Attention, Perception, & Psychophysics (2021) 83:3162–3182
https://doi.org/10.3758/s13414-021-02334-w

Rhythmic and speech rate effects in the perception of durational cues

Jeremy Steffman[1]

Accepted: 25 May 2021 / Published online: 12 July 2021

Abstract

Listeners' perception of temporal contrasts in spoken language is highly sensitive to contextual information, such as variation in speech rate. The present study tests how rate-dependent perception is also mediated by distal (i.e., temporally removed) rhythmic patterns. In four experiments the role of rhythmic alternations and their interaction with speech rate effects are tested. Experiment 1 shows proximal speech rate (contrast) effects obtain based on changes in local context. Experiment 2 shows that these effects disappear with the addition of distal rhythmic alternations, indicating that rhythmic grouping shifts listeners' perception, even when proximal context conflicts. Experiments 3 and 4 explore how orthogonal variation in overall speech rate impacts these effects and finds that trial-to-trial (i.e., global) speech rate variation eliminates rhythmic grouping effects, both with and without variation in proximal (immediately preceding) context. Together, these results suggest a role for rhythmic patterning in listeners' processing of durational cues in speech, which interacts in various ways with proximal, distal, and global rate contexts.

Keywords Speech perception · Durational processing · Speech rhythm · Speech rate · Perceptual grouping

Introduction

The temporal structure of speech is highly variable. Both within and across individuals, the rate at which speech is produced varies from utterance to utterance (Quené, 2008; Quené, 2013; Miller, Grosjean, & Lomanto, 1984). Different languages have also been shown to exhibit characteristic speech rates (Pellegrino, Coupé, & Marsico, 2011). This pervasive temporal variability modulates the distribution of acoustic cues over time, and accordingly, listeners must take temporal structure into account when they perceive speech. At the same time, spoken language is characterized by variations in temporal structure that are *rhythmic*, defined by perceived recurring patterns (Hay & Diehl, 2007; Hawkins & Smith, 2001; Lehiste, 1977). Rhythmic structure constitutes another source of contextual variation in the speech stream, and a body of literature suggests rhythmic patterning plays an important role in word segmentation and other domains of speech processing (Dilley & McAuley, 2008; Dilley, Mattys, & Vinke, 2010; Kidd, 1989; Morrill, Dilley, McAuley, & Pitt, 2014; Quené & Port, 2005). The present study tests how rhythmic patterning mediates perception of temporal cues in speech, and how these influences interact with local and non-local variations in speech rate, building on past work in this domain.

Speech rate effects

Perception of durational cues in context can be characterized as rate-dependent in the sense that listeners' categorization of a phonetic continuum ranging between two phonemic categories shifts on the basis of contextual speech rate.[1] Listeners' ability to factor the temporal structure of

✉ Jeremy Steffman
jeremy.steffman@northwestern.edu

[1] Northwestern Department of Linguistics, 2016 Sheridan Road, Evanston, IL 60208, USA

[1] Cues that have been shown to be rate-dependent include voice onset time (Miller & Volaitis, 1989; Toscano & McMurray, 2015), formant transition duration as a manner cue (Wade & Holt, 2005; Miller & Liberman, 1979), vowel duration as a cue to coda obstruent voicing (Heffner, Newman, & Idsardi, 2017; Steffman, 2019), and vowel duration in a language with contrastive vowel length (Bosker, 2017; Reinisch & Sjerps, 2013). Rate-dependent perception also extends to syllable identification and word segmentation, discussed below (Bosker, Sjerps, & Reinisch, 2020; Dilley & Pitt, 2010; Reinisch, Jesse, & McQueen, 2011).

speech into their perception of durational cues can therefore be taken to play an important role in the comprehension of spoken language. These sorts of contextual speech rate effects have been shown to occur on the basis of both "proximal" and "distal" speech rate, where proximal refers to rate changes that are temporally adjacent or close by in terms of some unit of speech (e.g., segments, syllables). The term distal, though defined in various ways throughout the literature (Heffner et al., 2017), can be taken in general to mean the rate of speech of material further removed from a given target sound, though distal can also be used to refer to overall rate does not necessarily exclude proximal context (if rate changes are manifested over the entirety of a precursor). In addition to proximal and distal rate, listeners track the long-term temporal patterns of an experiment as a whole (across trials), that is, the *global* speech rate context. See, for example, Stilp (2020) for an overview of temporal context effects in perception.

The ways in which proximal, distal, and global rate interact in speech perception are complex, and studies comparing these effects offer a nuanced picture of their relative importance.[2] Below, some relevant interactions between proximal, distal and global contexts are outlined.

Proximal and distal rate

Comparing distal and proximal rate, in a series of experiments (Bosker, 2017) crossed the rate of repetition of a pure tone precursor (slow/fast) with the duration of each tone in the precursor (short/long), and tested how this 2 × 2 manipulation modulated listeners' perception of contrastive vowel length in Dutch. Bosker found that when distal rate (i.e., fast/slow repetitions of tones) varied, proximal contrast effects (based on short/long preceding tone durations) were not observed, suggesting that distal rate cues take precedence when varying orthogonally with proximal changes. This presents a case wherein distal temporal cues are more relevant than proximal ones. However, other studies have shown that proximal contexts can be weighted more heavily than distal ones. For example, Newman and Sawusch (1996) tested how proximal and distal rate variation following a given target sound impacted perception of various temporal contrasts. The authors observed robust proximal effects, while finding that increased temporal separation between a given target sound reduces and eliminates rate-dependent adjustments (though unlike Bosker they did not cross proximal and distal rates).[3] Adding further nuance, Heffner, Newman, and Idsardi (2017) examined how proximal and distal speech rate contexts (testing various definitions of distal) influenced listeners' perception of word-initial consonant voicing as cued by voice onset time (e.g., "coat"-"goat"), and word-final consonant voicing as cued by preceding vowel duration (e.g., "coat"-"code"). Heffner et al. (2017) found that variation in distal speech rate did not exert any influence on word-initial voicing perception (cued by voice onset time), though it did impact perception of vowel duration as a cue to word-final voicing, under certain definitions of distal. In the domain of word segmentation, Reinisch, Jesse, and McQueen (2011) paint a complex picture of proximal and distal effects as well. The authors put preceding proximal and distal rates in conflict, such that in one condition proximal rate was fast and distal rate was slow, and the in another condition this relationship was reversed.[4] In this case of conflict, listeners shift categorization in line with *proximal speech rate*. Nevertheless, conflicting distal rate weakened the proximal effect, in comparison to an experiment where both distal and proximal agreed. Moreover, the authors show that, when proximal and distal contexts do *not* conflict, distal effects are robust even when followed by an interval of uninformative neutral proximal rate. Thus overall, evidence for distal speech rate effects on segmental categorization and segmentation are somewhat mixed, and it is clear that distal and proximal contexts can interact (as in e.g., Reinisch et al. 2011).

