioral rate effect: Individuals who showed stronger entrainment to the speech rhythm of the distal context also showed a larger perceptual bias in word categorization in the expected direction (i.e., slow-rate entrainment to short-vowel target words and fast-rate entrainment to long-vowel target words). This is in line with psychoacoustic findings that only rhythms in the theta range (3–9 Hz) induce these distal rate effects (Bosker & Ghitza, 2018), and that destroying the rhythm in the distal context also eliminates distal rate effects (Bosker, 2017a). This is accounted for by a temporal sampling framework, whereby entrained theta oscillations impose periodic phases of neuronal excitation and inhibition, thus sampling the input signal at the appropriate temporal granularity.

An important question for further work is whether neural entrainment can also explain global speech rate effects. Previous studies on global speech rate tracking have suggested that their effects had neural correlates similar to the ones found for distal speech rate (Baese-Berk et al., 2014). However, Alexandrou, Saari-

nen, Kujala, and Salmelin (2018), using MEG, observed spatial differentiation between the neural regions involved in processing global speech rate (temporal cortex bilaterally and right parietal cortex) and those processing distal speech rate variation (left parietal regions). The present outcomes extend the finding of spatially distinct neural regions that underlie global and distal speech rate processing by providing evidence for temporal differentiation as well: We show that global rate effects have a distinct time course in perceptual processing relative to distal rate effects (i.e., arise later). Hence, it would seem unlikely that one and the same neurobiological mechanism could account for both global and distal speech rate processing without additional principles.

In fact, previous research on the global speech rate effect has shown that this global effect is subject to constraints that have not been found for the distal rate effect. Specifically, the global speech rate effect is talker-specific (Maslowski et al., 2019a, 2018; Reinisch, 2016b) and global rate tracking fails with considerable speech rate variation within a given talker (Maslowski et al., 2019a; Reinisch, 2016b), whereas the distal rate effect seems automatic and obligatory. Thus, while general-auditory mechanisms like sustained neural entrainment have been proposed to underlie the distal rate effect (Bosker, 2017a; Wade & Holt, 2005), the global speech rate effect is unlikely to be explained by domain-general auditory principles.

The findings of the current study corroborate this; the difference in time courses between the global and distal speech rate effects shows that participants took longer to take global speech rate into account, compared to distal speech rate. This indicates that higher-level factors are considered after the first perceptual normalization for distal rate. Consequently, the global rate effect in the present study does not fit in a straight-forward manner with current theories of neural entrainment to speech rate, in which brain oscillations adapt to the rhythm of an auditory signal independent from talker identity (Bosker, 2017a).

Therefore, in our view, accounts of neural entrainment (in their present form) are unlikely to suffice to explain how the talker-specific and relatively late global rate effect influences perception of temporally ambiguous cues. These theories may be adapted to explain the global rate effect by including a system for talker recognition that feeds into the mechanism that tracks the speech envelope. The question remains, however, whether it is plausible that such mechanisms, one for rate tracking and one for talker tracking, feed into each other, since proximal and distal speech rate effects seem to involve general-auditory mechanisms without the need for a system that incorporates feedback about the talker's iden-

tity (Newman & Sawusch, 2009; Bosker, 2017b). It may also be the case that, although different types of speech rate effects have typically been described as falling into the same category of rate normalization, they involve distinct computations that are not yet well described.

In sum, this study measured online language processing with an eye-tracking paradigm to test the time course of lexical activation, varying distal and global contextual speech rates. The results illuminate timing differences between the two speech rate effects, supporting the idea that the distal speech rate effect may underlie an early perceptual mechanism, whereas the global speech rate effect may be controlled by a different cognitive adjustment mechanism. This is in line with predictions from the two-stage model of acoustic context effects (Bosker et al., 2017). Future work may investigate which neurobiological mechanisms underlie global speech rate processing, joining the distal and global rate effects that have parallel consequences for speech perception in a single theoretical, neurobiologically plausible framework.

Appendix

Stimulus sentences. Two talkers were recorded producing a set of eight Dutch stimulus sentences (English paraphrases 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].

	Sentences and translations
1	[Peter fluisterde Ilse iets verkeerd in en toen hoorde] **Ilse het** tak-/taak**je** [gezegd worden]. *Peter whispered something in Ilse's 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]. *While 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.*

6 | General discussion

Imagine driving on the motorway at high speed, minutes before turning in another road with a lower speed limit. Before actually making the turn, you can already predict how you may experience driving on that slower road. Even if driving on the slower road is on average fast (e.g., 60 $^{km}/_{h}$), it will seem moderately slow, as a result of having just driven at a faster speed. However, the same speed will seem fast after coming from a city road with an even lower speed limit (e.g., 30 $^{km}/_{h}$), because your brain "tricks you into" experiencing the different speeds relative to each other.

This doctoral thesis examined how listeners perceive different rates of speech. Specifically, it looked into the effects of distal speech rate (i.e., non-adjacent within-sentence rate) and global speech rate (i.e., beyond-sentence rates from multiple talkers) on word recognition, to shed light on the underlying mechanisms at play during speech rate processing. This chapter summarizes the core findings of the preceding empirical chapters and relates these findings to mechanisms that may underlie speech rate tracking. It also discusses future research directions.

6.1 Summary of main findings

Chapter 2 tested the prediction that distal rate tracking is automatic and obligatory for word recognition by investigating whether and how speech rate normalization influences lexical access, without an explicit recognition task. We measured cross-modal repetition priming on Dutch /ɑ, a:/ words to assess whether participants had most likely heard an /ɑ/-word or an /a:/-word in a prime sentence. That is, participants listened to ambiguous prime words (e.g., /m?t/) embedded in fast and slow prime sentences, after which they had to perform a lexical decision task on a visual target word on a computer screen ("mat", "maat", or "zon"). A fast sentence was hypothesized to lead to perceiving more /a:/ words in the prime sentence, thus facilitating lexical decision when the tar-

get word also had a long /a:/-vowel (as opposed to its minimal pair twin with a short /ɑ/-vowel). The results revealed that participants normalized the prime words for speech rate, even in a task where no explicit attention was drawn to the temporally ambiguous words. This suggests that the processing of distal rate is indeed automatic and obligatory, involving perceptual processes independent of decision making.

Chapters 3, *4*, and *5* investigated speech rate effects of global, more distant speech rate contexts and contrasted these contexts with distal within-sentence rate contexts. In these chapters, the same experimental design was used: One participant group listened to two talkers, with Talker A speaking at a neutral speech rate and Talker B speaking at a fast speech rate (high-rate group). In the other participant group, Talker A again spoke at a neutral rate, but here Talker B was slow (low-rate group). The global rate effect was measured by comparing the two participant groups on their perception of Dutch /ɑ, a:/-words in Talker A's neutral rate speech. *Chapters 3*, *4*, and *5* all demonstrated robust effects of global speech rate: Participants in the high-rate group reported fewer long /a:/ words in Talker A's neutral rate speech than participants in the low-rate group. This suggests that neutral rate sounded slow when it was surrounded by fast speech, but fast when it was surrounded by slow speech. This contrastive effect of global speech rate was found both with artificially compressed and expanded speech (*Chapter 3*, Experiment 2) and with naturally produced fast and slow speech (*Chapter 4*, Experiment 3).

Chapter 3's Experiment 3 also revealed a constraint on the global rate effect. In this experiment, participants listened to Talkers A and B both speaking at a neutral rate and a fast rate (high-rate group) or at a neutral rate and a slow rate (low-rate group), such that there was no distinction between the rate variation in the speech produced by Talker A and Talker B. No difference in perception of neutral rate speech between the two groups was found. This indicates that global rate tracking fails with a reasonable amount of within-talker speech rate variation.

Chapter 4 uncovered another constraint to the global speech rate effect. Two groups of participants again categorized Dutch temporally ambiguous /ɑ-a:/ words in Talker A's neutral rate speech, but this time, the surrounding global context was their own fast or slow speech. The results showed no effect of global rate when the surrounding speech context was self-produced. This was true both when listeners heard their own voice during production (Experiment 1), and when they were listening to playback of their own voices (Experiment 2).

Thus, *Chapters 3* and *4* found that the global speech rate effect is conditioned by factors that have not been found for distal speech rate effects. That is, the global rate effect is not induced by self-produced speech, whereas the distal speech rate effect arises independent of talker identity (carriers in one voice can influence perception of targets in another voice; Newman & Sawusch, 2009) and can be self-induced (Bosker, 2017b). These findings suggest that different mechanisms may underlie these two types of effect. Whereas distal speech rate processing seems to involve automatic perceptual normalization, global speech rate processing seems to involve additional higher-level cognitive adjustments, that may arise at a later point in time.

The results of *Chapters 3* and *4* led us to investigate when in time the distal and global rate effects become observable. *Chapter 5* again compared two participant groups that listened to either neutral and fast speech or to neutral and slow speech. An eye-tracking paradigm was used to assess the time courses of the global and distal speech rate effects. Participants saw two response options on a screen and their looks to these two words were measured to evaluate when in time a preference arose for one word over the other. The results revealed timing differences between the global and distal speech rate effects, with the global speech rate effect manifesting itself more than 250 ms later than the distal rate effect. This late effect of global speech rate supports the idea that it involves higher-level cognitive adjustments, compared to the distal speech rate effect, which seems to be perceptual.

6.2 A processing model of speech rate tracking

When listening to speech, listeners frequently encounter temporally ambiguous words. This is because there is no one-to-one correspondence between acoustic cues to duration and temporal phonological contrasts such as the Dutch vowel contrast between /ɑ/ and /a:/. Comprehension is facilitated by keeping track of the context speech rate: Listeners use the acoustic rate cues in the context speech and relate these cues to a temporally ambiguous word, in order to arrive at the word the talker most likely intended to produce. The results of the research in this thesis showed that listeners take into account a variety of speech rate cues in perception of temporally ambiguous speech. Furthermore, the results highlight the complexity of tracking global speech rate cues (relative to distal rate cues), the use of which depends on additional factors.

