Applied Psycholinguistics **37** (2016), 1397–1415
doi:10.1017/S0142716415000612

Speaker-specific processing and local context information: The case of speaking rate

EVA REINISCH
Ludwig Maximilian University Munich

Received: March 23, 2015 Accepted for publication: November 8, 2015

ADDRESS FOR CORRESPONDENCE
Eva Reinisch, Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich, Schellingstraße 3, Munich 80799, Germany. E-mail: evarei@phonetik.uni-muenchen.de

ABSTRACT
To deal with variation in the speech signal, listeners rely on local context, such as speaking rate in a carrier sentence directly preceding a target, as well as more global properties of the speech signal, such as speaker-specific pronunciation variants. The present study addressed whether, despite its variability even within one speaker, habitual speaking rate can be tracked as a speaker-specific property and how such speaker-specific tracking of habitual rate would interact with effects of local-rate normalization. In two experiments, listeners were exposed to a 2-min dialogue between a fast and a slow speaker. At test, listeners categorized minimal word pair continua differing in the German /a/–/a:/ duration contrast spoken by the same two speakers. The results showed that listeners responded with /a:/ more often for the fast speaker but only when words were presented in isolation and not when presented with additional local-rate information. That is, despite the general assumption that duration cues and speaking rate are too variable to be used in a speaker-specific fashion, tracking habitual speaking rate may help speech perception. The results are discussed in relation to a belief-updating model of perceptual adaptation and exemplar models.

To be able to understand spoken language, listeners must deal with the fact that no two words are ever spoken in exactly the same way, especially when produced by different speakers. Speakers differ not only due to differences in their anatomy (e.g., a male vs. female voice) but also in habitual speech characteristics such as the way they produce certain segments (e.g., Kraljic & Samuel, 2007; Norris, McQueen, & Cutler, 2003), habitual speaking rate (Koreman, 2006; Quené, 2008; Tsao & Weismer, 1997), or even the speaking rate in a given situation (Miller, Grosjean, & Lomanto, 1984; Quené, 2013). This is a problem because a fast speaker's realization of the word "path" may sound like a slow speaker's "bath." The English sounds /p/ and /b/ differ among other cues most saliently in the time it takes between opening the lips and the beginning of the following /a/ (voice-onset time [VOT]; Abramson & Lister, 1985; Lisker & Abramson, 1964, 1967). In fast speech, however, durations are compressed (e.g., Crystal & House, 1982, 1988;

© Cambridge University Press 2015 0142-7164/15

Gay, 1978). As a result, a short VOT in /b/ could be perceived as a compressed /p/ in fast speech. It is important that many studies have shown that listeners compensate for this variation by interpreting durations such as VOT relative to the speaking rate of the context (e.g., Kidd, 1989; Miller, 1987; Miller & Dexter, 1988; Miller & Liberman, 1979; Sawusch & Newman, 2000, to name just a few; note that throughout this paper the term "speaking rate" will be used in line with these studies referring to articulation rate as pause rate and hesitations will be controlled for in the experiments). That is, relative to the fast context, the target VOT sounds relatively longer than following a slow context. Similar rate effects have been found with regard to the perception of vowel duration (Reinisch & Sjerps, 2013), word segmentation (Reinisch, Jesse, & McQueen, 2011a), lexical stress (Reinisch, Jesse, & McQueen, 2011b), and even the perception of function words (Dilley & Pitt, 2010).

In addition to using such "local" information like the speaking rate in a carrier sentence directly preceding a target word, listeners have been shown to track the global rate within an experimental session (Baese-Berk et al., 2014). Listeners have been shown to track properties that are specific to certain speakers in order to improve/facilitate speech perception. For example, listeners have been shown to take into account whether or not speakers have been heard before (i.e., they remember their voices; Goldinger, 1996, 1998; Nygaard & Pisoni, 1998; Nygaard, Somers, & Pisoni, 1994), what words speakers are likely to say (Creel & Tumlin, 2011), or how speakers are likely to pronounce certain sounds (e.g., in a foreign accent; Baese-Berk, Bradlow, Wright, 2013; Bradlow & Bent, 2008). These types of information are likely not independent processes but rather context effects that interact during the speech perception process (see, e.g., Sjerps & Reinisch, 2015). Because, however, little is known about these interactions, and the number of possible combinations is enormous, the present study set out to test the combination of two of them: speaker-specific processing and speaking rate. Specifically, it will be addressed whether, despite its variability even within one speaker, habitual speaking rate can be tracked as a speaker-specific property and how such a speaker-specific tracking of habitual rate would interact with effects of local-rate normalization.

Although speakers can be grouped into fast and slow speakers according to their habitual speaking rates (Koreman, 2006; Tsao & Weismer, 1997), corpus studies suggest that tempo variation within speakers can be substantial (Miller et al., 1984) and tends to be considerably larger within than between speakers (Quené, 2008). This variability has been used to explain how listeners tune in to an individual speaker's pronunciation of spectral contrasts but not, or to a lesser extent, to the speaker's pronunciation of duration contrasts. For example, Kraljic and Samuel (2007) showed that adaptation to deviating pronunciation variants of fricatives is perceived as speaker specific (e.g., /s/–/ʃ/, where the main cue is spectral center of gravity), but adaptation to unusual/ambiguous pronunciation variants in stop voicing (/d/–/t/, where the main cue is duration) is generalized across speakers. However, other studies demonstrated that listeners can remember duration properties in a speaker-specific fashion, for example, that a certain speaker tends to produce /p/ with a short VOT whereas another speaker produces the /p/ with a long VOT (e.g., Allen, Miller, & deSteno, 2003). At least when the respective duration characteristics are salient and consistent within an experiment, listeners use this knowledge to judge how typical a production of a word is for a speaker (Allen & Miller, 2004). Important to the present investigation, these authors show that speaker-specific variation in VOT depends on the speakers' individual speaking rates (see Theodore, Miller, & deSteno, 2009). It therefore seems likely that listeners also track speaker-specific rate information (i.e., each speaker's habitual rate) in speech perception, which then allows them to perceive words as intended at various rates.