[2] Proximal, distal, and global contexts all shape listeners' perception of durational cues in speech, though recent work suggests the mechanisms responsible for these effects may not be the same (Bosker, 2017; Bosker and Ghitza, 2018; Maslowski, Meyer, & Bosker, 2020). For example, a durational contrast account (Diehl & Walsh, 1989; Wade & Holt, 2005) is often offered to explain proximal effects, whereby the "perceived length of a given acoustic segment is affected contrastively by the duration of adjacent segments" (Diehl & Walsh 1989, p 2154), such that a given segment is perceived as *shorter* when following a segment that is relatively long. This sort of localized contrast account is supported by a variety of speech perception findings in which only proximal context is manipulated (e.g., Miller & Liberman 1979; Miller & Volaitis 1989). On the other hand, a growing body of literature suggests that more distal rate effects may be best accounted for by *entrainment* (Doelling, Arnal, Ghitza, & Poeppel, 2014; Luo & Poeppel, 2007), a model in which oscillators encode rate information neurally on the basis of the rate of repetition of roughly syllable-sized envelope fluctuations in the signal (Bosker, 2017; Peelle & Davis, 2012; Pitt, Szostak, & Dilley, 2016). This premise further has recent neurobiological support (Kösem et al., 2018; Kösem, Bosker, Jensen, Hagoort, & Riecke, 2020).

[3] It should be noted more generally that in the domain of *spectral* contrast, the evidence favors a clear precedence of proximal, over distal, context (Stilp 2018, Stilp 2020 for a review)

[4] Reinisch et al. (2011) tested how variation in rate influenced segmentation of ambiguous sequences in Dutch in which, for example, a durational event (e.g., closure duration for [t]) signaled a sequence of two /t/s across a word boundary, or a single /t/, as in "nooit rap" versus "nooit trap" ("never quick"/"never staircase"). A faster contextual rate in this case leads to increased perception of /t/ initial words, that is, closure duration is perceived as relatively long in relation to fast rate, signaling a geminate at the word boundary.

Global rate

Listeners also track temporal information that extends beyond what is heard in a given trial: the *global* temporal traits of the stimuli to which they are exposed in an experiment (Baese-Berk et al., 2014; Jones & McAuley, 2005; Maslowski et al., 2020). One recent example of global effects in rate-dependent perception comes from Maslowski, Meyer, and Bosker (2020), who tested how both distal and global speech rate influenced listeners' perception of Dutch vowel length contrasts. The authors created three contextual speech rates (fast, neutral, slow) in carrier phrases with words which listeners identified as containing long or short vowels. One set of listeners heard only fast and neutral speech rates, while another set of listeners heard only neutral and slow rates, thus varying the global temporal structure across groups. In addition to the expected effect of preceding rate in a stimulus, the global rate manipulation also impacted listeners' perception such that an overall faster global rate (fast and neutral stimuli) led to *fewer* long vowel responses in the neutral condition, as compared to an overall slower global rate (neutral and slow stimuli). In other words, the neutral rate presented in a faster global context sounded relatively slow (reducing long vowel responses), as compared to the neutral rate in slower global rate context, which sounded relatively fast. These and related findings show that tracking of global, long-term patterns can shape rate-dependent perception, with effects that strengthen over time as listeners accumulate exposure to a global rate (Baese-Berk et al., 2014).

Another established role of global temporal context is in how much variation it introduces to listeners across trials in an experiment. For example, Jones and McAuley (2005) implemented a time judgment task in which listeners compared the duration of a standard inter-onset interval (IOI) with a comparison interval, with the standard preceded by base IOIs. Listeners' accuracy in comparing IOIs was assessed while the overall rate of precursor base IOIs was manipulated. Crucially, trial-to-trial variation in rate impacted listeners' accuracy in the time judgment task. When rate was more variable globally (i.e., across trials), accuracy decreased, showing that listeners' tracking of the interval timing characteristics in the stimuli was hindered when the rate of base IOIs varied within the experiment. Accuracy was also directly impacted by the rate of the immediately preceding trial: when a two-trial sequence showed a larger change in rate, accuracy was lower. The authors postulated changes in rate may be disruptive if "listeners become accustomed to a certain 'average pace'" over the course of exposure to multiple trials (see also Baese-Berk et al. 2014; Maslowski et al. 2019), and then must adapt to a departure from this pace, as captured in entrainment models (Large & Jones, 1999). In this view, global variation in speech may hinder listeners tracking of certain patterns (including, potentially, rhythmic patterns), a point discussed further in Section "Experiment 3".

In summary, the relationship between distal and proximal rate effects remains somewhat unsettled in the literature. Moreover, the way in which global temporal structure mediates other rate effects is an active area of research. The present set of experiments will examine these issues in testing how distal *rhythmic* context interacts with proximal, distal, and global variation in speech rate (described in "The present study").

Rhythmic grouping effects

Rhythmic patterns are defined here as a property of the temporal organization of speech, perceived in terms of repeating structure conveyed by modulations in F0 (the fundamental frequency of the speech signal, reflecting the number of glottal cycles per second and corresponding to the perception of pitch), amplitude, or duration (Barry, Andreeva, & Koreman, 2009; Handel, 1993; Hayes, 1995; Jun, 2012). The relevance of rhythmic patterns in speech processing has been well-established in various domains (Brown, Salverda, Dilley, & Tanenhaus, 2015; Cutler & Darwin, 1981; Dilley & McAuley, 2008; Morrill et al., 2014; Quené & Port, 2005).[5] Among these, variation in rhythmic patterns has been shown to shape word segmentation and processing of durational information in the speech signal. Much of the research in this domain has been couched in the *perceptual grouping hypothesis*, which postulates that listeners' grouping of speech material in the signal is influenced by the structure of alternating rhythmic patterns that precede it. For example, Dilley and McAuley (2008) found that word segmentation in an ambiguous string was modulated by distal alternations of F0. Consider an example: the words "foot note book worm" might be segmented as (1) "footnote # bookworm" (where # indicates a word boundary), or as (2) "foot # notebook # worm". The authors found that the preferred grouping of the words shifted based on distal rhythmic context. Specifically, with alternating low (L) and high (H) pitch targets in words preceding the ambiguous string, listeners exhibited a preference to parse out ambiguous strings such that a HL

[5] For example, isochronous timing for linguistic units (e.g., metrically prominent, or stressed syllables) has been hypothesized to aid speech processing (Lehiste, 1977; Hawkins & Smith, 2001), as related to the more general theory of dynamic attending (e.g., Jones 1976; Large & Jones 1999) in which recurrent patterns in a stimulus guide attentional resources and expectations for incoming auditory material. Indeed, regular timing for trochaic (strong-weak) and iambic (weak-strong) syllabic patterns were shown by Quené and Port (2005) to facilitate processing in a phoneme monitoring task (see also (Cutler & Darwin, 1981)), regardless of the sequence type, or its deviation from previous sequences (i.e., an iamb following a series of trochees).

or LH sequence formed a unit. When "foot" carried a HL contour (preceded by a sequence of HL patterning), the following two syllables were parsed as in (2) being grouped together in "notebook". Conversely, when the word "foot" carried only a L target (preceded by a sequence of LH patterning) it was grouped perceptually with the following "note" such that parse (1) above was obtained. In general terms, listeners' preference to perceptually group syllables into words followed a preference to group alternating pitch patterns as a binary unit (LH or HL), and to generate periodic expectations about upcoming material on this basis (see also e.g., Brown et al. 2015; Dilley et al. 2010; Morrill et al. 2014). These effects are hypothesized in this literature to derive from domain-general perceptual organizational principles. Most relevant to the present study, as eluded to above, units alternating in some property (i.e., pitch, duration, amplitude) are perceived as having a sequenced structure (Handel, 1993). For example, a sequence of tones alternating in high and low pitch (HLHLHL) or (LHLHLH) is perceived by listeners to be a repeating sequence of binary tonal units, i.e., (HL) (HL) (HL) or (LH) (LH) (LH), where parentheses indicate a group (e.g., Woodrow 1909; Woodrow 1911). Alternations in the temporal domain have also been shown shape the perception subsequent intervals (Jones, 1976; Jones & McAuley, 2005; McAuley & Jones, 2003).