Previous research has argued that distal speech rate effects involve early automatic perceptual processes. Effects of distal speech rate have been found to arise early in perceptual processing, between 300 and 400 ms after target sound onset (Reinisch & Sjerps, 2013; Toscano & McMurray, 2015; Kaufeld et al., 2019). They are not sensitive to changes in talker voice (Newman & Sawusch, 2009; Bosker, 2017b), and have also been found in infants (Eimas & Miller, 1980) and bird species (Dent et al., 1997; Welch, Sawusch, & Dent, 2009). Even non-speech contexts such as pulse trains can induce distal speech rate effects (Gordon, 1988; Diehl & Walsh, 1989; Wade & Holt, 2005; Bosker, 2017a), and such effects are unaffected by attention (Bosker et al., 2017).

The results in this thesis add to these findings on distal speech rate tracking. The findings support the involvement of automatic processes in distal speech rate tracking: Without exception, the experiments in *Chapters 2* to *5* found effects of distal speech rate on perception of the Dutch vowel contrast between /ɑ/ and /a:/. Moreover, *Chapter 5* found that distal rate effects were statistically reliably observable at 300 ms after vowel offset (400 ms after vowel onset; cf. Reinisch & Sjerps, 2013; Toscano & McMurray, 2015). Finally, *Chapter 2* demonstrated that distal rate normalization of prime words takes place even without attention being drawn to temporally ambiguous words in an implicit task, where categorization of another stimulus (i.e., the target) took place.

These findings for distal speech rate contrast with the findings for global speech rate, for which several constraints were uncovered in this thesis. Firstly, *Chapter 3* found that the global rate effect is talker-specific (i.e., a global rate effect was only found when Talker A had a different speech rate from Talker B; cf. Experiment 3). This indicates that the system responsible for global rate tracking also keeps track of which talker produced a specific rate and keeps this information in memory.

Additionally, *Chapter 3* found that the global rate effect emerges only when talkers' speech rates are relatively stable. In *Chapter 3*'s Experiment 3, the rate variation between the fast and neutral conditions and the neutral and slow condition was large, blocking the global rate effect. *Chapter 4*'s Experiment 3 showed that some within-talker variability in global speech rate is allowed before the tracking mechanism fails, as naturally produced fast and slow speech induced global rate effects. These results point to the global rate effect being driven by habitual speech rates, which listeners use only when these rates are relatively stable. These findings suggest that the global rate tracking system only sends

speech rate cues to memory that are consistent over a longer time and neglects global rate cues from talkers speaking at highly variable speech rates.

Chapter 4 showed that cues in one's own global speech rate are not used for processing of subsequent speech, not even when passively listening back to recordings of one's own voice. This indicates that global rate effects are not only talker-specific, but also sensitive to talker identity.

Finally, *Chapter 5* found that it takes longer to process global speech rate cues than distal speech rate cues: The global rate effect arose more than 250 ms later than the distal rate effect. This result, together with the findings in *Chapters 3* and *4*, suggests that the perceptual outcome of the acoustic signal based on vowel internal cues and distal speech rate cues is merged with higher-level information, which takes approximately 250 ms.

The challenge is to explain all rate effects within one coherent model. As noted above, the findings of this thesis indicate that listeners need to keep a memory representation of previous global rate cues that is contingent on additional higher-level factors, such as talker identity. Thus, to be able to account for both distal and global rate effects, a distinction needs to be made between an online tracking component and a memory component. The online tracking component estimates the phonetic information in the surrounding sentence context and then integrates these within-sentence rate cues with the information stored in working memory. *Chapters 3* and *4* suggest that this memory representation at the very least must comprise information about talker identity and habitual rate.

The merger of rate cues with higher-level information may work in a similar fashion as how listeners process other types of linguistic information. That is, similar problems may arise with syntactic ambiguity in incremental sentence processing (e.g., when listeners need to interpret the referent of a pronoun; Gordon & Scearce, 1995), where world knowledge and syntactic cues need to be combined to comprehend the meaning of a sentence. Similarly, comprehenders may encounter situations where they have to integrate cues from different sensory modalities (e.g., the McGurk effect; McGurk & MacDonald, 1976). It has been proposed that comprehenders may calculate statistical contingencies that evaluate the probability that several cues co-occur (e.g., Saffran et al., 1999; Kirkham, Slemmer, & Johnson, 2002; Demberg & Keller, 2008). Applying this proposal to the use of talker-specific global speech rate cues, I suggest that listeners calculate the likelihood of a specific talker speaking at a specific speech rate, and only retain global cues when the co-occurrence of talker voice and speech rate is above

a certain probability. Listeners have some (individual) threshold for relying on the outcome. Future studies may investigate this hypothesis by systematically varying co-occurrences of specific talkers and speech rates.

Postulating an effect of memory in addition to a low-level perceptual mechanism not only explains how global effects may arise. It can also explain how a following within-sentence context can influence perception of an ambiguous speech sound or word just heard. An example of this is found in Newman and Sawusch (2009), who showed that segments both adjacent and non-adjacent to a preceding target can influence perception of that target. This result may be explained by the percept of an ambiguous sound or word just heard being held active in working memory. This memory trace can then be altered online. This allows the listener to change the percept of the ambiguous sound, based on the following context. As such, there may be an alteration process taking place, unfolding from online acoustic processes to the memory component, where it takes longer to commit to lexical activation for more ambiguous sounds.

An efficient way to conceptualize these principles with the findings of this thesis is to adopt a cue integration framework, as proposed by, for instance, Martin (2016) and Toscano and McMurray (2012) (also introduced in *Chapter 2*) with a statistical learning mechanism. Statistical learning involves detecting regularities in the acoustic signal and coupling these regularities to knowledge about the world (e.g., Talker A usually speaks slowly). Cues in the signal are weighted, depending on their probability to be informative for comprehension based on previous experience. In cue integration frameworks, multiple cues in the signal (e.g., speech rate, speaker voice, vowel length, and vowel quality) are used as soon as they have been processed. Different acoustic cues may have different weights, based on how reliable they are estimated to be. For instance, in a noisy environment, vowel quality may be weighted as less important than vowel length, as the former is less salient when the speech signal is distorted. When listeners listen to within-sentence rate cues, these cues will always be processed because they are reliable for speech rate estimation and consequently they are weighted heavily. Talker-specific global rate cues may start off as cues with a low weight, because of greater within-talker speech rate variation than between-talker rate variation (Quené, 2008). That is, the rate of the sentence is more predictive of following vowel durations than the speech rate of the talker, because rates vary more within talkers than between. However, talker-specific cues may be upweighted as they become reliable over time. Similarly, such cues may be downweighted when the global rate tracking system observes that they

are highly variable, which causes the global rate effect to disappear with unstable global speech rates.

A related framework is Kleinschmidt and Jaeger's (2015) belief-updating model of perceptual adaptation (also discussed in *Chapter 3*), where listeners' 'beliefs' about cue distributions that are based on previous experience determine which cues in the signal are used during processing of the incoming signal. These beliefs can be updated with increasing evidence for a particular pattern. For instance, if Talker A always speaks at a slow rate, the listener builds up the expectation that she will do so again, consequently upweighting talker-specific cues for Talker A, whereas a less specific talker-general expectation will be built up for Talker B, if he occasionally speaks slowly and otherwise fast, with talker-specific cues thus being downweighted.

Both approaches can account for the emergence of the distal and the global speech rate effects, but they cannot explain the timing differences between the two effects. The fact that the global rate effect comes with a delay compared to the distal rate effect may either suggest that there are control processes that license the use of global cues, or that it takes longer to decide between competing words when cues are more ambiguous (i.e., the global rate may compete with the distal rate, causing it to be more ambiguous), because the evidence accumulation for one word or the other takes longer. I propose that the latter option is more likely, given that *Chapter 5* found that the distal rate effect arose somewhat later when target words were temporally more ambiguous (i.e., when the ambiguous target vowels were in the middle of the vowel continua), compared to when target words were temporally less ambiguous.

On a neurobiological level, the effects may be implemented by neural entrainment, where the brain tracks the syllabic speech rhythm (Ghitza, 2012; Giraud & Poeppel, 2012; Peelle & Davis, 2012; Kösem et al., 2018). In the literature, distal rate effects have been proposed to be a result of sustained neural entrainment to the fast or slow speech rhythm of the context. For instance, Kösem et al. (2018) tested whether neural oscillations actively shape speech perception. In their magnetoencephalography (MEG) experiment, participants were presented with fast and slow sentences, followed by a temporally ambiguous target word that they had to categorize. Kösem et al. observed that neural oscillations in the theta band entrained to the rhythm of the sentences. These oscillations persisted for a few cycles into the target word time window, after termination of the context sentence. Moreover, the extent to which participants showed sustained entrainment was related to their behavioral categorization biases: High entrain-

ment was associated with a stronger categorization bias than low entrainment. This suggests that the perceived length of a segment (e.g., a vowel such as /ɑ/ or /a:/) is driven by the number of oscillatory cycles during a temporally ambiguous target word. These, in turn, are driven by the entrained rhythm in the context sentence (hence: sustained entrainment).

Alexandrou et al. (2018), using MEG, tested which oscillatory components encoded within-sentence distal speech rate and which encoded global average speech rate. Participants listened to 40-second fragments differing in global speech rate, with within-sentence rate variations. Alexandrou et al. found that distal speech rate was associated with modulations in the theta band (4–7 Hz) and global speech rate with modulations in the delta band (2–4 Hz). Moreover, they also found that different neural regions were involved during distal and global rate processing: Left parietal regions were involved during distal speech rate processing, and bilateral temporal and right parietal cortex regions were involved during global speech rate processing. Consistent with the results of the present thesis, this study suggests that distal and global rate tracking may involve different processes. However, in Alexandrou et al., participants performed no categorization or identification task. Therefore, it is hard to interpret their findings in light of the results of the present thesis.