Speaker-specific processing of habitual rate has to "compete" with another process that listeners use to deal with variability in temporal properties of the speech signal, namely, rate normalization. Listeners rely on local temporal information to interpret a following sound or word (where "local" mostly refers to the range of one context sentence though not necessarily the segments adjacent to a target). It is critical that normalization for local rate has been shown to apply across different speakers (Green, Tomiak, & Kuhl, 1997; Newman & Sawusch, 2009; Sawusch & Newman, 2000). When listeners hear the beginning of a sentence or syllable spoken by one speaker and finished by another (usually a male and a female voice), they take the rate of the first speaker into account when judging what the second speaker said. This evidence has been used to argue that normalization for local speaking rate takes place before other early perceptual processes such as stream segregation (i.e., the perceptual separation of voices) occur. It also implies that local-rate information is taken into account prelexically before word forms are accessed. Reinisch and Sjerps (2013) showed that rate context is taken into account as early as phones in the unfolding speech signal are being interpreted. The question now arises whether such local-rate normalization would override any speaker-specific processing of habitual rate that had been learned due to longer term exposure. In other words, would a previously experienced habitual rate of a speaker modulate the magnitude of the local-rate normalization effect?

A model of perceptual adaptation, the belief updating model (Kleinschmidt & Jaeger, 2015), suggests that whenever listeners recognize consistencies in the speech signal for a given situation, they will track situation- or speaker-specific distributions of acoustic cues. These specifically adapted models of cue distributions will be reapplied in perception when the situation or the speaker is recognized again. That is, upon encountering a situation or speaker that is similar or the same as one that has been experienced before, adaptation does not have to start over from baseline assumptions. Rather, the previously established cue distributions that optimally predicted the categories will be used as the new stating point for perception and further adaptation. What types of cue distributions and situational properties are being tracked is an empirical matter, but given some consistency within a certain situation the belief updating model predicts situation-specific tracking of these cues. That is, within an experiment, speaker-specific habitual rate information may be tracked and subsequently reapplied to categorize words differing in a duration contrast in a speaker-specific fashion. Note that in this case similar predictions would be made by exemplar-based theories of speech perception where rich acoustic detail is stored such that speaker- and situation-specific information may be used in a second encounter (Goldinger, 1996, 1998; Johnson, 1997, 2006; Pierrehumbert, 2001).

What is less clearly defined in the belief updating model are presumed "low-level" general auditory effects such as normalization for local speaking rate. Overall, the belief updating model does not make any direct assumptions about a processing hierarchy or timing of different adaptation processes (see Sjerps & Reinisch, 2015, for a discussion). Kleinschmidt and Jaeger merely state that "[i]n order to make good use of bottom-up information from acoustic cues, listeners require the appropriate likelihood function for the current situation" (2015, p. 160). That is, the interpretation of the speech signal is ideally modulated by appropriate top-down information about a given situation. The more variable local linguistic contexts are, the more the speech perception system has an incentive to track these local statistical distributions in conjunction with the more global nonlinguistic context (i.e., a speaker, situation, etc.). That is, listeners try to predict the variable input signal from global, possibly less variable, situations. In relation to the present question about the interaction of speaker-specific rate effects and "local" rate normalization within a carrier sentence, the belief updating model would then suggest an interaction of the effects; at least if the speaker-specific habitual rate information were distinct enough to be tracked and reapplied upon recognition of the speakers. If this is the case, even early perceptual processes (local-rate normalization) should be interpreted relative to or "predicted" from top-down knowledge (see Clark, 2013; Farmer, Brown, & Tanenhaus, 2013).

Alternatively, however, the duration of a carrier sentence may be sufficient for listeners to retrieve cue distributions for "fast" and "slow" speech independently of the speaker. As discussed above, studies on local-rate normalization show early, immediate, and partly speaker-independent effects. Moreover, despite "distal" rate effects (in the sense that speaking rate context does not have to be immediately adjacent to the target; e.g., Dilley & Pitt, 2010; Reinisch et al., 2011a; Summerfield, 1981), rate context is taken to have stronger effects the closer it is to a target. Newman and Sawusch (1996; Sawusch & Newman, 2000) go as far as to suggest a running time-window of approximately 250 to 300 ms that carries the main weight for rate normalization. In this view, speaker specificity may not affect local-rate normalization. The present study hence addressed two questions: first, whether global/habitual rate information can be used in a speaker-specific fashion at all, and second, if so, whether and how a speakers' habitual rate interacts with the process of local-rate normalization.

Experiment 1 addressed whether habitual speaking rate can be tracked in a speaker-specific fashion. We asked whether listeners "remember" the typical/habitual rate of two speakers and use this information when interpreting these speakers' speech. Specifically, listeners were presented with two female speakers in conversation. Two female speakers rather than a male and a female speaker were chosen to test *speaker*-specific processing rather than effects of possible gender differences. Listening to dialogue may be considered a reasonably natural situation, while allowing for a direct assessment of habitual rate differences between speakers. Note that the goal here was to assess the use of global/habitual rate information. Individual microvariations in speech timing such as potential differences in the timing relations between certain segments or stressed versus unstressed syllables could theoretically be used to identify the speakers. However, potential influences of microtiming on the overall perception of speech tempo were taken care of by counterbalancing the roles of the speakers (i.e., the words they said) as well as their overall habitual rate (i.e., across listeners both speakers were the fast or the slow one; see Methods section for details). Following exposure to a 2-min dialogue, listeners were asked to categorize minimal word pairs that contained a critical duration contrast: German /a/ versus /a:/. Note that unlike the respective Dutch vowel contrast (as used, e.g., in Reinisch & Sjerps, 2013), German /a/–/a:/ is a "real" duration contrast without consistent spectral differences (Jessen, 1993; Pätzold & Simpson, 1997). If listeners take into account the previously experienced habitual rate of the speaker, more /a:/ responses may be expected for the fast than for the slow speaker (because at a fast rate shorter durations may be sufficiently long to cue the long vowel). In terms of the belief updating model, this would suggest that listeners establish speaker-specific models for the distributions of duration cues and apply these to the interpretation of the test words accordingly.