The present study

Thus far, we have reviewed some findings that pertain to distal, proximal, and global speech rate effects in rate-dependent perception, as well the as influence of rhythmic alternations in grouping units in the speech stream. This section describes two directly relevant studies that test how rhythmic and speech rate influences combine in listeners' perception of durational information.

Morrill, Dilley, McAuley, and Pitt (2014) provide clear evidence for a grouping role of rhythmic patterning which is additive with speech rate effects in durational processing. The authors tested how listeners interpret the presence or absence of a function word, where an ambiguous region in a sentence could be interpreted as having or lacking said function word. For example, in the sentence "Zach plans he'll be here for an hour or more", the lack of a function word would be "Zach plans he'll be here for an hour more". An acoustically ambiguous rendition in which the function word is reduced and coarticulated with the preceding word could be interpreted as either of the sentences above based on (temporal) context. Morrill et al. (2014) manipulated preceding speech rate and found that slower rates lead to fewer perceived function words: when preceding rate is slow, an ambiguous region sounds relatively fast, too fast to include a function word (see also Dilley and Pitt 2010). Moreover, the authors manipulated rhythmic context, using f0 alternations distal to the critical region, which alternated across syllables in either a binary (HL) (HL) (HL) or ternary (HLL) (HLL) pattern. Preceding ternary grouping increased function word reports, such that listeners expected the ambiguous region to contain three syllables including a function word, which shared the relevant ternary pattern (HLL, over three syllables). On the other hand, preceding binary f0 alternations decreased function word reports, as listeners expected the ambiguous region to contain only two syllables with f0 matching the preceding pattern (HL). Importantly, these rhythmic effects were *additive* with the effects of preceding rate: a faster context with ternary rhythm showed more function word reports than a faster context with binary rhythm, which showed more than slower contexts with ternary and binary rhythm (with the same effect of rhythm at slower rates). These results thus lead to an expectation of additive rhythmic and rate effects, with both impacting listeners' interpretation of temporal information.

The picture is complicated somewhat by Kidd (1989), who demonstrated comparable rhythmic expectancy effects by varying alternating durational patterns in a stretch of speech and testing how it impacted listeners' perception of a following voice onset time (VOT) continuum (ranging from /k/ to /g/). Kidd manipulated a precursor in which stresses alternated, as indicated by capitalization here: BIRD in the HAND is worth TWO in the [target] (where the target was categorized as /ki/ or /gi/). The target sound is notably in the stressed position in relation to preceding context, immediately preceded by unstressed material. Kidd manipulated the rate of both stressed and unstressed portions in the utterance, both in the utterance overall and of just stressed or unstressed portions (such that, in some stimuli, stressed portions were fast, and unstressed portions were slow). As would be expected, an overall fast context generated lower VOT boundaries (i.e., more /k/ responses) as compared to an overall slow context. Notably, contexts in which *only* stressed location rate was fast (and unstressed location rate was slow) showed lower VOT boundaries as compared to stimuli in which stressed location rate was slow (and unstressed location rate was fast). In other words, listeners shifted categorization in line with rate of stressed syllable and the rhythmic pattern with which they were presented, disregarding the rate of the unstressed portion in the precursor (in Kidd's Experiments 2 and 3). This is particularly notable as it is the unstressed rates which are the most proximal to the target, and their rate differences are in competition with the rate conveyed by more distal stressed locations in the utterance, suggesting that distal rhythmic context is taking precedence over proximal contrasts (essentially the opposite of what was found by Reinisch et al. 2011). Kidd's analysis focuses on the overall effects of stressed and unstressed word rates across multiple conditions (pooling, e.g., overall fast and fast stressed conditions and comparing them to overall slow and slow stressed conditions to test the effect of stressed location rate). As such, the paper does not provide a direct statistical comparison of conditions which put proximal rate and distal (rhythmic) rate alternations in conflict (e.g., stressed fast and stressed slow conditions only). In Kidd's Experiment 1, these two conditions show essentially comparable VOT boundaries, however in two subsequent experiments (with the same stimuli) the fast stressed syllable condition (with slow proximal context) shows numerically lower VOT boundaries, as compared to the slow stressed syllable condition (with fast proximal context), suggesting that distal rhythmic rate alternations are taking precedence over proximal context, as described above. However, given that these conditions are not directly compared statistically with one another, and that this effect seemingly does not occur in Experiment 1, this is not entirely clear. The present study will present a direct comparison of the relative importance of proximal, and distal (rhythmic) contexts, when they conflict.

Kidd also created conditions in which proximal contexts deviated from established rhythmic patterns and speech rates. When established patterns were defied, listeners relied solely on the deviant proximal information, even when distal information was in clear conflict. For example, in one set of conditions, listeners heard an all-fast precursor, except for the immediately pre-target unstressed location, which was slow. In another, they heard an all-slow precursor, except for the immediately pre-target unstressed location which was fast. Here, only the proximal context mattered (Kidd's Experiment 3, cf. Reinisch et al. 2011). Similarly, reliance on proximal context was observed when it deviated from alternating rhythmic patterns (Kidd's Experiment 2), essentially reversing the rhythmic expectancy effect observed when preceding rhythmic patterns were consistent. This finding raises the pertinent question of how "fragile" effects of rhythmic expectancy are: what sorts of deviations in context are tolerated by listeners? Given that Kidd's manipulations of rhythmic alternations and speech rate were not orthogonal, we do not currently know the extent to which *durational* rhythmic alternations (conveyed by alternating durations as opposed to F0) and rate effects are additive in the sense of Morrill et al. (2014), and this remains a pertinent question given Kidd's finding that deviating proximal durations are capable of undoing distal rhythmic effects. Also of note, Kidd blocked the presentation of his stimuli by precursor type such that listeners did not hear trial-to-trial variation in the precursor (which varied in rate and rhythm), a point that will be returned to in Section "Experiment 3".

The present study will build on the literature outlined above by explicitly comparing cases where proximal and distal rhythmic cues are in conflict (Experiment 2), once independent proximal effects are established (Experiment 1). This will allow us to directly assess the relative importance of distal rhythmic alternations and proximal speech rate effects. Experiment 3 will then explore how orthogonal variation in overall speech rate (distal and proximal) interacts with rhythmic effects (also introducing global rate variation). Experiment 4 extends Experiment 3 to test this two-by-two rate-by-rhythm design with proximal rate controlled across conditions. Results from these Experiments will accordingly help us understand how rhythmic perceptual grouping mediates rate-dependent perception, and how speech rate effects (proximal, distal, and global) interact with (or interfere with) rhythmic grouping effects.

Experiment 1

The goal of Experiment 1 was to elicit a proximal rate (durational contrast) effect, based on the duration of a single syllable preceding the target sound. Experiment 2 will then test if the addition of distal context changes the observed effect. This will accordingly allow us to assess the relative importance of both contexts, building on the results of Kidd (1989) as outlined above.