Reviewing the global rate effects found in this thesis, it is evident that neural entrainment on its own cannot account for all speech rate effects: (1) The global speech rate effect is talker-specific (i.e., self-produced speech induces no global rate effect; *Chapter 4*), (2) it is sensitive to within-talker rate variation (*Chapter 3*), and it emerges late in time (*Chapter 5*). Moreover, the global 'habitual' speech rates that listeners track are partially based on other speech rates. Thus, the global speech rate is an inferred rate, because it is contrasted to other surrounding speech rates. For instance, the high-rate groups in *Chapters 3* to *5* recognized 'neutral' Talker A as a slow talker because Talker B spoke faster, whereas the low-rate groups recognized Talker A as fast. That is, the high-rate groups inferred a lower speech rate for Talker A than her actual speech rate (as evidenced by fewer long vowel responses for neutral Talker A), whereas the low-rate groups inferred a higher speech rate for that particular talker. Similarly, the high-rate groups perceived fast-speaking Talker B to be 'even faster' and the low-rate groups slow-speaking Talker B to be 'even slower' than their actual speech rates. As a result, the global rates that form the basis for listeners' perception of ambiguous sounds and words do not reflect actually perceived rates, calculated as the number of syllables per second. Rather, the global rates are derivatives

of the perceived speech rates that have been recognized as 'relatively fast' or 'relatively slow', determined by other surrounding speech rates. The inferred global speech rate, based on expectancies from speech rates previously heard, is taken into account in perception of subsequent speech from the same (Reinisch, 2016b) or other (*Chapters 3* and *4*) talkers. If the neural entrainment account is true, neurons entrain to the rhythm of the acoustic signal (independent from talker identity), not the inferred global rate. Since the recognized global speech rate is inferential, the actual rhythm of the speech signal and the abstract global speech rate may not match. Crucially, this means that neurons cannot entrain to the global speech rates tracked by the listener without additional principles such as the memory effect described above. Therefore, in order to construct a psychological model that unifies all speech rate effects, it is essential to incorporate a memory component that feeds into the mechanism tracking the speech envelope.

6.3 Future research directions

An integrative model of speech rate tracking must not only account for the results reported in the current thesis but should have a wider scope. For instance, such a model should also be able to account for speech rate effects on semantic effects and word segmentation. Therefore, an important challenge for future research is to compare the global speech rate effect on phonetic boundary shifts (/ɑ/ vs. /a:/; as repeatedly observed in this thesis) to the global speech rate effect on function word perception, as described in Baese-Berk et al. (2014). Baese-Berk et al. compared three participants groups, who each listened to a variety of speech rates that differed in their average speech rate across groups. They found an effect of the global speech rate on perception of function words (lexical rate effect), but the direction of this effect was opposite to the effects found in the present thesis. That is, instead of 'neutral' rate speech sounding slow in the presence of fast speech (i.e., a contrastive effect), neutral rate speech sounded slightly faster in the presence of fast speech (i.e., an assimilative effect). Interestingly, it has been suggested that phoneme-level rate effects (e.g., Summerfield, 1981; Nooteboom, 1981; Miller & Baer, 1983; Kidd, 1989; Reinisch et al., 2011) and lexical rate effects (Dilley & Pitt, 2010; Dilley et al., 2013; Morrill et al., 2014; Pitt et al., 2016; Lai & Dilley, 2016; O'Dell & Nieminen, 2018) involve differential mechanisms because they behave differently (Heffner, Newman, & Idsardi, 2017). For instance, it has been found that lexical rate effects are

only induced by intelligible speech (Pitt et al., 2016), whereas phoneme-level effects also arise after unintelligible speech (Kluender, 1984) and even non-speech (Gordon, 1988; Diehl & Walsh, 1989; Wade & Holt, 2005; Bosker, 2017a). In an attempt to explain the difference between the phonetic boundary shift effect and the lexical rate effect, Pitt et al. (2016) proposed that rate (of any kind) by necessity has a contrastive effect on two phonemes differing in their temporal properties only, whereas the lexical rate effect "seems more akin to perceptual assimilation than contrast" (Pitt et al., 2016, p. 342). The implications of this statement remain to be formulated; therefore, how these two effects of global speech rate on segmental ambiguities and lexical ambiguities, respectively, relate to each other remains unresolved. Attempts to compare distal cues to word segmentation and segments have already been made (see Heffner et al., 2017). Future work may also want to compare the two global effects directly, to identify the source for the different perceptual outcomes in global speech contexts.

In this thesis, and in most other relevant research, the stimuli were semantically unbiased. That is, both members of a minimal pair were possible interpretations in a given carrier phrase. Because disambiguation in natural language can often be done based on the meaning of the sentence, generalization of the results of this thesis to natural language is limited. Therefore, another interesting avenue for future work is to explore interactions between speech rate tracking and top-down lexical influences such as (semantic) prediction, to see how well the effects found in this thesis scale up to speech processing in more natural settings. Because natural language interpretation is expectation-driven, one could test, for instance, how the brain responds to conflicting cues from bottom-up acoustic speech rate cues and top-down semantic cues in a priming paradigm.

6.4 General conclusion

In sum, the findings in this doctoral thesis demonstrate that the speech rate context in which words are uttered (and who utters them) can systematically change the way speech is perceived. The results suggest that speech rate effects take place in everyday communication, as they are induced by naturally produced fast and slow speech and speech rate tracking is highly automatic, taking place without explicit attention being drawn to temporally ambiguous words. These findings highlight the complexity of speech comprehension in conversation, which is facilitated by keeping track of different speech rate cues. The distance of speech rate cues to an ambiguous word plays a part in how listeners

process them. Current models of word recognition have not yet implemented explicit mechanisms that can account for both distal and global speech rate effects. This thesis suggests that several components are required in order to account for all speech rate effects. First, there must be an online acoustic tracking mechanism that estimates phonetic information in the surrounding sentence context. Second, memory has to be invoked for the global rate effect to influence speech recognition, because the global rate effect depends on talker identity. The online acoustic tracking component may involve statistical learning that integrates cues with memorized information about previous rate cues and talker voice. These principles may be implemented in a neural entrainment account. Future work may explore the neurobiological bases of the proposed mechanisms for distal and global speech rate processing in more detail.

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. doi: 10.1121/1.1779271

Alexandrou, A. M., Saarinen, T., Kujala, J., & Salmelin, R. (2018). Cortical tracking of global and local variations of speech rhythm during connected natural speech perception. *Journal of Cognitive Neuroscience*, *30*(11), 1704–1719. doi: 10.1162/jocn_a _01295

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. doi: 10.1121/1.1701898

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439. doi: 10.1006/jmla.1997.2558

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. doi: 10.1016/S0010-0277(99)00059-1

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi: 10.1016/j.jml.2007.12.005

Baese-Berk, M. M., Dilley, L. C., Henry, M. J., Vinke, L., & 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. doi: 10.3758/s13414-018-1626-4

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. doi: 10.1177/0956797614533705

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory*

and Language, *68*(3), 255–278. doi: 10.1016/j.jml.2012.11.001

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., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: Behavioural evidence from rate normalisation. *Language, Cognition and Neuroscience*, *33*(8), 955–967. doi: 10.1080/23273798.2018.1439179

Bosker, H. R., & Reinisch, E. (2015). Normalization for speech rate in native and nonnative speech. In *18th International Congress of Phonetic Sciences 2015 [ICPhS XVIII]*.

Bosker, H. R., & Reinisch, E. (2017). Foreign languages sound fast: Evidence from implicit rate normalization. *Frontiers in Psychology*, *8*, 1063. doi: 10.3389/fpsyg.2017.01063

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

Brehm, L., & Goldrick, M. (2017). Distinguishing discrete and gradient category structure in language: Insights from verb-particle constructions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(10), 1537–1556. doi: 10.1037/xlm0000390

Brown, M., Dilley, L. C., & Tanenhaus, M. K. (2012). Real-time expectations based on context speech rate can cause words to appear or disappear. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34).

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

Cho, S.-J., Brown-Schmidt, S., & Lee, W.-y. (2018). Autoregressive generalized linear mixed effect models with crossed random effects: An application to intensive binary time series eye-tracking data. *Psychometrika*, *83*(3),

751–771. doi: 10.1007/s11336-018-9604-2

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. doi: 10.1016/j.cognition.2008.04.004

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. doi: 10.3758/s13414-015-0855-z

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. doi: 10.1080/01690965.2011.610597

Creel, S. C., & Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass*, *5*(5), 190–204. doi: 10.1111/j.1749-818X.2011.00276.x

Cunningham, S. J., Turk, D. J., Macdonald, L. M., & Macrae, C. N. (2008). Yours or mine? Ownership and memory. *Consciousness and Cognition*, *17*(1), 312–318. doi: 10.1016/j.concog.2007.04.003

de Brouwer, A. J., Smeets, J. B., & Plaisier, M. A. (2016). How heavy is an illusory length? *i-Perception*, *7*(5), 1–5. doi: 10.1177/2041669516669155

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. doi: 10.1016/j.cognition.2008.07.008

Dent, M. L., Brittan-Powell, E. F., Dooling, R. J., & Pierce, A. (1997). Perception of synthetic/ba/–/wa/speech continuum by budgerigars (Melopsittacus undulatus). *The Journal of the Acoustical Society of America*, *102*(3), 1891–1897. doi: 10.1121/1.420111

Diehl, R. L., Souther, A. F., & Convis, C. L. (1980). Conditions on rate normalization in speech perception. *Perception & Psychophysics*, *27*(5), 435–443. doi: 10.3758/BF03204461

Diehl, R. L., & Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *The Journal of the Acoustical Society of America*, *85*(5), 2154–2164. doi: 10.1121/1.397864

Dilley, L. C., Morrill, T. H., & Banzina, E. (2013). New tests of the distal speech rate effect: Examining cross-linguistic generalization. *Frontiers in Psychology*, *4*, 1002.