Experiment 2 then tested whether and how speaker-specific rate information interacts with local contrastive effects of rate normalization. That is, listeners listened to the same dialogue as before and categorized the same minimal word pairs differing in the /a/–/a:/ duration contrast. However, this time the target words for the categorization task were presented at the end of rate-manipulated carrier sentences, mimicking typical local-rate normalization experiments. Given previous findings that rate information is the more important the closer it is to a target (Newman & Sawusch, 1996; Reinisch et al., 2011a; Wayland, Miller, & Volaitis, 1994), strong effects of the local-rate information were expected. The question was whether the previously experienced habitual rate of the speaker as heard in the dialogue would modulate this effect of local rate. That is, would the effect be stronger if the previously experienced speaker-specific rate matched the local rate (e.g., fast speaker from dialogue produces a fast context sentence) than when the two types of rate did not match (i.e., fast speaker from dialogue produces a slow context sentence). The results will be discussed with regard to perception models such as the belief updating model or exemplar models.

EXPERIMENT 1

Method

Participants. Sixteen native speakers of German participated for a small payment. They were recruited from the student population at the University of Munich.

Materials. A dialogue between two female speakers talking about their holidays was scripted such that the occurrence of the phonemes /a/ and /a:/ was minimized without losing naturalness of the utterances. Note that the critical segments could not be avoided altogether because they occur abundantly in function words like *haben* "have," which, for example, is used for forming the present perfect, which is the typical grammatical form in spoken German when talking about events in the (temporal) past. Eventually, a total of 433 words in the dialogue contained 23 tokens of /a/ and 15 tokens of /a:/ (see below for counterbalancing roles and rates, and hence the distribution of critical vowels across speakers). Two female native speakers of German were recorded reading both roles of the dialogue. They were instructed to read at a comfortable rate, which they should keep as steady as possible throughout the recordings. They were also asked to avoid changes in voice (quality or pitch) when switching roles. Turns in which the speaker misread part of a sentence were rerecorded, as were turns with perceptible changes in rate as judged by the experimenter. These were either due to hesitations (when the speaker slowed down or even stopped speaking within a turn) or if a speaker appeared to speed up over the course of the recordings. In these cases, the experimenter reminded the speaker to keep her tempo constant. Overall only a few turns ($<$10%) per speaker had to be rerecorded.

The two recordings of the dialogue were cut at phrase boundaries. Breaks and hesitations were excluded. Phrase durations were measured and subsequently changed using PSOLA as implemented in PRAAT (Boersma & Weenink, 2009). The amount of duration/rate change was calculated such that for each utterance a fast version would be 15% shorter, and a slow version 10% longer than the average of the two speakers' natural duration for this given utterance. Manipulated phrases were spliced back together leaving 300 ms silence between utterances. This amount of rate change and interutterance gap was arrived upon by informal pretesting ensuring that the resulting dialogues sounded natural while rates were distinctive enough to be recognized as fast and slow. Four versions of the dialogue were created such that within each version, the two speakers were characterized by different speaking rates, and across versions, each speaker would speak both roles, once fast and once slow (e.g., speaker 1 has Role 1 and speaks fast, while speaker 2 takes Role 2 and speaks slowly). Using all possible combinations of role and rate per speaker should counteract potential influences of each speaker's specific timing of segments and syllables within an utterance on the overall perception of speaking rate. The dialogue was just over 2 min long.

Both speakers recorded 16 German minimal word pairs differing in the /a/–/a:/ duration contrast. All words were recorded in semantically unconstraining carrier sentences spoken at a neutral speaking rate (i.e., speakers were reminded to speak at the same comfortable rate that had been used for the dialogue). Five minimal word pairs were selected in which the durations of target words, and especially those of the target vowels, matched most closely between speakers (see Table 1). These word pairs were *bannen–bahnen* ("banish"–"to channel"), *rammen–Rahmen* ("drive by impact"–"frame"), *Ratte–Rate* ("rat"–"installment"), *schlaff–Schlaf* ("saggy"–"sleep"), and *Wall–Wahl* ("ridge"–"election"). These target words were spliced out of their sentences, and /a/–/a:/ vowel continua were created. The longest value (i.e., the /a:/ endpoint) for each word was taken to be the average between the two speakers' /a:/ duration for that word. Sixteen shorter steps were then created by duration adjustment using the PRAAT duration tier and PSOLA resynthesis (i.e., resulting in a total of 17 steps). Across words, the step sizes ranged from 8.3 to 10.4 ms. All other segments in the words were set to an average value between the two speakers' segments. Two pretests were run, first, to test what range of the vowel continuum would be sufficient to allow for responses from fewer than 5% to more than 95% /a:/ responses at the continuum endpoints. Second, they tested whether the rate differences chosen for the dialogue (i.e., slow = 10% slower and fast = 15% faster than normal) would result in reasonably sized rate effects when implemented in local-rate contexts.

Table 1. *Carrier sentences and their target words*

Beim Wortspiel wählte sie immer den Begriff	rammen–Rahmen
In the word game she always chose the term	*drive by impact–frame*
Im Kreuzworträtsel suchten sie den Begriff	bannen–bahnen
In the crossword puzzle they were looking for the term	*banish–to channel*
Der neue Film hieß “die letzte	Ratte–Rate
The new film was titled “the last	*rat–installment*
Du kennst doch die Bedeutung von dem Wort	Wall–Wahl
You do know the meaning of the word	*ridge–election*
Der Stotterer mühte sich mit dem Wort	schlaff–Schlaf
The stutterer had trouble with the word	*saggy–sleep*

Note: English translations mimic German sentence structure.