The test case adopted in all experiments reported here is that of vowel duration as a cue to coda obstruent voicing in American English. Vowels are substantially longer before voiced obstruents, and this is a reliable cue to voicing for listeners (Chen, 1970; Heffner et al., 2017; Raphael, 1972; Steffman, 2019), though as noted by Heffner et al. (2017), relatively few studies have examined this as a rate-dependent cue. In all experiments reported here, listeners categorized a vowel duration continuum ranging from the English word "coat" to the English word "code". These words were chosen to be roughly frequency matched, based on frequency counts from the SUBTLEX-US corpus (Brysbaert & New, 2009).[6]

Materials for Experiment 1 and 2

Following Hay and Diehl (2007) and Hoequist and Kohler (1986), synthetic non-word speech was used to create various rhythmic patterns. Stimuli were created by Pitch-Synchronous Overlap Add (PSOLA) re-synthesis (Moulines & Charpentier, 1990) of the speech of a male speaker of American English, using the software Praat

[6] The $\log^{10}$ frequency of "coat" is 3.33, the $\log^{10}$ frequency of "code" is 3.43.

(Boersma & Weenink, 2020).[7] In creating the vowel duration continuum, the word "code" was excised from the carrier phrase "I'll say code now". Audible voicing after closure was removed, such that perception of voicing was dependent on vowel duration. The vowel duration of the original token was approximately 140 ms. Prior to the manipulation of vowel duration, pitch on the target was monotonized to have a constant pitch value of 131 Hz. The continuum was synthesized by manipulating the vocalic portion of the target word (all continuum steps were created by re-synthesis). The continuum had five steps, each separated by 15 ms. The shortest endpoint of the continuum was set to be 90 ms (corresponding to a "coat" response), the longest endpoint of the continuum was set to be 150 ms (corresponding to a "code" response). These endpoint durations were determined based on a pilot experiment.

To allow for tight control of the durational properties of the precursor, a CV syllable [$\text{t}^\text{h}\alpha$], produced by the same speaker, was re-synthesized to have one of two vocalic durations, 75 ms or 150 ms. Only vowel duration was manipulated, VOT was identical in all precursor syllables. In Experiment 1, only a single precursor syllable preceded the target sound. In the *short-long* condition, which will be referenced by the same name in Experiment 2, the target was preceded by a single short (75 ms vowel duration) syllable. In the *long-short* condition, the target was preceded by a single long (150 ms vowel duration) syllable. The stimuli used for Experiment 1 are shown as the *boxed region* in Fig. 1. The goal of Experiment 1 is accordingly to confirm that this localized durational change is sufficient to generate a proximal rate (contrast) effect (Bosker, 2017; Wade & Holt, 2005). This predicts that the target vowel would be perceived by listeners as relatively short in the long-short condition, leading to *decreased "code" responses* therein.

In Experiment 2, an extended precursor of seven syllables preceded the target sound. Short and long syllables were iterated in two different patterns, as shown in Fig. 1. The number of precursor syllables was chosen to be comparable to previous studies in the perceptual grouping literature (Dilley & McAuley, 2008; Morrill et al., 2014). In creating the short-long condition, a short syllable was placed preceding a long syllable to create at short-long sequence. This pattern was then repeated three times. In a final, fourth foot, a short syllable was followed by the target sound (i.e., the two syllables from the stimuli in that condition in Experiment 1). The target was thus grouped with a preceding short syllable to form the second syllable of a short-long sequence (see Fig. 1). In creating the long-short condition, the relative ordering of long and short precursor syllables was switched such that three trochaic feet preceded the final foot, which consisted of a long syllable and the target sound (see Fig. 1). The two conditions thus present different rhythmic structures preceding the target, and differ in the implied status of the target, as either the second syllable in a short-long, or long-short unit. All syllables were separated by 50 ms of silence, and the temporal alignment of syllables was by their acoustic onset. So that duration alone distinguished the conditions, the average intensity and pitch (which was monotonized at 131 Hz) of every syllable in the stimulus was manipulated to be the same. The seven-syllable precursor had a total duration of approximately 2 s. Predictions related to this extended precursor used in Experiment 2 will be discussed in Section Experiment 2. In both experiments there were thus a total of ten unique stimuli (2 precursors × 5 continuum steps).

Participants and procedure

Thirty-two participants were recruited for Experiment 1. All were students at UCLA and received course credit for their participation. The procedure was a simple 2AFC task, in which participants were presented with an auditory stimulus and categorized the target sound as "coat" or "code". The platform Appsobabble (Tehrani, 2020) was used to control stimulus presentation and collect participant responses. Participants completed the task seated in front of a desktop computer a sound-attenuated booth. Stimuli were presented at a mean level of 70 dB binaurally via a PELTOR™3M™listen-only headset. The target words were represented orthographically on the computer monitor, with each target word centered in each half of the monitor. The side of the screen on which the target words appeared was counterbalanced across participants, such that for half of the participants "code" was on the left, and for half "code" was on the right. Each unique stimulus was presented ten times, in randomized order, for a total of 100 trials in the experiment. The procedure took approximately 10 to 15 min to complete.

Results and discussion

Results were assessed statistically using a Bayesian logistic mixed effects regression model implemented in the *brms* package in R (Bürkner, 2017). The model predicted listeners' categorization response (with "coat" mapped to 0

[7] PSOLA synthesis allows for manipulation of pitch and duration by analyzing the speech signal into pitch-synchronous Hanning-windowed sub-units for voiced portions of speech. Duration is manipulated by the duplication or reduction of windowed units. Pitch is manipulated by moving units closer together, raising pitch, or further apart, lowering pitch. The output signal is constructed via convolution of units with the overlap add technique (Crochiere, 1980; Oppenheim & Schafer, 1975). PSOLA is used frequently in perception experiments where pitch and/or duration are manipulated (e.g., Bosker 2017; Dilley & McAuley 2008; Reinisch & Sjerps 2013; Steffman & Jun 2019).

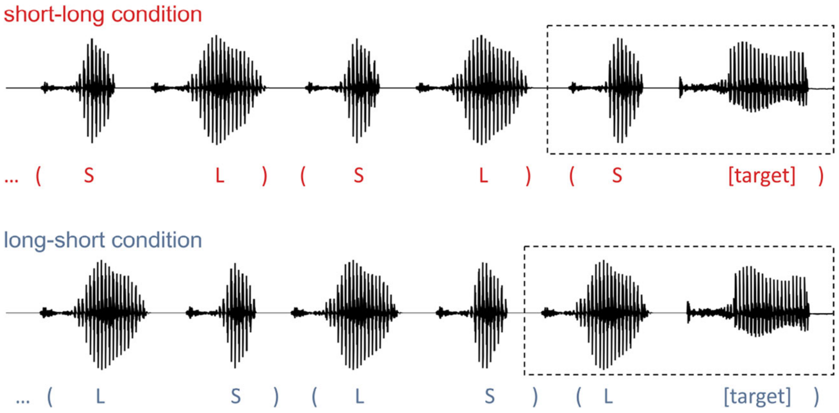
short-long condition
... (S L) (S L) (S [target])
long-short condition
... (L S) (L S) (L [target])

Fig. 1 Waveform representations of each of the two conditions used in Experiments 1 and 2, showing five (of seven total) precursor syllables. L (long) and S (short) refer to vowel duration in the precursor syllables, all [$t^h\alpha$]. Listeners' hypothesized perceptual grouping is represented by ( ) surrounding the precursor syllable labels. The proximal context, which was the only precursor present in Experiment 1, is boxed. The longest step from the vowel duration continuum is shown as the target sound in these examples

and "code" mapped to 1), as a function of contrast-coded rhythm condition (short-long mapped to -0.5, long-short mapped to 0.5), and continuum step which was scaled and centered at zero. The interaction between these two fixed effects was also included in the model. The default prior distribution, an improper uniform distribution over real numbers, was used. In assessing the model output, if the 95% credible interval (CI) for a given effect is observed to exclude zero, an effect is taken to have a meaningful (credible) impact on categorization responses (see Bürkner 2017; Vasishth et al. 2018 for details on Bayesian models). Random effects in the model were specified as by-participant random intercepts and random slopes for all fixed effects. The full model output is shown in Table 2, in Appendix 2 which contains model summaries for all analyses reported here.