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words

to appear or disappear. *Psychological Science*, *21*(11), 1664–1670. doi: 10.1177/0956797610384743

Dink, J. W., & Ferguson, B. (2015). *eyetrackingR: An R library for eye-tracking data analysis*. Retrieved from `http://www.eyetrackingr.com/`

Eger, N. A., & Reinisch, E. (2019). The impact of one's own voice and production skills on word recognition in a second language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(3), 552–571. doi: 10.1037/xlm0000599

Eimas, P. D., & Miller, J. L. (1980). Contextual effects in infant speech perception. *Science*, *209*(4461), 1140–1141. doi: 10.1126/science.7403875

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Attention, Perception, & Psychophysics*, *67*(2), 224–238. doi: 10.3758/BF03206487

Forbach, G. B., Stanners, R. F., & Hochhaus, L. (1974). Repetition and practice effects in a lexical decision task. *Memory & Cognition*, *2*(2), 337–339. doi: 10.3758/BF03209005

Forrin, N. D., & MacLeod, C. M. (2018). This time it's personal: The memory benefit of hearing oneself. *Memory*, *26*(4), 574–579. doi: 10.1080/09658211.2017.1383434

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680–698. doi: 10.1037/0278-7393.10.4.680

Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: Intelligibility of speech with a manipulated modulation spectrum. *Frontiers in Psychology*, *3*, 238. doi: 10.3389/fpsyg.2012.00238

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. doi: 10.1038/nn.3063

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. doi: 10.1037/0033-295X.105.2.251

Gordon, P. C. (1988). Induction of rate-dependent processing by coarse-grained aspects of speech. *Perception & Psychophysics*, *43*(2), 137–146. doi: 10.3758/BF03214191

Gordon, P. C., & Scearce, K. A. (1995). Pronominalization and discourse coher-

ence, discourse structure and pronoun interpretation. *Memory & Cognition*, *23*(3), 313–323. doi: 10.3758/BF03197233

Graux, J., Gomot, M., Roux, S., Bonnet-Brilhault, F., Camus, V., & Bruneau, N. (2013). My voice or yours? An electrophysiological study. *Brain Topography*, *26*(1), 72–82. doi: 10.1007/s10548-012-0233-2

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, *39*(5), 350–365. doi: 10.1016/j.jcomdis.2006.06.013

Heffner, C. C., Newman, R. S., Dilley, L. C., & 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. doi: 10.1044/2015_JSLHR-H-14-0239

Heffner, C. C., Newman, R. S., & Idsardi, W. J. (2017). Support for context effects on segmentation and segments depends on the context. *Attention, Perception, & Psychophysics*, *79*(3), 964–988. doi: 10.3758/s13414-016-1274-5

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, *5*, 82. doi: 10.3389/fnhum.2011.00082

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: An MEG study. *Journal of Cognitive Neuroscience*, *14*(8), 1125–1138. doi: 10.1162/089892902760807140

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*(4), 460–482. doi: 10.1016/j.jml.2007.02.001

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171. doi: 10.1016/j.actpsy.2010.11.003

Jacewicz, E., Fox, R. A., & 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. doi: 10.1121/1.3459842

Jardri, R., Pins, D., Bubrovszky, M., Despretz, P., Pruvo, J.-P., Steinling, M., & Thomas, P. (2007). Self awareness and speech processing: An fMRI study. *Neuroimage*, *35*(4), 1645–1653. doi: 10.1016/j.neuroimage.2007.02.002

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. doi: 10.3758/BF03193320

Kaufeld, G., Ravenschlag, A., Meyer, A. S., Martin, A. E., & Bosker, H. R. (2019). Knowledge-based and signal-based cues are weighted flexibly during spoken language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. (Advance online publication) doi: 10.1037/xlm0000744

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. doi: 10.3758/BRM.42.3.643

Keyes, H., & Brady, N. (2010). Self-face recognition is characterized by "bilateral gain" and by faster, more accurate performance which persists when faces are inverted. *Quarterly Journal of Experimental Psychology*, *63*(5), 840–847. doi: 10.1080/17470211003611264

Kidd, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4), 736–748.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. doi: 10.1016/S0010-0277(02)00004-5

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. doi: 10.1037/a0038695

Kluender, K. R. (1984). Physical acoustic information for magnitude estimation of rate of speech. *The Journal of the Acoustical Society of America*, *75*(S1), S41–S41. doi: 10.1121/1.2021425

Knoblich, G., & Flach, R. (2001). Predicting the effects of actions: Interactions of perception and action. *Psychological Science*, *12*(6), 467–472. doi: 10.1111/1467-9280.00387

Knoblich, G., Seigerschmidt, E., Flach, R., & Prinz, W. (2002). Authorship effects in the prediction of handwriting strokes: Evidence for action simulation during action perception. *The Quarterly Journal of Experimental Psychology*, *55A*(3), 1027–1046. doi: 10.1080/02724980143000631

Koreman, J. (2006). Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *The Journal of the Acoustical Society of America*, *119*(1), 582–596. doi: 10.1121/1.2133436

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., & Hagoort, P. (2018). Neural entrainment determines the words we hear. *Current Biology*, *28*(18), 2867–2875. doi: 10.1016/j.cub.2018.07.023

Lai, W., & Dilley, L. (2016). Cross-linguistic generalization of the distal rate effect: Speech rate in context affects whether listeners hear a function word in Chinese Mandarin. In *Proceedings of Speech Prosody* (Vol. 8, pp. 1124–1128). doi: 10.21437/SpeechProsody.2016-231

Marian, V., Bartolotti, J., Chabal, S., & 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., & 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. doi: 10.3389/fpsyg.2016.00120

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2018). Listening to yourself is special: Evidence from global speech rate tracking. *PloS one*, *13*(9), e0203571. doi: 10.1371/journal.pone.0203571

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019a). 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

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019b). Listeners normalize speech for contextual speech rate even without an explicit recognition task. *The Journal of the Acoustical Society of America*, *146*(1), 179–188. doi: 10.1121/1.5116004

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134. doi: 10.1111/j.1467-7687.2007.00653.x

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. doi: 10.3758/BF03193773

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. doi: 10.1038/264746a0

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126. doi: 10.1207/s15516709cog0000_79

McQueen, J. M., & Viebahn, M. C. (2007). Tracking recognition of spoken words by tracking looks to printed words. *The Quarterly Journal of Experimental Psychology*, *60*(5), 661–671. doi: 10.1080/17470210601183890

Miller, J. L. (1981). Some effects of speaking rate on phonetic perception. *Phonetica*, *38*(1–3), 159–180. doi: 10.1159/000260021

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(94)90031-0

Miller, J. L., & 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. doi: 10.1121/1.389399

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. doi: 10.1159/000261728

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, *25*(6), 457–465. doi: 10.3758/BF03213823

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. doi: 10.5334/labphon.66

Monsell, S., Patterson, K. E., Graham, A., Hughes, C. H., & 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

Morrill, T. H., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2014). Distal rhythm influences whether or not listeners hear a word in continuous speech: Support for a perceptual grouping hypothesis. *Cognition*, *131*(1), 69–74. doi: 10.1016/j.cognition.2013.12.006

Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception & Psychophysics*, *58*(4), 540–560. doi: 10.3758/BF03213089

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. doi: 10.1016/j.wocn.2008.09.001

Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. *Journal of Neuroscience*, *33*(41), 16110–16116. doi: 10.1523/JNEUROSCI.2137-13.2013

Nooteboom, S. G. (1981). Speech rate and segmental perception or the role of words in phoneme identification. In *Advances in psychology* (Vol. 7, pp. 143–150). Elsevier. doi: 10.1016/S0166-4115(08)60188-0

O'Dell, M., & Nieminen, T. (2018). Distal rate effect for finnish epenthetic vowels. In *Proc. 9th International Conference on Speech Prosody 2018* (pp. 646–650). doi: 10.21437/SpeechProsody.2018-131

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, *3*, 320. doi: 10.3389/fpsyg.2012.00320

Pickett, J., & Decker, L. R. (1960). Time factors in perception of a double consonant. *Language and Speech*, *3*(1), 11–17. doi: 10.1177/002383096000300103

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. doi: 10.1016/0167-6393(93)90063-Q

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. doi: 10.3758/s13414-015-0981-7

Pufahl, A., & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology*, *70*, 1–30. doi: 10.1016/j.cogpsych.2014.01.001

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. doi: 10.1121/1.2821762

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. (2016a). Natural fast speech is perceived as faster than linearly time-compressed speech. *Attention, Perception, & Psychophysics*, *78*(4), 1203–1217. doi: 10.3758/s13414-016-1067-x

Reinisch, E. (2016b). Speaker-specific processing and local context information: The case of speaking rate. *Applied Psycholinguistics*, *37*(6), 1397–1415. doi: 10.1017/S0142716415000612

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), 978–996. doi: 10.1037/a0021923

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. doi: 10.1016/j.wocn.2013.01.002

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. doi: 10.1126/science.274.5294.1926

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. doi: 10.1016/S0010-0277(98)00075-4

Sawusch, J. R., & Newman, R. S. (2000). Perceptual normalization for speaking rate II: Effects of signal discontinuities. *Perception & Psychophysics*, *62*(2), 285–300. doi: 10.3758/BF03205549

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human perception and performance*, *3*(1), 1–17. doi: 10.1037/0096-1523.3.1.1

Schachtenhaufen, R. (2010). Schwa-assimilation og stavelsesgrænser. *NyS, Nydanske Sprogstudier*(39), 64–92.

Schuerman, W. L. (2017). *Sensorimotor experience in speech perception* [Doctoral dissertation]. Radboud University, Nijmegen.