Pretests. Eight participants who did not take part in the main experiments participated for a small monetary compensation. The listeners’ task was to listen to the minimal-pair continua in sentence-final position and indicate by button press which of two words they heard. Words were presented at the end of the carrier sentences in which they had been recorded (see Table 1). Sentences were set to a speaking rate that was either 10% slower than the average of the two speakers’ rates or 15% faster, hence matching the rate manipulation of the dialogue. Rate was matched on the basis of the overall sentence durations. In the first pretest, participants were presented with 11 of the 17 continuum steps. Six steps were omitted through sparser sampling of steps close to the endpoints of the continuum (where every other step was dropped). This reduced the number of trials to 220 (2 speakers × 2 rates × 5 sentences/minimal pairs). The left panel of Figure 1 shows the categorization functions for the minimal pairs, following the fast and slow sentences. It can be seen that several steps of the continuum close to the endpoints were identified with close to ceiling performance, leaving little room for the effect of rate. For the middle steps, however, a rate effect was clearly evident, suggesting that the amount of rate manipulation (10% slower vs. 15% faster than the average of the two speakers’ rates) was sufficient to trigger reliable effects of rate normalization from the local context.

A second pretest further explored the possibility of decreasing the number and range of continuum steps in favor of increasing the number of repetitions per step to be used in the main experiments while still retaining a categorization function from a clear /a:/ to a clear /a/. In addition, the minimal pair *schlaff–Schlaf*, whose categorization function wasn’t as clear as the other pairs’ functions, was dropped to reduce the number of words. Participants listened twice to all combinations of the remaining four sentences, the two speakers, two rates, and seven continuum steps. The right panel of Figure 1 provides the results. The continuum still ranged from fewer than 5% /a:/ responses to more than 95% /a:/ responses while showing a substantial rate effect for all but the newly established continuum endpoints. In this way, any expected effects for the main experiments should have been maximized. Note that in Experiment 1 the minimal pair continua were used in isolation and in Experiment 2 they were used in their carrier sentences.

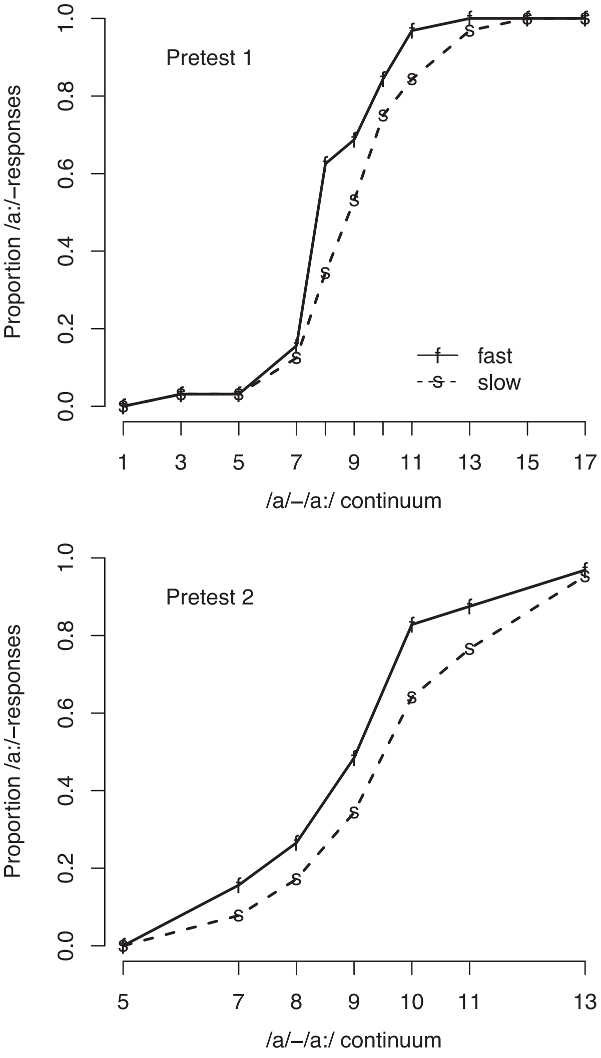

Figure 1. Proportion /a:/ responses in the two pretests over the /a/–/a:/ continuum following fast and slow carrier sentences.

Design and procedure. Participants were seated in a sound-attenuated booth wearing headphones to listen to the speech materials. They were randomly assigned to one of the four versions of the dialogue (i.e., two roles in the dialogue crossed with two habitual speaking rates) such that four participants listened to each version. They were instructed to listen for content because they may be asked about it after the experiment. After the end of the dialogue, participants had to answer the question whether one of the speakers was on holiday in Norway or

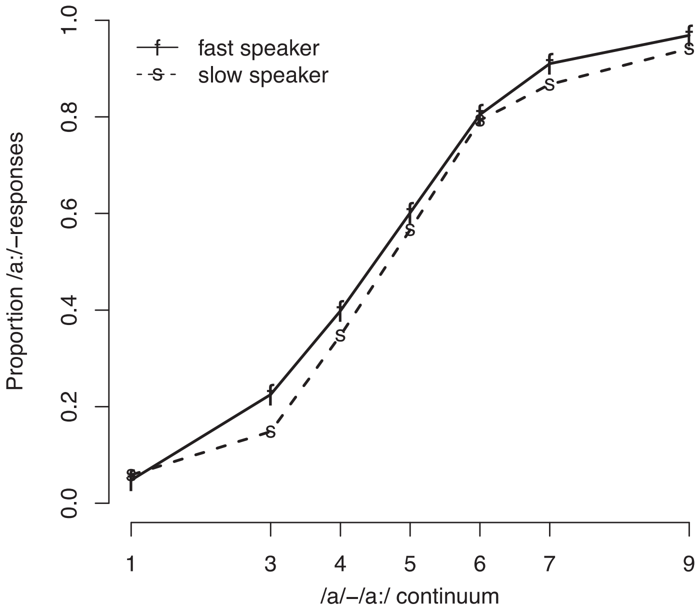

Figure 2. Proportion /a:/ responses over the /a/–/a:/ continuum for fast (solid line) versus slow (dashed line) speaker from the dialogue.