Categorization responses in Experiment 1 are shown in the left panel of Fig. 2. An expected effect of the vowel duration was observed, whereby increasing vowel duration along the continuum increased the log-odds of a "code" response (β=1.14, 95% CI= [0.79, 1.50]). A credible effect of precursor was also observed, such that the long-short condition exhibited significantly decreased "code" responses (β=-0.36, 95% CI= [-0.73, -0.004]). The interaction between these two fixed effects was not credible. The results of Experiment 1 thus evidence a canonical

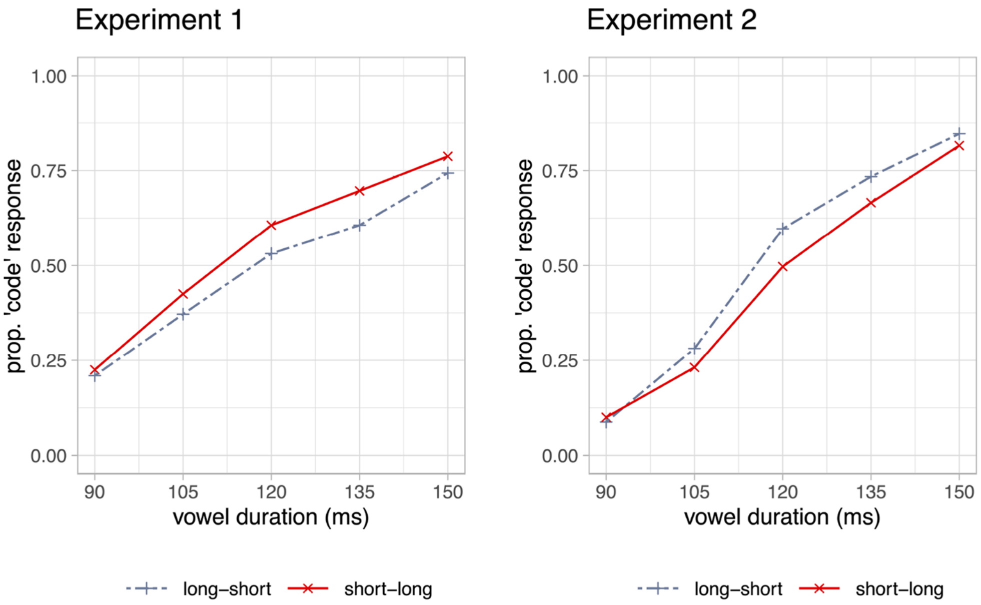

Fig. 2 Categorization responses in Experiment 1 (*left panel*) and Experiment 2 (*right panel*), at each vowel duration step on the continuum, split by precursor condition

proximal rate (contrast) effect, showing that when a longer syllable precedes the target, overall longer vowel duration is required by listeners to perceive voicing, or put differently, the target vowel sounds shorter following a longer preceding syllable (resulting in decreased "code" responses overall).

Experiment 2

The goal of Experiment 2 was to test if the addition of distal context exerts an influence on their categorization of the target sound, in line with Kidd (1989). If listeners group the precursor sequence into two-syllable units as discussed in "Rhythmic grouping effects", the target sound's status in such a grouping varies across conditions. In the short-long condition, the target forms the implied second longer unit in a short-long sequence, where the opposite is true in the long-short condition. If this exerts an effect on their perception of target vowel duration, it is predicted that overall *longer* vowel durations should be required for a voiced "code" percept in the short-long condition, where the expectation of a relatively long target is generated by the precursor. On the other hand, in the long-short condition, overall *shorter* vowel durations should be required for a voiced percept, given that the rhythmic context implies a short target vowel duration. If this result is obtained it would effectively replicate Kidd (1989), while showing definitively this effect is taking precedence over proximal contrasts (seen in Experiment 1). Empirically, this prediction can be stated as *increased "code" responses* in the long-short condition (where overall shorter vowel durations should be required for a voiced percept), the opposite of the effect observed in Experiment 1.[8]

Participants and procedure

32 participants were recruited for Experiment 2, from the same population as Experiment 1. None of these participants had taken part in Experiment 1. The procedure was identical to Experiment 1.

Results and discussion

Statistical assessment of results in Experiment 2 was the same as that in Experiment 1. Fixed and random effects were specified in the same fashion. The model output is shown in Table 3 (in Appendix 2) and categorization responses are plotted in the right panel of Fig. 2. A credible effect for continuum step was found such that "code" responses increased as vowel duration increased, as expected (β=1.79, 95% CI = [1.48, 2.11]). The rhythmic precursor (short-long versus long-short) also showed a credible effect whereby a preceding long-short context showed increased "code" responses (β=0.33, 95% CI = [0.09, 0.57]). Comparing these results to Experiment 1, we can see a complete reversal in the effect (compare panels in Fig. 2). Importantly, in both of these experiments, proximal contexts were identical, showing that when distal rhythmic context is present, it takes precedence over proximal difference in duration, effectively wiping out the contrast effect seen in Experiment 1. We've thus seen so far that, in analogous fashion to Kidd (1989), rhythmic patterns and grouping mediate perception of durational cues. Experiment 3 builds on these findings by introducing orthogonal variation in overall speech rate, allowing us to address questions of additivity raised by Morrill et al. (2014), and more generally, to examine the relationship between overall rate and the effects observed thus far.

Experiment 3

Experiment 3 crossed the rhythmic manipulation from Experiment 2 (short-long/long-short) with a speech rate manipulation which varied the rate of all syllables in the precursor (to be neutral or fast). As described in "The present study", Morrill et al. (2014) give us a clear expectation that rate effects and rhythmic grouping effects should be additive. In the simplest case this would present itself as a four-way distinction across the four conditions in which the most "code" responses are obtained for the long-short and fast rate condition, while the least are obtained for the short-long and neutral rate condition, with the other two conditions falling in between. This outcome would essentially replicate Morrill et al. (2014), in showing that these two influences jointly combined to shape perception

[8] The Experiment 2 stimuli can be considered in terms of the so-called "iambic/trochaic law" (e.g., Hayes 1995). This refers to a tendency for listeners to perceive alternating sequences as iambic or trochaic on the basis of the alternating acoustic medium. Alternations in intensity have been suggested to be perceived generally as *trochaic* (strong-weak) while alternations in duration are generally perceived as *iambic* (weak-strong). This is relevant to the present design in the sense that the alternations employed are purely durational, and accordingly, if the iambic/trochaic law obtains, one might predict that both conditions here would be perceived as alternating in a weak-strong fashion. However, previous research has suggested this pattern is only a tendency (Crowhurst & Olivares, 2014), and can be overridden. In particular, Hay and Diehl (2007) found a "strong tendency" for listeners' perception of rhythm to be based on the starting pattern of a given sequence, i.e., the structure of the first two units in the pattern. This led the authors in that study to create onset masking in which stimuli were gradually faded in to obscure the starting point of the sequence. In the absence of such masking, as in the present stimuli, it is assumed that listeners' perception of sequence structure will be based largely on the starting pattern in the sequence. The results from Experiment 2 further support this conclusion.

of duration. It is worth remarking here again that Morrill et al. implemented their rhythmic variations in F0, such that the rhythmic manipulation didn't introduce *temporal* variation in the stimuli. Rate information and rhythmic information therefore never conflicted in the way they did in Kidd (1989), or in Experiment 2.