Schuerman, W. L., Meyer, A., & McQueen, J. M. (2015). Do we perceive others better than ourselves? A perceptual benefit for noise-vocoded speech produced by an average speaker. *PloS ONE*, *10*(7), e0129731. doi: 10.1371/journal.pone.0129731

Schwab, S. (2011). Relationship between speech rate perceived and produced by the listener. *Phonetica*, *68*(4), 243–255. doi: 10.1159/000335578

Seedorff, M., Oleson, J., & McMurray, B. (2018). Detecting when timeseries dif-

fer: Using the Bootstrapped Differences of Timeseries (BDOTS) to analyze Visual World Paradigm data (and more). *Journal of Memory and Language*, *102*, 55–67. doi: 10.1016/j.jml.2018.05.004

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(3), 243–261. doi: 10.1017/S0142716400001417

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention, Perception, & Psychophysics*, *73*(4), 1195–1215. doi: 10.3758/s13414-011-0096-8

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... others (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. doi: 10.1073/pnas.0903616106

Sui, J., He, X., & Humphreys, G. W. (2012). Perceptual effects of social salience: Evidence from self-prioritization effects on perceptual matching. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1105. doi: 10.1037/a0029792

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074–1095. doi: 10.1037/0096-1523.7.5.1074

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. doi: 10.1126/science.7777863

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. doi: 10.1121/1.3106131

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284–1301. doi: 10.3758/s13414-012-0306-z

Toscano, J. C., & 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. doi: 10.1080/23273798.2014.946427

Treille, A., Vilain, C., Kandel, S., Schwartz, J.-L., & Sato, M. (2015). Speech

in the mirror? Neurobiological correlates of self speech perception. In *Seventh Annual Meeting of the Society for the Neurobiology of Language* (pp. 220–221).

Truong, G., Roberts, K. H., & Todd, R. M. (2017). I saw mine first: A prior-entry effect for newly acquired ownership. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 192–205. doi: 10.1037/xhp0000295

Truong, G., & Todd, R. M. (2017). SOAP opera: Self as object and agent in prioritizing attention. *Journal of Cognitive Neuroscience*, *29*(6), 937–952. doi: 10.1162/jocn_a_01083

Turk, D. J., Van Bussel, K., Brebner, J. L., Toma, A. S., Krigolson, O., & Handy, T. C. (2011). When "it" becomes "mine": Attentional biases triggered by object ownership. *Journal of Cognitive Neuroscience*, *23*(12), 3725–3733. doi: 10.1162/jocn_a _00101

Ventura, M. I., Nagarajan, S. S., & Houde, J. F. (2009). Speech target modulates speaking induced suppression in auditory cortex. *BMC Neuroscience*, *10*(1), 58. doi: 10.1186/1471-2202-10-58

Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics*, *67*(6), 939–950. doi: 10.3758/BF03193621

Welch, T. E., Sawusch, J. R., & Dent, M. L. (2009). Effects of syllable-final segment duration on the identification of synthetic speech continua by birds and humans. *The Journal of the Acoustical Society of America*, *126*(5), 2779–2787. doi: 10.1121/1.3212923

Xu, M., Homae, F., Hashimoto, R.-i., & Hagiwara, H. (2013). Acoustic cues for the recognition of self-voice and other-voice. *Frontiers in Psychology*, *4*. doi: 10.3389/fpsyg.2013.00735

Nederlandse samenvatting

Iedereen kent wel iemand die erg snel of erg langzaam spreekt. De snelheid waarop mensen spreken verschilt echter niet alleen tussen mensen; mensen variëren individueel ook in de snelheid waarop ze spreken, afhankelijk van bijvoorbeeld hun gesprekspartner of de situatie waarin ze verkeren. Zo is men in een luidruchtige omgeving of tegen een tweedetaalspreker geneigd langzamer te spreken dan in een stille omgeving of tegen een moedertaalspreker. We hebben als luisteraars geleerd ons aan te passen aan al die variatie in spreeksnelheid. Toch kan de grote variatie ook problemen opleveren. Dat komt omdat de duur van spraakklanken betekenisonderscheidend is. Zo heeft de klinker in "mat" een kortere duur dan de klinker in "maat" en dit is een belangrijk verschil; als de klinkers even lang zouden zijn, zouden de twee moeilijk uit elkaar te houden zijn. Er zijn dus twee typen duur waar een luisteraar tegelijkertijd rekening mee moet houden: die van de spraakklanken zelf, en de snelheid waarop de spreekcontext wordt uitgesproken.

De vraag die dit proefschrift bestudeert, is hoe luisteraars de variatie in spreeksnelheid gebruiken om te beoordelen of een spraakklank kort of lang is. Oftewel: hoe bepaalt de luisteraar of de spreker het heeft over een "mat" of over een "maat"? Om deze vraag te kunnen beantwoorden, zijn in dit proefschrift de invloeden van twee soorten spreeksnelheidscontext getoetst, namelijk de *distale* context en *globale* context. Wat deze soorten contexten omvatten, kan worden uitgelegd aan de hand van het volgende voorbeeld:

> **Spreker A**: Welk woord heb je net tegen mij gezegd?
> **Spreker B**: *Ik heb zojuist het* woord [mat/maat] ge*zegd*.

In dit voorbeeld is de schuingedrukte spraak van Spreker B de distale context van het ambigue woord "mat"/"maat": het is de directe zinscontext, afgezien van de twee lettergrepen aangrenzend aan het ambigue woord. De spraak van Spreker A is de globale context, nog verder van het ambigue woord verwijderd en gesproken door een andere spreker. Distaal betekent hier dus 'verder ver-

wijderd', terwijl globaal betrekking heeft op de volledige context. Afhankelijk van de snelheden waarmee de distale en de globale contexten zijn uitgesproken, kan de luisteraar zowel "mat" als "maat" verstaan. Luisteraars 'normaliseren' dus voor spreeksnelheid; ze bepalen aan de hand van de duur van de klinker én de spreeksnelheid van de context of een klinker als kort of lang is bedoeld.

Dit proefschrift heeft een aantal bevindingen gedaan over in hoeverre de distale en de globale spreekcontext effecten van spreeksnelheidsnormalisatie teweeg brengen. De eerste belangrijke bevinding die werd gedaan is dat luisteraars de distale spreekcontext automatisch verwerken (zie *Hoofdstuk 2*). Dat wil zeggen dat luisteraars de snelheid van de distale spreekcontext onbewust gebruiken om te bepalen of ze een kort of lang woord hebben gehoord. Daarom is het waarschijnlijk dat luisteraars ook in hun dagelijkse communicatie spraak normaliseren voor de snelheid waarop het is uitgesproken. De distale spreekcontext heeft daarbij een contrastief effect op spraakperceptie. De duur van een spraakklank zoals die wordt waargenomen is namelijk relatief aan de spreeksnelheid van de distale context: een klank tussen de korte "a" in "mat" en de lange "aa" in "maat" wordt in snelle spraak eerder waargenomen als lang (= "aa"), omdat de klinker *relatief lang* klinkt. In langzame spraak wordt dezelfde klank echter eerder waargenomen als kort (= "a"), omdat de klinker relatief kort klinkt in een langzame context.

De tweede belangrijke bevinding in dit proefschrift is dat, naast het effect van de distale spreekcontext, er ook een effect optreedt van de globale spreeksnelheid. Om het effect van globale spreeksnelheid te meten, werden in *Hoofdstuk 3*, *Experiment 2*, twee groepen met elkaar vergeleken. Een hoge-snelheidsgroep luisterde naar Spreker A met een snelle spreeksnelheid en Spreker B met een neutrale spreeksnelheid. De andere lage-snelheidsgroep luisterde ook naar Spreker B met een neutrale snelheid, maar naar Spreker A met een langzame spreeksnelheid. In de zinnen van Sprekers A en B waren ambigue woorden opgenomen met het klinkercontrast "a" en "aa", die verschillen in duur. De perceptie van de neutrale spraak van Spreker B werd tussen de twee groepen vergeleken. In de hoge-snelheidsgroep hoorden proefpersonen minder lange "aa"-klanken in de neutrale spraak van Spreker B dan in de lage-snelheidsgroep. Dat wil zeggen dat Spreker B relatief langzaam klonk, wanneer Spreker A sneller sprak, en omgekeerd dat Spreker B relatief snel klonk, wanneer Spreker A langzamer sprak. Concreet betekent dit dat de luisteraar geneigd is het ambigue woord te horen als kort (= "mat") in de spraak van Spreker B, wanneer Spreker A sneller spreekt. Net als het distale spreeksnelheidseffect is het globale effect dus contrastief. Dit

effect kan worden vergeleken met hoe je rijsnelheid kunt ervaren. Wanneer je net van de snelweg komt, kan het rijden op een 60-kilometerweg langzaam aandoen, maar wanneer je net van een grindweg afkomt waar je maar 30 km per uur mag, kan het rijden op dezelfde 60-kilometerweg juist als snel aanvoelen.

Dit proefschrift vond ook dat het globale spreeksnelheidseffect aan meer factoren onderhevig is dan het distale spreeksnelheidseffect. Er werden drie verschillen vastgesteld tussen de twee effecten. Ten eerste toonde *Hoofdstuk 3, Experiment 3,* aan dat effecten van globale spreeksnelheid alleen optreden wanneer er niet te veel spreeksnelheidsvariatie is binnen sprekers. Dat wil zeggen dat het globale effect sprekerspecifiek is, omdat het iets uitmaakt welke spreker langzaam of snel sprak, waar het distale effect onafhankelijk is van sprekers.

Ten tweede vond *Hoofdstuk 4* dat de spreeksnelheid van de luisteraar zelf geen globaal spreeksnelheidseffect teweeg brengt. In *Hoofdstuk 4* werden net als in *Hoofdstuk 3* twee groepen met elkaar vergeleken. Echter, deze keer werd snelle/langzame Spreker A vervangen door de proefpersoon zelf. Zodoende luisterde de hoge-snelheidsgroep nog steeds naar neutrale spraak van Spreker B, maar spraken de proefpersonen tussen het luisteren door zelf zinnen uit op hoge snelheid. De lage-snelheidsgroep luisterde ook naar neutrale Spreker B, maar sprak zelf op lage snelheid. De groepen werden weer vergeleken in hun perceptie van de ambigue woorden in de spraak van Spreker B. Dit keer was er geen verschil tussen de groepen in hoeveel korte en lange klinkers ze hoorden. Een volgend experiment liet zien dat dit ook gold wanneer proefpersonen zichzelf alleen *hoorden* (van opnames). Eigen spraak brengt dus geen globaal spreeksnelheidseffect teweeg. Hierin verschilt het globale spreeksnelheidseffect van het effect van de distale context, waar eerder onderzoek aantoonde dat eigen spraak in de distale context wél een effect heeft op de perceptie van een andere spreker.