Sweden.[1] Immediately thereafter all participants performed the same phonetic categorization task, categorizing each of the four minimal pair continua in each of the speakers' voices. The two speakers' word items were presented intermixed. On every trial participants saw the response options for the upcoming minimal pair on a computer screen for 500 ms before the sound was played. Upon hearing the word, participants were instructed to press the 1 key or 0 key on a computer keyboard to indicate which of the two words they perceived. The key layout (left–right) matched the left and right word on the screen. The word on the left always contained /a:/. Participants were informed that their response was logged by seeing the chosen response option move upward on the screen, where it stayed for 400 ms. The next trial started 300 ms later (black screen). Each participant received a total of 224 trials; that is, each step of each continuum for each of the two speakers was presented four times. Items were presented in random order with the restriction that all tokens were presented once before any of the tokens was repeated. The experiment was controlled by ePrime software (Psychology Software Tools, Inc.) and took approximately 20 min to complete.

Experiment 1 results

Figure 2 shows the proportion /a:/ responses over continuum steps separately for the speaker that was heard as the fast speaker in the dialogue and the speaker that was heard speaking slowly. Note that this measure is an aggregate over the two actual speakers (i.e., the two voices) because across conditions/participants each speaker was the fast and the slow speaker in both roles (and additional analyses with voice/speaker as a factor showed no difference with regard to the effect of rate). As can be seen in Figure 2, the solid line, representing the respective fast speaker, is slightly above the dashed line that represents the slow speaker. That is, more "long vowel" (/a:/) responses were given for the fast than for the slow speaker; hence, listeners appeared to take into account the previously experienced, habitual rate of a speaker when categorizing this speaker's minimal pair continua.

Statistical analyses confirmed this observation. A generalized linear mixed-effects model was fit with response (/a:/ coded as 1, /a/ coded as 0) as a dependent variable and Continuum Step (centered on 0, recoded to range from –0.5 to 0.5), Speaker Rate (fast speaker from dialogue = 0.5, slow speaker = –0.5), and their interaction as fixed factors (using the lme4 package, v. 1.1–7 in R, v. 3.1.2). Participant was entered as a random factor with random slopes for all (within-participant) fixed factors (i.e., a full random-effects structure was used; Barr, Levy, Scheepers, & Tily, 2013). A logistic linking function was used to account for the dichotomous dependent variable.

The results showed significant effects of the intercept term ($b_{\text{Intercept}} = 0.38$, $z = 2.70$, $p < .01$), indicating an overall preference for /a:/ responses; Continuum Step ($b_{\text{ContinuumStep}} = 7.63$, $z = 17.59$, $p < .001$), with more /a:/ responses the longer the vowel, and, critically, they showed an effect of Speaker Rate ($b_{\text{SpeakerRate}} = 0.25$, $z = 2.64$, $p < .01$). More /a:/ responses were given for the fast than for the slow speaker. The interaction between Continuum Step and Speaker Rate was not significant ($b_{\text{Step:SpeakerRate}} = 0.19$, $z = 0.18$, $p = .86$); that is, the effect of Speaker Rate was stable over the whole continuum rather than restricted to, for example, the most ambiguous steps, or larger at one end of the continuum than the other.

Experiment 1 discussion

Experiment 1 showed that listeners take into account the habitual speaking rate of speakers when interpreting these speakers' utterances upon their next encounter. During a 2-min dialogue, listeners were presented with the speakers' habitual rates in direct contrast. The rates were manipulated to the extent that differences were clearly audible, and pretests established that the same amount of local-rate change in carrier sentences immediately preceding the target words led to the expected shift in a categorization function for the /a:/–/a/ vowel duration contrast. What is important to note about the rate manipulation in the dialogues is that it was implemented linearly by compression or expansion of entire utterances. That is, the microstructure of each speaker's timing patterns (between segments and syllables) was kept intact. While it cannot be excluded that this kind of microtiming contributes toward the overall perception of speech tempo, in the present case it can be assumed that a rate change of 10% slower versus 15% faster than the average rate is substantial enough to be effective in spite of any contributions of microtiming. Moreover, across participants, both speakers were heard in both roles of the dialogue and as both the fast and the slow speaker. This should counteract any specific effects of microtiming. It is important that explicit instructions as well as the setting of a conversation in general should have drawn listeners' attention to the content rather than the form (or rate) of the dialogue.

To sum up, listening to a 2-min dialogue between two female speakers allows listeners to tune in to their specific habitual speaking rates and to use this information in a subsequent phonetic categorization task of an /a:/–/a/ duration contrast. This confirms suggestions made by the belief updating model (Kleinschmidt & Jaeger, 2015) that listeners track relatively stable properties of a speech signal to create situation- and/or speaker-specific models of cue distributions that can be used for speech processing when a similar situation/speaker is recognized. Experiment 2 explored how this categorization in terms of speaker-specific habitual rate would be modulated by local-rate context.

EXPERIMENT 2

Method

Participants. Twenty-four native speakers of German participated for a small payment. They were recruited from the student population at the University of Munich. None had participated in Experiment 1 or in the pretests.

Materials. The same dialogue as described in the Methods section of Experiment 1 was used for exposure to familiarize participants with the fast versus slow habitual speaking rate of the two female speakers. For the phonetic categorization judgments at test, however, words were not presented in isolation but were instead embedded in semantically unconstraining carrier sentences (see Table 1) that were also manipulated in rate (i.e., local rate). Fast sentences were made 15% faster than the average duration of this sentence as produced by the two speakers; slow sentences were made 10% slower than the average. These were the same carrier sentences the minimal word pairs had been recorded in and that were used in the pretests. The pretests already indicated that this magnitude of rate change in the sentences affects vowel categorization in the minimal pairs.