Relatedly, the introduction of variable temporal structure raises an alternative possibility. As described in "Introduction" and 1 when a portion of a stimulus (as in Kidd 1989) or stimuli across multiple trials (as in Jones and McAuley 2005), deviate from regularity, listeners' tracking of temporal patterns is diminished. In the present case, this can be considered on two scales: first within a presented stimulus, a fast preceding rate is substantially faster than the unaltered duration of the target vowel continuum overall (see Fig. 3). Changing from a fast precursor target sound might inhibit temporal tracking across the precursor (as in Kidd), predicting diminished rhythmic effects in the fast condition, as compared to the neutral condition. On a global scale, by varying rate across trials within an experiment, the global temporal structure of the experiment has become variable (as compared to Experiment 2). Alteration of global temporal traits (i.e., over the course of the experiment and trial to trial) has been shown to impact listeners' sensitivity to temporal patterns in speech (Barnes & Jones, 2000; Jones & McAuley, 2005; Jungers, Palmer, & Speer, 2002; Warren, 1985). Recall that Jones and McAuley (2005) found reduced accuracy in a time judgment task when rate was variable, both at a global scale but also based on the rate of an immediately preceding trials (the more divergent in rate a two-trial sequence was, the less accurate temporal judgments were). Though the present study utilizes a very different task than that of Jones and McAuley (2005), these findings bear on Experiment 3, with overall (and randomized, trial-to-trial) speech rate variation. If perceptual grouping in the temporal domain (cf. Morrill et al. 2014) is hindered by global rate variation, we might expect to see a reduction of the rhythmic effects seen in Experiment 2 in both rate conditions. This is especially important to consider in relation to Kidd (1989), who *blocked* stimulus presentation by precursor type in all of his Experiments, such that listeners heard (randomized) continuum steps in a single precursor condition only, as a single block. Contextual trial-to-trial temporal variability was thus not present in Kidd's study.

Materials

The materials for Experiment 3 were the same as in Experiment 2, with the added overall speech rate

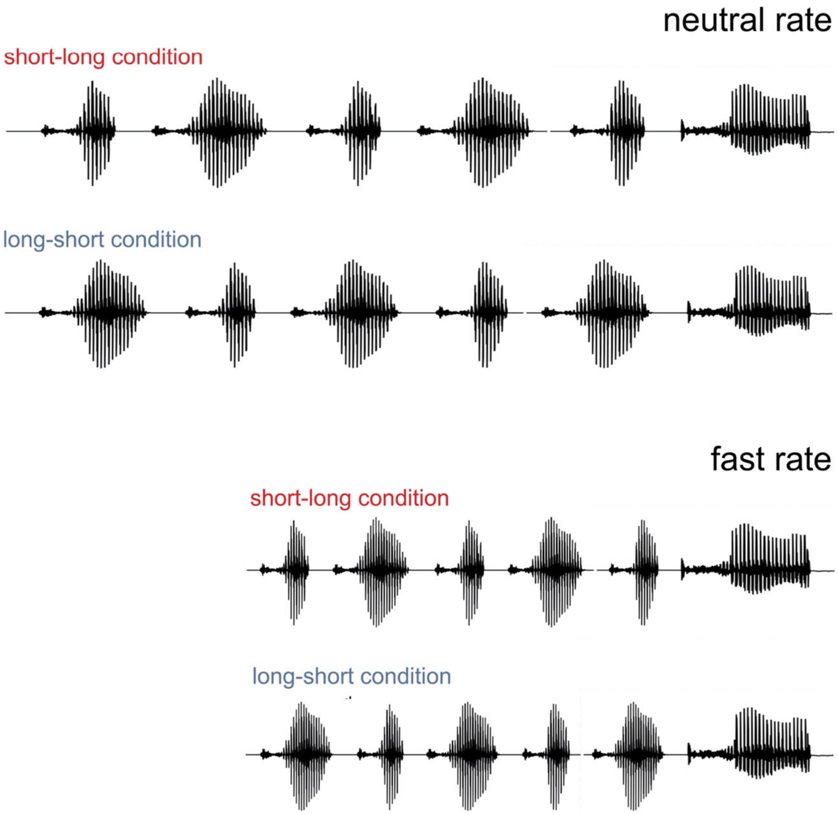

Fig. 3 Waveform representations of the four precursor conditions in Experiment 3, labeled by rate and rhythmic status, and showing five (of seven total) precursor syllables. Note the neutral rate condition is identical to the stimuli in Experiment 2. See Fig. 1 for reference

manipulation in the precursor. This manipulation was accomplished by linear compression of *all precursor syllables*, including proximal material. The newly created "fast" condition was set to be 66% of the duration of the original precursor. The original precursor constituted the "neutral" condition. The fast condition precursor had a total duration of approximately 1.3 s, with a rate of approximately 5.38 syllables per second (both fast and neutral conditions have syllable rates that are within the range that rate effects occur (Bosker & Ghitza, 2018)). This additional manipulation resulted in 20 unique stimuli that were used in Experiment 3 (2 rate conditions × 2 rhythm conditions × 5 continuum steps). These manipulations are shown in Fig. 3.

Participants and procedure

Thirty-two participants were recruited from the same population as previous experiments. None of these participants had participated in Experiment 1 or 2. The procedure was the same as previous experiments (of note, presentation of all stimuli was completely randomized as in previous experiments). There were twice as many trials (200) due to 2×2 crossing of precursor conditions. The procedure took 20 to 25 min for participants to complete.

Results and discussion

The statistical assessment of the results in Experiment 3 was the same as that in previous experiments. The model was specified to include precursor rate as a fixed effect (fast mapped to -0.5, neutral mapped to 0.5), as well as all previous fixed effects, and all interactions. The random effect structure in the model included all of these fixed effects and interactions as random slopes. The model output is shown in Table 4 in Appendix 2. Results are plotted in Fig. 4.

As in previous experiments, the vowel duration continuum showed an expected credible effect on categorization (β=2.24, 95% CI = [1.86, 2.64]). Precursor rate and rhythm were observed to enter into a credible two-way interaction (β=-0.38, 95% CI = [-0.77, -0.03]). To inspect the interaction, model contrasts were extracted from the interacting terms using the package *emmeans* (Lenth, Singmann, Love, Buerkner, & Herve, 2018). The estimated marginal effects obtained with this method provide the median of the posterior distribution for a given contrast accompanied by 95% highest posterior density credible intervals. The output for this assessment, testing the effect of rate in each rhythm condition, and the effect of rhythm in each rate condition, is shown in Table 1. First, with respect to the effect of rate, we can see that changes in precursor rate only showed a credible effect in the long-short rhythmic condition, also visible in Fig. 4. The effect is in the expected direction, such that a neutral (relatively slower) precursor decreased listeners' "code" responses (in the long-short condition).