Ten derde liet *Hoofdstuk 5* zien dat het distale spreeksnelheidseffect en het globale spreeksnelheidseffect verschillende tijdlijnen hebben. Dit werd vastgesteld met een eye-trackingexperiment, waarbij de oogbewegingen van proefpersonen werden gemeten. Afhankelijk van wanneer de ogen fixeren op de ene of de andere antwoordoptie op een scherm, kan namelijk worden bepaald wanneer proefpersonen een beslissing maakten over welk woord ze hadden gehoord in een bepaalde zin. Het distale spreeksnelheidseffect werd waargenomen na ongeveer 300 ms na afloop van een ambigue klinker, terwijl het globale spreeksnelheidseffect pas werd waargenomen na ongeveer 570 ms na afloop van de

klinker. Luisteraars hebben dus meer tijd nodig om de globale spreekcontext te verwerken dan de zinscontext.

Hoofdstuk 6 vatte de bevindingen van het proefschrift samen en verbond de resultaten aan wat al bekend was over spreeksnelheidsnormalisatie en spraakperceptie. Samenvattend liet dit proefschrift zien dat de spreeksnelheidscontext waarin woorden worden uitgesproken (en wie ze uitspreekt) de manier waarop spraak wordt waargenomen systematisch kan veranderen. De resultaten suggereren dat spreeksnelheidseffecten plaatsvinden in dagelijkse communicatie, omdat natuurlijke snelle en langzame spraak spreeksnelheidseffecten teweeg brachten en omdat de effecten optraden zonder dat er aandacht werd gevestigd op de ambigue woorden. Deze bevindingen benadrukken de complexiteit van spraakperceptie, hetgeen vergemakkelijkt wordt doordat de luisteraar de spreeksnelheden van sprekers bijhoudt. De afstand van de spreekcontext (distaal of globaal) tot een ambigu woord speelt ook een rol in hoe luisteraars die context verwerken. Huidige modellen van woordherkenning hebben nog geen expliciete mechanismen geïmplementeerd die zowel distale als globale spreeksnelheidseffecten kunnen verklaren. Dit proefschrift suggereert dat verschillende componenten nodig zijn om alle spreeksnelheidseffecten te verklaren. Ten eerste moet er een mechanisme zijn dat online de fonetische informatie in de context van de omringende zinnen schat; wordt er snel of wordt er langzaam gesproken? Ten tweede moet een beroep gedaan worden op het geheugen voordat het globale snelheidseffect spraakherkenning kan beïnvloeden, omdat het globale snelheidseffect afhangt van de identiteit van de spreker. Het eerste mechanisme behelst mogelijk statistisch leren dat informatie integreert met reeds opgeslagen informatie over spreeksnelheid en spreekstem. Deze principes kunnen worden geïmplementeerd in een model waarin de hersengolven zich voegen naar het inkomende spraaksignaal. Toekomstig onderzoek zou zich in meer detail kunnen richten op de neurobiologische grondslagen van de in dit proefschrift voorgestelde mechanismen voor distale en globale spreeksnelheidsverwerking.

Dansk sammenfatning

Alle kender vel én, som enten taler meget hurtigt eller meget langsomt. Det tempo, man taler i, varierer imidlertid ikke kun fra person til person: En person skifter også selv mellem forskellige taletempi, afhængig af hvem personen taler med eller den situation, som personen befinder sig i. For eksempel har de fleste en tendens til at snakke langsommere, når der er meget støj i baggrunden eller når de taler med en person der har et andet modersmål. Som lyttere har vi lært at tilpasse os variation i andre personers taletempo. Dog kan den hyppige variation i taletempi skabe problemer, da det påvirker sproglydenes længde, og lydlængde kan være betydningsadskillende; vokalen i "kulde" (/kulə/) er for eksempel kortere end vokalen i "kugle" (/ku:lə/), og i dette eksempel er det vokalens længde, der er afgørende for, at ordene kan skelnes fra hinanden. Således er der to varighedskarakteristika, som har betydning for, hvordan man som lytter opfatter ord: Sproglydenes længde og det tempo de udtales i.

Spørgsmålet, som denne doktorafhandling behandler, er hvordan personer, der lytter til tale, bruger variationen i taletempoet i deres bedømmelse af sproglydenes længde. Hvordan bestemmer lytteren for eksempel, om taleren snakker om "en gul kæde" eller "en guldkæde"? For at kunne besvare dette spørgsmål, har denne afhandling testet hvordan to forskellige typer taletempo påvirker lytteres opfattelse af sproglyde, nemlig den fjerne konteksts tempo og den globale konteksts tempo. Hvad disse to typer af kontekst inkluderer, illustreres med følgende eksempel:

Taler A: Hvilket ord var det du sagde til mig?
Taler B: *Jeg sagde ord*et [kuld/kugl]e.

I dette eksempel er Taler B's kursiverede tale den fjerne kontekst af den tvetydige vokal kort/langt "u" i ordet "kulde"/"kugle": Det er sætningskonteksten, foruden de to stavelser som grænser op til den tvetydige stavelse. Taler A's tale er den globale kontekst, fjernet yderligere fra den tvetydige vokal og talt af en anden person. Her refererer "global" altså til hele konteksten. Afhængig af

de fjerne og de globale konteksters tempi, kan lytteren høre enten "kulde" eller "kugle". Det vil sige at lyttere "normaliserer" taletempi; de bestemmer på baggrund af både vokalens længde og kontekstens taletempi, om vokalen må være kort eller lang.

I denne afhandling er en række fund om, i hvilket omfang den fjerne og den globale talekontekst frembringer effekter af taletemponormalisering. Det første vigtige fund er, at lyttere automatisk processerer den fjerne talekontekst (se *kapitel 2*). Det vil sige, at lyttere ubevidst bruger tempoet i den fjerne talekontekst til at afgøre, om de har hørt en kort eller en lang vokal i et ord, og dermed ordets betydning. Derfor er det sandsynligt at lyttere også normaliserer tale for det tempo, der tales i i daglig kommunikation. Den fjerne talekontekst pavirker taleperception på en *kontrastiv* måde. Det betyder, at en sproglyds længde opfattes relativt til taletempoet: En tvetydig sproglyd midt imellem kort hollandsk "a" og langt hollandsk "aa", som er den vokalkontrast denne afhandling har undersøgt, opfattes oftest som langt "aa" i hurtig tale, fordi vokalen i dette tilfælde lyder *relativt lang*. I langsom tale opfattes den samme sproglyd imidlertid som kort "a", fordi vokalen lyder relativt kort i en langsom kontekst.

Det andet vigtige fund i denne afhandling er, at der ud over effekten af den fjerne talekontekst også er en effekt af det globale taletempo. For at måle denne effekt, blev der i *kapitel 3, eksperiment 2,* sammenlignet to forskellige lyttergrupper: 1) en højt-tempogruppe, der lyttede til Taler A med et hurtigt taletempo og Taler B med et neutralt taletempo, og 2) en lavt-tempogruppe, som også lyttede til Taler B med et neutralt tempo, mens Taler A her talte langsomt. I sætningerne fra talere A og B blev der inkluderet tvetydige ord med den hollandske vokalkontrast mellem kort "a" og langt "aa", som adskiller sig i længde. Opfattelsen af Taler B's neutrale tale blev sammenlignet mellem de to grupper. I højt-tempogruppen hørte testpersonerne færre lange "aa"-vokaler i den neutrale tale fra Taler B i forhold til lavt-tempogruppen. Det vil sige, at Taler B lød relativt langsom, når Taler A talte hurtigere, og omvendt lød Taler B relativt hurtig, når Taler A talte langsommere. Konkret betyder dette, at lyttere har en tendens til at høre tvetydige ord som korte i Taler B's tale, når Taler A taler hurtigere. Ligesom det fjerne taletempo, har det globale taletempo altså også en kontrastiv effekt. Den globale effekt kan sammenlignes med, hvordan man oplever kørehastighed. Når man lige er kommet af motorvejen, kan det føles langsomt at køre 60 $^{km}/_{t}$, men hvis man lige er kommet fra en grusvej, hvor man kun må køre 30 $^{km}/_{t}$, kan kørsel på den samme vej med 60 $^{km}/_{t}$ føles som meget hurtig.

Denne afhandling fandt også, at den globale taletempoeffekt er underlagt flere faktorer end den fjerne taletempoeffekt. Der blev fundet tre forskelle mellem de to effekter. For det første blev det demonstreret (*kapitel 3*, *eksperiment 3*), at effekter af globalt taletempo kun opstår, når der ikke er for meget variation i den individuelle talers taletempo. Det vil sige, at den globale effekt er talerspecifik, fordi det betyder noget, hvem der talte langsomt eller hurtigt, hvor den fjerne effekt ikke er betinget af, hvem der talte i hvilket tempo.

For det andet blev det demonstreret (*kapitel 4*), at lytteres egne taletempi ikke frembringer en global taletempoeffekt. Dette blev igen vist ved at sammenholde to lyttergrupper. Denne gang blev hurtigt/langsomt talende Taler A erstattet af testpersonens egen stemme. Højt-tempogruppen lyttede således stadig til Taler B's neutrale tale, men mellem disse sætninger udtalte testpersonerne selv sætninger i et højt tempo. Lavt-tempogruppen lyttede også til neutral Taler B, men talte selv i et lavt tempo. De to gruppers opfattelse af de tvetydige ord i Taler B's tale blev igen sammenlignet. Denne gang var der ingen forskel mellem grupperne i forhold til hvor mange korte og lange vokaler de hørte. Et andet eksperiment viste, at dette også gjaldt, når testpersonerne kun *hørte* sig selv (fra optagelser). Ens egen tale frembringer altså ingen global taletempoeffekt. Heri adskiller den globale taletempoeffekt sig fra den fjerne kontekssteffekt, da tidligere forskning har vist, at ens egen tale i den fjerne kontekst påvirker perceptionen af en anden taler.