Design and procedure. Design and procedure were similar to Experiment 1. Listeners listened to one of the four versions of the dialogue (i.e., 2 roles × 2 rates) and then performed phonetic categorization of the minimal pairs, now at the end of fast or slow carrier sentences. That is, in addition to the speakers' habitual speaking rates that listeners had experienced in the dialogues, they had local-rate information in the carrier sentences. It is important that for both speakers the local-rate information was fast in half of the trials and slow in the other half, that is, either matching or mismatching their habitual rate. Each participant received a total of 224 categorization trials. Each step of each of the four word continua for each of the two speakers was presented twice in each, the fast and slow, version of the carrier sentence. Items were presented in random order with the restriction that all tokens were presented once before any of the tokens was repeated. The experiment was controlled by ePrime software (Psychology Software Tools, Inc.) and took approximately 30 min to complete.

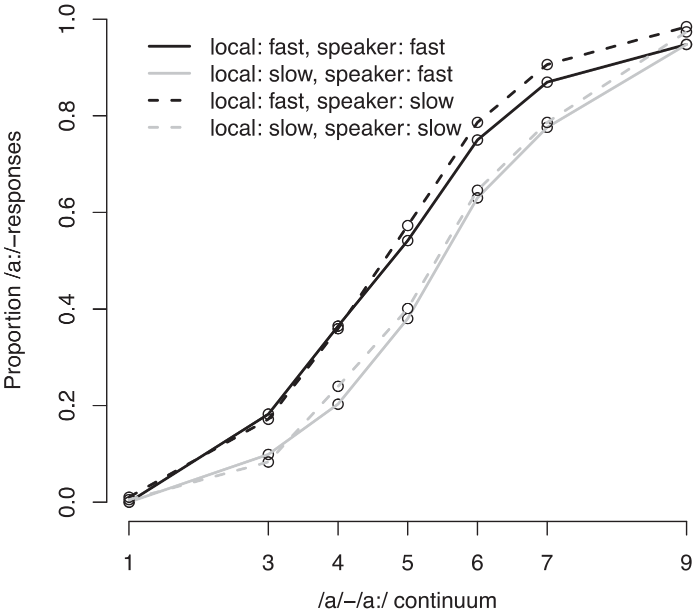

Figure 3. Proportion /a:/ responses over the /a/–/a:/ continuum for fast versus slow local rates (black vs. gray lines) and fast versus slow speakers from the dialogue (solid vs. dashed lines).

Experiment 2 results

Figure 3 shows the proportion of /a:/ responses along the vowel duration continuum for Local Rate (i.e., of the carrier sentence at test) and Speaker Rate (i.e., the habitual rate of the speakers heard in the dialogue) factors. As can be seen, there is a substantial effect of Local Rate (difference in black vs. gray lines), but differences between the fast and slow speakers (solid and dashed lines) appear rather small, and, in addition, are in the opposite than expected direction (dashed lines are above the solid lines, i.e., more /a:/ responses for the slow than the fast speaker).

The statistical results confirm this observation. A generalized linear mixed-effects model was fit with response (/a:/ coded as 1, /a/ coded as 0) as a dependent variable and Continuum Step (centered on 0, recoded to range from –0.5 to 0.5), Speaker Rate (fast speaker from dialogue = 0.5, slow speaker = –0.5), Local Rate (fast = 0.5, slow = –0.5), and all interactions as fixed factors. Participant was entered as a random factor with random slopes for all (within-participant) fixed factors. A binomial linking function was used to account for the dichotomous dependent variable. Table 2 shows the results. Listeners gave more /a:/–responses the longer the vowel, and following a fast than a slow Local Rate. No effect of Speaker Rate was found (with the tendency in the opposite than expected direction also evidenced by the negative regression weight), and there was no interaction between Local Rate and Speaker Rate as could be expected if the two types of information were used jointly or effects were influencing one another. None of the

Table 2. *Results of Experiment 2*

	b	*z*	*p*
Intercept	−0.11	−0.98	.33
Continuum step	8.19	27.37	<.001
Local rate	0.73	9.32	<.001
Speaker rate	−0.16	−1.23	.22
Continuum step: local rate	0.01	0.02	.99
Continuum step: speaker rate	−0.65	−0.85	.39
Local rate: speaker rate	−0.07	−0.47	.64
Continuum step: local rate: speaker rate	−0.69	−0.86	.39

other interactions reached significance; that is, as in Experiment 1, the effects of rate did not linearly vary over continuum steps.

An additional analysis was carried out to test whether listeners stopped using habitual speaking rate over the course of the test phase (i.e., started using local rate instead). This analysis was restricted to the five middle steps of the duration continua where effects of local rate were most evident (see Figure 3). A generalized linear mixed-effects model was fit with Speaker Rate, Local Rate, and Trial Number as fixed factors (all coded as before) and a full random effects structure. The only significant effects were Local Rate ($b_{\text{LocalRate}} = 0.52$, $z = 7.90$, $p < .001$; more /a:/ responses following the fast than the slow carrier sentences) and Trial Number ($b_{\text{TrialNumber}} = 0.26$, $z = 2.14$, $p < .05$; more /a:/ responses later in the experiment). All other factors and interactions had a *p* value of >.39. The effects were thus not modulated over the course of the experiment.

Finally, the impression from Figures 2 and 3 was confirmed that the effect of Local Rate in Experiment 2 was larger than the effect of habitual Speaker Rate in Experiment 1. This was done by fitting linear regression models separately for each participant with /a:/ responses as the dependent variable and Continuum Step and Rate (Habitual/Speaker Rate in Experiment 1, Local Rate in Experiment 2) as factors. In this way, regression weights for Rate (and Step) were obtained for each participant. Given the coding of factors described above, regression weights can be used as a measure of effect size. Regression weights for rate (as a measure for the magnitude of the rate effect) from participants in Experiment 1 versus Experiment 2 were then compared in an independent samples *t* test. A significant difference in the effects of habitual Speaker Rate in Experiment 1 versus Local Rate in Experiment 2 was found, $t\ (23.37) = 4.3$, $p < .001$.