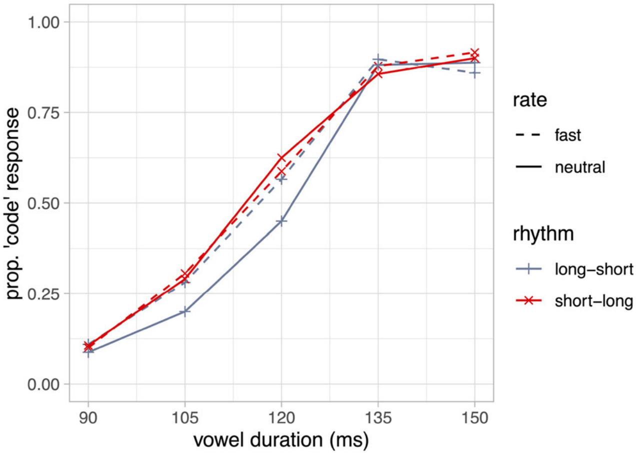

Fig. 4 Categorization responses in Experiment 3, showing categorization split by rhythm and rate conditions, labeled at right

Table 1 also shows that rhythm only has a credible effect in the neutral rate condition, though estimates in both conditions show the same directionality, giving weak evidence for an effect of rhythm when rate is fast. The directionality of the rhythmic effect is such that a preceding long-short context shows credibly decreased "code" responses (in the neutral rate condition). This is a total reversal of the effect seen in Experiment 2, and is instead in line with the proximal contrast effect observed in Experiment 1. This is particularly notable for the effect of rhythm in the neutral rate condition, because the stimuli in the neutral rate condition in Experiment 3 were identical to the stimuli in Experiment 2 (Experiment 3 additionally including the fast stimuli). This point is discussed further below. Both rhythm and rate also evidenced credible main effects in the model (rate: β=-0.19, 95% CI = [-0.34, -0.04]; rhythm: β=-0.32, 95% CI = [-0.50, -0.14]). However, the interaction and comparisons shown in Table 1 would suggest that both main effects are driven primarily by an

Table 1 Contrasts showing the effect of rate in each rhythm condition and rhythm in each rate condition in Experiment 3

Effect of rate			
Rhythm condition	Estimate	L95% CI	U95% CI
Long-short	**−0.38**	**−0.63**	**−0.15**
Short-long	0	−0.23	0.25
Effect of rhythm			
Rate condition	Estimate	L95% CI	U95% CI
Neutral	**−0.50**	**−0.77**	**−0.24**
Fast	−0.13	−0.36	0.12

effect within a certain condition (particularly for the effect of speech rate).

What might explain this pattern of results? Figure. 4 shows that one particular condition, the long-short neutral condition, stands out from the rest, with decreased "code" responses. In fact, pairwise comparison of contrasts which compared across each combination of rhythm and rate found that the only condition which differed credibly from the others was the long-short neutral rate condition (see Table 5 in Appendix 2). In this light, we can consider how this condition differs from the others. In comparison to the short-long neutral condition, which is matched for rate, it evidences the effect of proximal contrast found in Experiment 1, described in more detail below. In comparison to both fast conditions, it varies in overall rate (including distal rate), but notably also varies in proximal context. One possibility is accordingly that these results are reducible to proximal context effects, where a longer preceding vowel in the long-short neutral condition leads to decreased "code" responses relative to all other conditions. The lack of a difference across these other conditions could be reduced to proximal context as well under the assumption that differences between proximal durations are small enough not to generate effects. The pre-target vowel in the short-long neutral condition is 75 ms in duration, as compared to approximately 50 ms in the short-long fast condition, where we find weak evidence for a rate effect (see Table 1). In comparison, the pre-target vowel is approximately 100 ms in the long-short fast condition. If differences between these very short vowel durations are less salient, the lack of an effect across them could result from (a lack of) proximal context effects, in comparison to the 150 ms pre-target vowel in the long-short neutral condition, which is substantially longer than all other pre-target durations. Experiment will 4 test this idea by controlling the duration of the immediately pre-target syllable.

As mentioned above, Experiments 2 and 3 present an interesting point of comparison given that they contain similar stimuli (and identical stimuli, in the case of the neutral rate condition in Experiment 3), yet show a reversal of the effect of rhythmic context. This is shown concretely in Fig. 5, which plots overall "code" response (collapsed across the continuum), for the effect of rhythm in Experiment 2, and the Effect of rhythm in Experiment 3 at both rates. As noted above, the neutral rate stimuli in Experiment 3 are identical to the stimuli in Experiment 2, however the inclusion of fast rate stimuli in Experiment 3 has completely reversed the rhythmic effect. This finding suggests that global rate variability has effectively eliminated the rhythmic grouping effects seen in Experiment 2, such that listeners revert to their sensitivity to proximal context. In line with Jones and McAuley (2005), this suggests the variability across trials in the temporal domain hinders listeners' ability to track rhythmic patterns, a clear argument for considering global contexts in testing rhythmic effects. A statistical comparison of Experiments 2 and 3 was carried out with a combined model, and is reported on in Appendix 1.

Experiment 3 offers a notable departure from the additive effects of rhythm (cued by F0) and speech rate documented by Morrill et al. (2014). It also speaks to Kidd (1989), who blocked stimulus presentation by precursor type (where precursors varied in both rate and rhythmic structure). In comparison to Kidd in particular, the present result suggests that an absence of blocking (by rate or rhythm), introducing global, trial-to-trial variation, is responsible for eliminating the effects of rhythmic context seen in

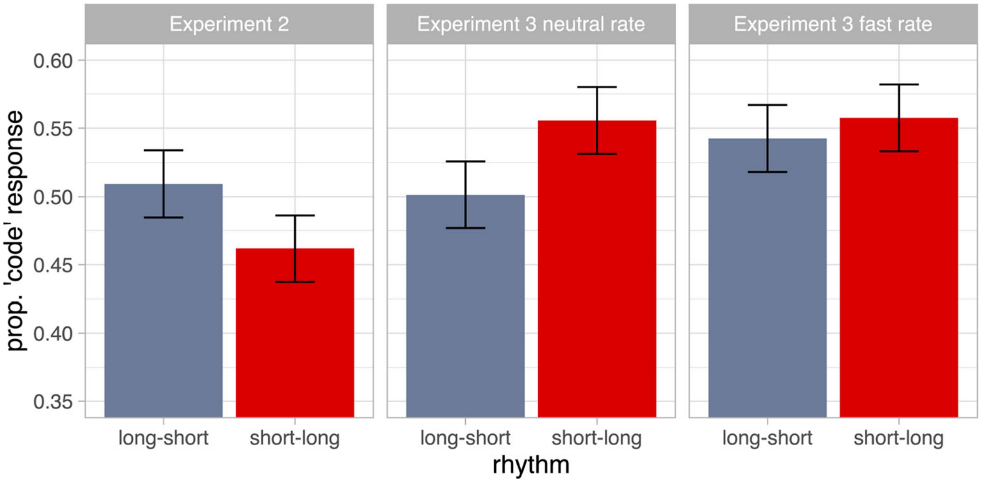

Fig. 5 Overall responses pooled by continuum step, for Experiment 2 (*left panel*), the neutral rate condition in Experiment 3 (*middle panel*), and the fast rate condition in Experiment 3 (*right panel*), split by rhythm condition. *Error bars* show 95% CI. Note that Experiment 2 and Experiment 3 neutral rate are acoustically identical stimuli

Experiment 2. We thus have evidence that global speech rate variation effectively impedes (temporal) rhythmic grouping, a point that is discussed further in "General discussion". The context effect which was observed instead is in line with the proximal contrast effect seen in Experiment 1. However, as discussed above, the extent to which this effect is attributable to proximal versus distal rate is unclear given that both proximal and distal context varied together in Experiment 3.