For det tredje blev det i *kapitel 5* vist, at den fjerne taletempoeffekt og den globale taletempoeffekt har forskellige tidslinjer. Dette blev fastslået med et øjesporingseksperiment, hvor testpersoners øjenbevægelser blev målt. Ud fra hvornår deres øjne fæstnede sig på den ene eller den anden svarmulighed på skærmen, kan det afgøres, hvornår testpersonerne tog en beslutning om, hvilket ord de havde hørt i en bestemt sætning. Den fjerne taletempoeffekt blev observeret ca. 300 ms efter den tvetydige vokal, mens den globale taletempoeffekt først blev observeret ca. 570 ms efter afslutning af vokalen. Det vil sige, at lyttere har brug for mere tid til at processere den globale talekontekst end sætningskonteksten.

I *kapitel 6* drøftes afhandlingens resultater i forhold til den viden, der eksisterer om normalisering af taletempo, og hvordan det påvirker taleperception. Overordnet har denne afhandling vist, at det taletempo som ord udtales i (og hvem der udtaler dem) systematisk kan ændre måden sproglyde opfattes på. Resultaterne antyder, at effekter af taletempo optræder i daglig kommunikation, både fordi naturlig hurtig og langsom tale fremkalder disse effekter, men også fordi effekterne opstod, uden at lytteres opmærksomhed var rettet mod de

tvetydige ord. Disse fund understreger kompleksiteten af taleperception, hvilket lettes ved, at lyttere holder styr på forskellige taleres taletempi. Ydermere spiller det en rolle i lytteres taleperception af tvetydige sproglyde, om talekonteksten er fjern eller global. Nuværende ordgenkendelsesmodeller har endnu ikke implementeret eksplicitte mekanismer, der kan forklare både fjerne og globale taletempoeffekter. Denne doktorafhandling antyder, at forskellige komponenter er nødvendige for at forklare alle taletempoeffekter. For det første bør der være en mekanisme, der estimerer den fonetiske information i de omgivende ytringer i realtid; tales der hurtigt eller langsomt? For det andet skal informationer om taletempo og talestemme lagres i hukommelsen, før det globale taletempo kan påvirke ordgenkendelsen, da effekten af det globale taletempo afhænger af talerens identitet. Den første mekanisme omfatter muligvis statistisk læring, som integrerer nye informationer med forudgående informationer om taletempo og talestemme, som allerede er gemt i hukommelsen. Disse principper kan implementeres i en model, hvor hjernen indstilles til frekvenser i det indgående talesignal. Fremtidig forskning bør fokusere mere detaljeret på de neurobiologiske faktorer i mekanismerne, der processerer fjerne og globale taletempi, som foreslået i denne afhandling.

Acknowledgements

It is finished. Tout est accompli. Es ist vollbracht.

Before starting my PhD project, I was told that it was going to be tough, nerve-racking, and lonely. Now, four years later, I have learned that doing a PhD does not have to be any of these things. This is, of course, in large part due to the guidance and support from everybody around me, both professionally and personally.

A great thanks to my promotor and supervisor Antje, whose clear and direct feedback and strategic plans of action I could not have gone without. An even greater thanks to my other supervisor and motivating force Hans Rutger, who fueled me with new energy and insights in every single meeting we had. His ability to do so makes me doubt that there are or ever will be better supervisors than him.

My brilliant paranymphs, Jonathan and Saoradh, thank you for our friendship and for the drinks and dinners (and conferences) we shared throughout the last couple of years. Your company is one of the reasons I have occasionally regretted not living in Nijmegen during my time at the MPI. Joe, your giggles make up for all the mess you continuously made on my desk and my shelves, forgetting that those were mine and not yours. Saoradh, your calmness is admirable and contagious. Do you remember when when we first met in my office? After that bewildering moment, things could only get better. I will truly miss both of you.

To the Technical Group, whose members have helped me out many times when I had problems with electronic devices or software, thank you. I would specifically like to thank Johan for his help with a couple of online pilot experiments and Maarten who has helped me with Presentation on many occasions. I am also grateful to Operations who helped with everything other than electronic devices, and I am in particular grateful to Angela and Anique, who were always kind and supportive regarding PhD issues. Thanks to the Library staff, Karin and Meggie, for hard-to-find papers and guidelines for publishing.

To the IMPRS, first Els and Dirkje, later Kevin, thank you for the educational opportunities you have given me. Kevin, you were always up for a quick chat,

even though you had to run the IMPRS all by yourself. I really appreciate your engagement and commitment. The same goes for my fellow PhD reps Limor, Julia, Merel, and Merel. It has been truly great meeting you and brainstorming with you and you have made my time at the MPI all the more fun.

I also want to thank everybody from the Psychology of Language department. Especially Alastair, for his comforting words before my very first talk, and Amie for showing me around when I first arrived at the MPI and pointing out to me what a great place it was. Annelies, Eirini, Greta, Laurel, Markus, Miguel, Nina, and Renske, thank you for all the nice chats we had at and outside work about other things than work (and work). Also thanks to Ashley, Annika, Lisa, and Elliott, with whom I have also shared many good times.

Being at work had been very different (and perhaps difficult) without having friends there too. A few people from the institute have become very dear to me. Special thanks are due to my first two office mates, William and Johanne, who made me feel so very welcome when I first arrived. Knowing that they did not wish for a 'third member', I think they did an outstanding job enduring me. Will, thanks for making me coffee every morning. I am happy that I did not lose you when you moved back to the other side of the world. Johanne, your loud and cheerful appearance has made me laugh so many times, and I still miss our fictional podcast about PhD life in Nijmegen. A genius podcast that was. And Suzanne, although we never shared an office, you belong to this paragraph too. Thanks for our regular laughs and your hands-on advice about almost everything that can happen and must be done in life.

Dank aan mijn vrienden buiten het instituut, Amy, Arno, Femke, Jan-Willem, Kobe, Laura, Leon, Leonie, Marieke, Niels, Olivia, Sarah en Wouter (van wie sommigen zonder /ɑ/ en /a:/ in hun naam als personages in mijn stimuluszinnen verschenen). Jullie hebben bijgedragen aan een gezonde balans tussen mijn werk- en privéleven. Dank aan mijn partner, Tom. Tom, jij verscheen pas tegen het einde van het stuk ten tonele, maar dat maakt je rol niet minder belangrijk. Fijn dat je er bent.

Dank aan mijn Nederlandse moeders, Audrey en Marie-Anne og tak til mine danske forældre Morten og Malene. Ook dank aan mijn broertjes en zusjes: Tobias, Veerle, Demian, og selvfølgelig tak til mine halve halvsøstre Rebecca og Therese.

Speciale dank gaat uit naar mijn vader, aan wie ook dit boekje is opgedragen. Op de middelbare school was nagenoeg ieder tweede woord van elk werkstuk van jouw hand en zelfs gedurende het afronden van mijn eerste scriptie

heb ik wel eens beroep gedaan op je vakmanschap en wijsheid, als ik door de [tɑkken/ta:ken] de boom niet eens meer kon onderscheiden. In die zin heb jij jaren aan voorwerk gedaan voor het boek dat je nu in je handen houdt. Is het toch nog ergens goed voor geweest. Bedankt papseflaps, dat nooit je iets teveel was en ik altijd op je kan rekenen.

Curriculum Vitae

Merel Maslowski was born in Enschede, the Netherlands, in 1989. She grew up in Gug, Denmark, and in Apeldoorn, the Netherlands. In 2011, she first graduated an undergraduate degree in Fine Arts from Avans University of Applied Sciences in 's-Hertogenbosch, the Netherlands. Towards the end of her undergraduate degree, Merel started to carry out an undergraduate degree in General Linguistics at the University of Amsterdam, the Netherlands, completed in 2014. In 2015, she then obtained her degree of MSc by Research in Linguistics from the University of Edinburgh, United Kingdom. In 2015, Merel started her PhD project at the Max Planck Institute for Psycholinguistics in the Psychology of Language department, which she completed in 2019.

Publications

Maslowski, M., Meyer, A. S. & Bosker, H. R. (2018). 'Whether long-term tracking of speech rate affects perception depends on who is talking'. In *Proceedings of Interspeech 2017*, 586–590. doi: 10.21437/Interspeech.2017-1517.

Maslowski, M., Meyer, A. S. & Bosker, H. R. (2018). 'Listening to yourself is special: Evidence from global speech rate tracking'. *PloS one*, 13(9), e0203571. doi: 10.1371/journal.pone.0203571.

Maslowski, M., Meyer, A. S. & 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.

Maslowski, M., Meyer, A. S. & Bosker, H. R. (2019). 'Listeners normalize speech for contextual speech rate even without an explicit recognition task'. *The Journal of the Acoustical Society of America*, 146(1), 179–188. doi: 10.1121/1.5116004

Maslowski, M., Meyer, A. S. & Bosker, H. R. (under review). 'The time course of speech rate normalization depends on the distance of the context'.