Experiment 2 discussion

Experiment 2 tested whether the speakers' habitual speaking rate that had been shown to modulate listeners' categorization responses in a speaker-specific fashion in Experiment 1 would modulate the use of local-rate information in context sentences. As was expected from previous literature, listeners used local-rate information from the carrier sentences such that more /a:/ responses were given following the fast than the slow carrier sentences. However, this effect was not modulated by the speakers' habitual rate information that had been experienced in the dialogue. Rather, the effect of Speaker Rate showed a slight tendency in the opposite than expected direction. According to Cook and Campbell (1979), a tendency in the opposite than expected direction suggests that for this effect the null hypothesis may be accepted. Moreover, there was no decrease in the magnitude of any of the rate effects over the experiment that could have disguised an effect of Speaker Rate and hence a modulation of the Local Rate effect. Overall, the effect of local rate in Experiment 2 was even stronger than the effect of habitual rate in Experiment 1. Implications of these results for our understanding of speech processing will be discussed in the General Discussion.

GENERAL DISCUSSION

The present study adds another piece to the puzzle of understanding how listeners deal with variation in the speech signal. Previous evidence suggested that durational information in general and global speaking rate in particular may be too variable to be used in a speaker-specific fashion (e.g., Miller et al., 1984; Quené, 2013). However, the present results demonstrated that at least under certain circumstances, habitual speaking rate can be tracked as a speaker-specific property and used for speech perception. This is in line with the belief updating model of perceptual adaptation, which suggests that listeners track speaker- or situation-specific cue distributions that are deemed sufficiently stable in a given situation. Exemplar models, as will be discussed below, may also account for the present findings, because they posit the storage of individual renditions of words or sounds that preserve information about the speaker and may have additional information associated to them (e.g., Pierrehumbert, 2001).

Note that there was some previous evidence that listeners can track duration cues for more than one speaker (i.e., for two speakers) because after exposure, they can judge certain durations as more or less typical for that speaker (Allen & Miller, 2004; Theodore et al., 2009). This is why the present study tried to maximize the chance of finding possible speaker-specific effects of habitual rate by presenting listeners with a dialogue in which voices and rates could be compared directly. In addition, the speech tempo for each of the speakers remained stable over the course of the conversation, which does not entirely match what is usually found in dialogue research. It has been suggested that in the course of an interaction, interlocutors converge to each other in terms of rate (e.g., Jungers & Hupp, 2009; Wilson & Wilson, 2005). This may encourage speaker-independent processing. However, because in the present study both speakers kept their typical habitual rate over the whole dialogue, listeners may have been encouraged to track the speakers' habitual rates in a speaker-specific fashion. Note that the goal here was to demonstrate that tracking of speaker-specific rate is possible and to maximize chances of detecting joint effects of habitual and local rate in Experiment 2.

An alternative account of the present results would be that listeners did not actually adapt to the speaking rate of the two speakers but rather relied on the nature of the few examples of /a:/ and /a/ in the dialogue (they could not be avoided when scripting the dialogue). However, if it were the specific segments that listeners adapted to, then the effect of the speakers' habitual pronunciation of the vowels should have been present in Experiment 2. Instead, the results of Experiment 2 suggested that local-rate information overpowered any speaker-specific effects.

Because the effect of Speaker Rate in Experiment 2 showed a tendency in the opposite than expected direction, a few issues warrant further discussion. Experiment 2 addressed whether and how speaker-specific tracking of habitual rate would interact with effects of local-rate normalization. Although it is likely that listeners may have started to track speaker-specific cue distributions for duration contrasts during the dialogue in Experiment 2, local-rate information influenced phonetic categorization more strongly: even more than the effect of habitual rate in Experiment 1. That is, while the effect of Speaker Rate (i.e., habitual rate) may be just too weak in general to be measured in the presence of local-rate normalization, there are at least two explanations why local rate is so much stronger (apart from being local). First, normalization for local-rate information occurs too early during speech perception for it to be influenced by speaker information (Newman & Sawusch, 2009). Second, in terms of the belief updating model, listeners are likely to rely on cue distributions for fast versus slow speech that have been calibrated in a speaker-independent fashion through repeated exposure to different speaking rates. It is possible that these distributions for fast and slow speech constitute the starting point for the speaker-specific models. That is, listeners learned that for one speaker, the fast model applies; for the other speaker, the slow model. Then, during phonetic categorization, as soon as listeners realized that rate and speaker did not correlate anymore, they relied on their speaker-general distributions for fast versus slow speech. This could either happen immediately such that listeners appear to "switch off" the speaker-specific cue distributions or in a gradual fashion (though potentially quickly).

Specifically, the alternative suggests that in the rate normalization task, listeners quickly recalibrated their speaker-specific cue distributions for the two speakers such that they became much broader to incorporate both fast and slow rates that both were encountered during the rate normalization task. This in turn would leave the speaker-independent distributions for fast versus slow speech the better fitting cue distributions on every single trial. These then would drive the effect measured in Experiment 2. This modulation of speaker-specific cue distributions appears to happen quickly, because no change over the experiment could be detected for any of the rate effects (i.e., trial number did not interact with any of the rate effects). Although the present set of experiments is not able to disentangle the "switch off" speaker-specific information view from a modulation perspective, the general adaptive nature of the belief updating model speaks in favor of modulation. Listeners up- or down-weigh cues according to the affordances of a given listening situation. Details about the speed at which this adaptation can happen will have to remain for future research.