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda I. van Turennout*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*
4. The open-/closed-class distinction in spoken-word recognition. *Alette Petra Haveman*
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*
6. Gesture and speech production. *Jan-Peter de Ruiter*
7. Comparative intonational phonology: English and German. *Esther Grabe*
8. Finiteness in adult and child German. *Ingeborg Lasser*
9. Language input for word discovery. *Joost van de Weijer*
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*
11. Producing past and plural inflections. *Dirk J. Janssen*
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*
13. From speech to words. *Arie H. van der Lugt*
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*
15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*
16. Language-specific listening: The case of phonetic sequences. *Andrea Christine Weber*

17. Moving eyes and naming objects. *Femke Frederike van der Meulen*
18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*
19. Morphology in speech comprehension. *Kerstin Mauth*
20. Morphological families in the mental lexicon. *Nivja Helena de Jong*
21. Fixed expressions and the production of idioms. *Simone Annegret Sprenger*
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*
23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*
24. Contextual influences on spoken-word processing: An electrophysiological approach. *Danielle van den Brink*
25. Perceptual relevance of prevoicing in Dutch. *Petra Martine van Alphen*
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*
27. Producing complex spoken numerals for time and space. *Marjolein Henriëtte Wilhelmina Meeuwissen*
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl Jenny Judith Karin Kemps*
29. At the same time…: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*
30. A grammar of Jalonke argument structure. *Friederike Lüpke*
31. Agrammatic comprehension: An electrophysiological approach. *Marijtje Elizabeth Debora Wassenaar*
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*
33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*
34. Phonetic and lexical processing in a second language. *Mirjam Elisabeth Broersma*
35. Retrieving semantic and syntactic word properties: ERP studies on the time course in language comprehension. *Oliver Müller*
36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*

37. Sensitivity to detailed acoustic information in word recognition. *Keren Batya Shatzman*

38. The relationship between spoken word production and comprehension. *Rebecca Özdemir*

39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*

40. The acquisition of phonological structure: Distinguishing contrastive from non-constrative variation. *Christiane Dietrich*

41. Cognitive cladistics and the relativity of spatial cognition. *Daniel Haun*

42. The acquisition of auditory categories. *Martijn Bastiaan Goudbeek*

43. Affix reduction in spoken Dutch: Probabilistic effects in production and perception. *Mark Plumaekers*

44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Madalla Kooijman*

45. Space and iconicity in German sign language (DGS). *Pamela M. Perniss*

46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*

47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*

48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*

49. Phoneme inventories and patterns of speech sound perception. *Anita Eva Wagner*

50. Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*

51. A grammar of Savosavo: A Papuan language of the Solomon Islands. *Claudia Ursula Wegener*

52. Prosodic structure in speech production and perception. *Claudia Kuzla*

53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*

54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*

55. Processing the fine temporal structure of spoken words. *Eva Reinisch*

56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*

57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*
58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*
59. Lexical interactions in non-native speech comprehension: Evidence from electroencephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*
60. Processing casual speech in native and non-native language. *Annelie Tuinman*
61. Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Payton Robinson*
62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*
63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*
64. The meaning and use of ideophones in Siwu. *Mark Dingemanse*
65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*
66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*
67. The relevance of early word recognition: Insights from the infant brain. *Caroline Mary Magteld Junge*
68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias Johannes Sjerps*
69. Structuring language: Contributions to the neurocognition of syntax. *Katrien Rachel Segaert*
70. Infants' appreciation of others' mental states in prelinguistic communication: A second person approach to mindreading. *Birgit Knudsen*
71. Gaze behavior in face-to-face interaction. *Federico Rossano*
72. Sign-spatiality in Kata Kolok: How a village sign language of Bali inscribes its signing space. *Connie de Vos*
73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*
74. Lexical processing of foreign-accented speech: Rapid and flexible adaptation. *Marijt Witteman*
75. The use of deictic versus representational gestures in infancy. *Daniel Puccini*

76. Territories of knowledge in Japanese conversation. *Kaoru Hayano*

77. Family and neighbourhood relations in the mental lexicon: A cross-language perspective. *Kimberley Mulder*

78. Contributions of executive control to individual differences in word production. *Zeshu Shao*

79. Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing. *Patrick van der Zande*

80. High pitches and thick voices: The role of language in space-pitch associations. *Sarah Dolscheid*

81. Seeing what's next: Processing and anticipating language referring to objects. *Joost Rommers*

82. Mental representation and processing of reduced words in casual speech. *Iris Hanique*

83. The many ways listeners adapt to reductions in casual speech. *Katja Pöllmann*

84. Contrasting opposite polarity in Germanic and Romance languages: Verum Focus and affirmative particles in native speakers and advanced L2 learners. *Giuseppina Turco*

85. Morphological processing in younger and older people: Evidence for flexible dual-route access. *Jana Reifegerste*

86. Semantic and syntactic constraints on the production of subject-verb agreement. *Alma Veenstra*

87. The acquisition of morphophonological alternations across languages. *Helen Buckler*

88. The evolutionary dynamics of motion event encoding. *Annemarie Verkerk*

89. Rediscovering a forgotten language. *Jiyoun Choi*

90. The road to native listening: Language-general perception, language-specific input. *Sho Tsuji*

91. Infants' understanding of communication as participants and observers. *Gudmundur Bjarki Thorgrímsson*

92. Information structure in Avatime. *Saskia van Putten*

93. Switch reference in Whitesands. *Jeremy Hammond*

94. Machine learning for gesture recognition from videos. *Binyam Gebrekidan Gebre*

95. Acquisition of spatial language by signing and speaking children: A comparison of Turkish sign language (TID) and Turkish. *Beyza Sumer*

96. An ear for pitch: On the effects of experience and aptitude in processing pitch in language and music. *Salomi Savvatia Asaridou*

97. lncrementality and Flexibility in Sentence Production. *Maartje van de Velde*

98. Social learning dynamics in chimpanzees: Reflections on (nonhuman) animal culture. *Edwin van Leeuwen*

99. The request system in Italian interaction. *Giovanni Rossi*

100. Timing turns in conversation: A temporal preparation account. *Lilla Magyari*

101. Assessing birth language memory in young adoptees. *Wencui Zhou*

102. A social and neurobiological approach to pointing in speech and gesture. *David Peeters*

103. Investigating the genetic basis of reading and language skills. *Alessandro Gialluisi*

104. Conversation electrified: The electrophysiology of spoken speech act recognition. *Rósa Signý Gisladottir*

105. Modelling multimodal language processing. *Alastair Charles Smith*

106. Predicting language in different contexts: The nature and limits of mechanisms in anticipatory language processing. *Florian Hintz*

107. Situational variation in non-native communication. *Huib Kouwenhoven*

108. Sustained attention in language production. *Suzanne Jongman*

109. Acoustic reduction in spoken-word processing: Distributional, syntactic, morphosyntactic, and orthographic effects. *Malte Viebahn*

110. Nativeness, dominance, and the flexibility of listening to spoken language. *Laurence Bruggeman*

111. Semantic specificity of perception verbs in Maniq. *Ewelina Wnuk*

112. On the identification of FOXP2 gene enhancers and their role in brain development. *Martin Becker*

113. Events in language and thought: The case of serial verb constructions in Avatime. *Rebecca Defina*

114. Deciphering common and rare genetic effects on reading ability. *Amaia Carrión Castillo*

115. Music and language comprehension in the brain. *Richard Kunert*

116. Comprehending Comprehension: Insights from neuronal oscillations on the neuronal basis of language. *Nietzsche H. L. Lam*

117. The biology of variation in anatomical brain asymmetries. *Tulio Guadalupe*

118. Language processing in a conversation context. *Lotte Schoot*

119. Achieving mutual understanding in Argentine Sign Language. *Elizabeth Manrique*

120. Talking Sense: the behavioural and neural correlates of sound symbolism. *Gwilym Lockwood*

121. Getting under your skin: The role of perspective and simulation of experience in narrative comprehension. *Franziska Hartung*

122. Sensorimotor experience in speech perception. *Will Schuerman*

123. Explorations of beta-band neural oscillations during language comprehension: Sentence processing and beyond. *Ashley Lewis*

124. Influences on the magnitude of syntactic priming. *Evelien Heyselaar*

125. Lapse organization in interaction. *Elliott Hoey*

126. The processing of reduced word pronunciation variants by natives and foreign language learners: Evidence from French casual speech. *Sophie Brand*

127. The neighbors will tell you what to expect: Effects of aging and predictability on language processing. *Cornelia Moers*

128. The role of voice and word order in incremental sentence processing. Studies on sentence production and comprehension in Tagalog and German. *Sebastian Sauppe*

129. Learning from the (un)expected: Age and individual differences in statistical learning and perceptual learning in speech. *Thordis Neger*

130. Mental representations of Dutch regular morphologically complex neologisms. *Laura de Vaan*

131. Speech production, perception, and input of simultaneous bilingual preschoolers: Evidence from voice onset time. *Antje Stoehr*

132. A holistic approach to understanding pre-history. *Vishnupriya Kolipakam*

133. Characterization of transcription factors in monogenic disorders of speech and language. *Sara Busquets Estruch*

134. Indirect request comprehension in different contexts. *Johanne Tromp*

135. Envisioning language - An exploration of perceptual processes in language comprehension. *Markus Ostarek*

136. Listening for the WHAT and the HOW: Older adults' processing of semantic and affective information in speech. *Juliane Kirsch*

137. Let the agents do the talking: On the influence of vocal tract anatomy on speech during ontogeny and glossogeny. *Rick Janssen*

138. Age and hearing loss effects on speech processing. *Xaver Koch*

139. Vocabulary knowledge and learning: Individual differences in adult native speakers. *Nina Mainz*

140. The face in face-to-face communication: Signals of understanding and non-understanding. *Paul Hömke*

141. Person reference and interaction in Umpila/Kuuku Ya'u narrative. *Clair Hill*

142. Beyond the language given: The neurobiological infrastructure for pragmatic inferencing. *Jana Bašnáková*

143. From Kawapanan to Shawi: Topics in language variation and change. *Luis Miguel Rojas-Berscia*

144. On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions. *Linda Drijvers*

145. Linguistic dual-tasking: Understanding temporal overlap between production and comprehension. *Amie Fairs*

146. The role of exemplars in speech comprehension. *Annika Nijveld*

147. A network of interacting proteins disrupted in language-related disorders. *Elliot Sollis*

148. Fast speech can sound slow: Effects of contextual speech rate on word recognition. *Merel Maslowski*