Such a modulation of processing has been demonstrated with other aspects of context situations (Brouwer, Mitterer, & Huettig, 2012; McQueen & Huettig, 2012; Poellmann, Mitterer, & McQueen, 2014). That is, when listening to clear speech, listeners give special weight to segments in word onset to modulate lexical access and identify the word that is being said. However, if context information suggests that the segments in the word onset may be less reliable than expected because the context is partially masked by radio noise (McQueen & Huettig, 2012) or contains segmental reductions in casual speech (Brouwer et al., 2012; Poellmann et al., 2014), then listeners reduce their reliance on word-initial segments in spoken-word recognition. Segmental mismatches in word-onset position thus cause less bottom-up inhibition in accessing acoustically similar words. That is, listeners flexibly adjust their processing mechanisms to a given context situation.

Speaker-specific processing of information could also be modeled in exemplar models of speech perception (e.g., Pierrehumbert, 2001). In exemplar models, words or sounds are stored as clouds of heard/remembered tokens of a given category. These tokens are organized such that more similar instances are represented as closer together than dissimilar ones and categorization can follow multiple structuring schemes. That is, listeners could remember the associated speakers and rates of a given token. Perception then works by matching (the acoustics of) incoming tokens to the properties of the stored exemplars. Results of Experiment 1, the finding that listeners can track speaker-specific rate information, could be explained by the best match of the words in the categorization task to the recently added exemplars from the two speakers heard during the dialogue. Note that this would require exemplars to occur on at least a segmental level because none of the words used for categorization had been heard during the dialogue. Results of Experiment 2, the stronger influence of local-rate information, may be accounted for by the weakness or relatively low number of the newly stored exemplars from the two speakers. The much higher number of exemplars with the (global) labels “fast” versus “slow” should then diminish or eliminate the speaker-specific influence. In this regard, it remains to be shown whether highly familiar speakers, given their habitual rates are sufficiently different, would allow for an observable modulation of the effects.

In sum, the present study showed that listeners can use temporal information (here in the form of a speaker’s habitual speaking rate) in a speaker-specific fashion and can use this information in categorizing duration cues. This is in spite of previous suggestions that duration cues and speech tempo are too variable even within one speaker to be useful in speech perception. In the absence of contradicting local information, listeners appear to rely on information that has been experienced as stable for a particular speaker, including habitual speaking rate.

ACKNOWLEDGMENTS

The author was supported by a grant from the Alexander von Humboldt Foundation. I thank Carolin Sabath and Anna-Maria Meck for help with the preparation and running the experiments.

NOTE

[1] The correct answer was “Norway,” but because this information appeared within the first few sentences of the dialogue, the one participant who answered wrongly was nevertheless included in the data set.

REFERENCES

Abramson, A. S., & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. In V. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 25–33). New York: Academic Press.

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

Allen, J. S., Miller, J. L., & deSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 113*, 544–552.

Baese-Berk, M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign-accented speech. *Journal of the Acoustical Society of America, 133*, EL174–EL180.

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

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

Boersma, P., & Weenink, D. (2009). PRAAT, doing phonetics by computer (version 5.1) [Computer software]. Retrieved from http://www.praat.org

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*, 707–729.

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes, 27*, 539–571.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–253.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Creel, S. C., & Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language, 65*, 264–285.

Crystal, T. H., & House, A. S. (1982). Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America, 72*, 705–716.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America, 83*, 1553–1573.

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

Farmer, T. A., Brown, M., & Tanenhaus, M. C. (2013). Prediction, explanation, and the role of generative models in language processing: Commentary to Clark, A. *Behavioral and Brain Sciences, 36*, 211–212.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America, 63*, 223–230.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166–1183.

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

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception & Psychophysics, 59*, 675–692.

Jessen, M. (1993). Stress conditions on vowel quality and quantity in German. *Working Papers of the Cornell Phonetics Laboratory, 8*, 1–27.

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

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34*, 485–499.

Jungers, M. K., & Hupp, J. M. (2009). Speech priming: Evidence for rate persistence in unscripted speech. *Language and Cognitive Processes, 24*, 611–624.

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

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*, 148–203.

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

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*, 1–15.

Lisker, L., & Abramson, A. S. (1964). A cross language study of voicing in initial stops: Acoustic measurements. *Word, 20*, 384–420.

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech, 10*, 1–28.

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *Journal of the Acoustical Society of America, 131*, 509–517.

Miller, J. L. (1987). Rate-dependent processing in speech perception. In A. W. Ellis (Ed.), *Progress in the psychology of language* (Vol. 3, pp. 119–157). London: Erlbaum.

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 369–378.

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

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

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

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*, 46–65.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*, 204–238.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*, 355–376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*, 42–46.

Pätzold, M., & Simpson, A. P. (1997). Acoustic analysis of German vowels in read speech. In A. P. Simpson, K. J. Kohler, & T. Rettstadt (Eds.). *The Kiel Corpus of Read/Spontaneous Speech—Acoustic database, processing tools and analysis results* (AIPUK Vol. 32, pp. 215–247). Kiel, Germany: IPDS.

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

Poellmann, K., Mitterer, H., & McQueen, J. M. (2014). Use what you can: Storage, abstraction processes, and perceptual adjustments help listeners recognize reduced form. *Frontiers in Psychology: Language Sciences, 5*, 437.

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

Quené, H. (2013). Longitudinal trends in speech tempo: The case of queen Beatrix. *Journal of the Acoustical Society of America, 133*, EL452–EL457.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011a). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 978–996.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011b). Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue. *Language and Speech, 54*, 147–166.

Reinisch, E., & Sjerps, M. J. (2013). Compensation for speaking rate and spectral context take place at a similar point in time. *Journal of Phonetics, 41*, 101–116.

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

Sjerps, M. J., & Reinisch, , E. Reinisch (2015). Divide and conquer: How perceptual contrast sensitivity and perceptual learning cooperate in reducing input variation in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 710–722.

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

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

Tsao, Y.-C., & Weismer, G. (1997). Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component. *Journal of Speech, Language, and Hearing Research, 40*, 858–866.

Wayland, S. C., Miller, J. L., & Volaitis, L. E. (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America, 95*, 2694–2701.

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing on turn-taking. *Psychonomic Bulletin & Review, 12*, 957–968.