Experiment 4

The data thus far has shown that, in the face of global speech rate variation, listeners give up tracking rhythmic patterns in a stimulus, but what do they rely on instead? As described above, one possible explanation for the results of Experiment 3 is that they can be reduced to variation in proximal context: the syllable immediately preceding the precursor. To test this, Experiment 4 made a slight modification to the Experiment 3 stimuli, neutralizing the duration of this immediately pre-target syllable across the four conditions shown in Fig. 3.

Materials

The elimination of differences in proximal context was accomplished simply by taking the average of the four immediately pre-target syllable durations which varied across rhythm and rate conditions, resulting in a pre-target vowel duration of approximately 93 ms. The stimuli were otherwise identical to Experiment 3, and will be referred to by the same names.

With differences in distal context only, we will be able to address the extent to which the results of Experiment 3 can be reduced to variation in proximal duration as discussed in "Results and discussion". This will allow us to understand to what extent listeners are sensitive to distal and proximal speech rate variation when the co-vary together as they did in Experiment 3, and also to further test the interaction between rhythm and rate. We can consider several possible outcomes. First, we might expect distal rate to impact perception, while the influence of proximal differences (previously manifested across rhythm conditions) disappears, in line with e.g., Reinisch et al. (2011). Given that F0-based rhythmic grouping in word segmentation persists over a neutral proximal context (Dilley & McAuley, 2008), we can observe if the rhythmic grouping effect re-emerges when proximal context is controlled (though this would go against what was found by Kidd 1989).

Alternatively, we can note the rate-neutralized syllable preceding the target word presents a deviation from the rhythmic and rate patterns in the precursor, being a duration of no other preceding syllable: too long to match the short-long pattern; too short to match the long-short pattern (at both rates in each case). Deviation from the established alternation of long and short syllables, and also the preceding rate, might draw listeners' attention to proximal context, as found in Kidd (1989). If it is the case that listeners are focusing on proximal differences in this regard, distal effects might reduce or disappear, as was found by Kidd (1989).

Participants and procedure

Thirty-five participants were recruited from the same population as previous experiments. None of these participants had participated in Experiment 1, 2 or 3. The procedure was identical to Experiment 3.

Results and discussion

A model with the same structure as that in Experiment 3 was used. The model output is shown in Table 6 in Appendix 2. Results are plotted in Fig. 6. As in all previous experiments, the vowel duration continuum showed a credible effect on categorization (β=1.71, 95% CI = [1.35, 2.08]). Notably, in a departure from the results of Experiment 3 there was no effect of precursor rate (β=-0.06, 95% CI = [-0.31, 0.19]). However, the effect of precursor rhythm was credible, with the long-short condition showing credibly decreased "code" responses (β=-0.15, 95% CI = [-0.31, -0.01]), as was also seen in Experiment 3.

First, consider the lack of a rate effect seen here in Experiment 4. With proximal duration controlled in Experiment 4, we see no impact of (distal) speech rate. This

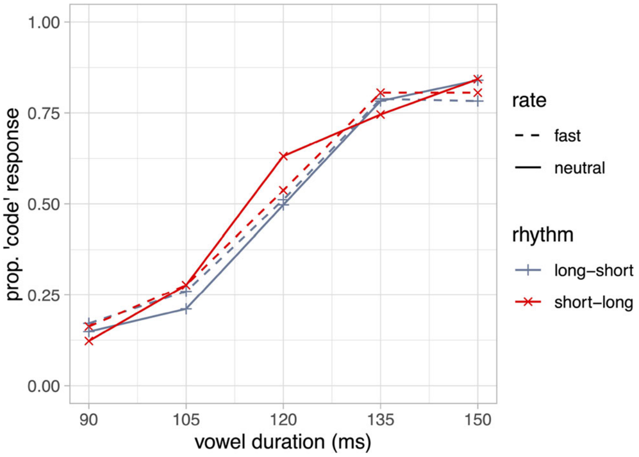

Fig. 6 Categorization responses in Experiment 4, showing categorization split by rhythm and rate conditions, labeled at right

suggests that in Experiment 3 the speech rate effect was primarily linked to proximal context such that it is no longer present when proximal context is invariant. This is notable in that it offers a departure from other studies in which there is a clear persistence of distal rate when proximal context is controlled (e.g., Reinisch et al. 2011). With this in mind, we can turn to the observed effect of rhythm condition. The directionality of the effect is consistent with the proximal rate effect observed in Experiments 1 and 3, which is perhaps surprising given that proximal context is invariant across conditions. However, Kidd (1989) offers a comparable finding which can explain this effect. As described in "The present study", Kidd finds that when proximal context deviated from rhythmic patterns established by preceding material, listeners shifted categorization in line with perceived proximal rate, even when its duration did not vary. Kidd states that "these effects are the result of an enhancement of the perceptual effect of the final section of speech due to the deviation from the articulatory rate established by the immediately preceding unstressed and stressed sections of speech" (p 742). For example, Kidd (in his Experiment 3) found that an all fast rate context differed from a mostly slow rate context with fast proximal rate (the proximal context being the same across conditions). The mixed-rate context showed *decreased* VOT category boundaries, such that listeners perceived a faster rate when the pre-target fast context deviated from preceding slow context, as compared to all fast preceding context. Put differently, the fast immediately pre-target context sounded even faster in contrast to preceding slow rate (and perhaps was focused on due to its deviation from the established rate), and this in turn shifted target categorization. Indeed, in Experiment 4, when listeners hear the transition between the pre-target syllable and the material that comes before it, the pre-target syllable will sound relatively short in the short-long conditions where it is preceded by a longer syllable. On the on other hand, the pre-target syllable in the long-short condition is preceded by shorter syllable, and should in contrast sound relatively long (a proximal contrast effect). A longer perceived pre-target syllable in the long-short condition, perceived as longer by virtue of its contrast with preceding material, would decrease "code" responses, as observed. As such, this result supports the claim that deviations from temporal patterns can draw listeners' attention, effectively eliciting a proximal contrast effect (where deviating proximal rate is interpreted relative to preceding material), analogous to what was found by Kidd (1989).

Experiment 4 therefore shows that the results of Experiment 3 are likely best explained as deriving primarily from proximal context, as rate condition effects disappear when proximal context is controlled. More interestingly, Experiment 4 shows that even with proximal context neutralized, listeners exhibit proximal contrast effects, effects that are hypothesized to result from a proximal syllable's deviation from established preceding patterns.

General discussion

The experiments reported in this paper have addressed the extent to which rhythmic pattern mediates listeners' perception of durational cues, and how these effects interact with speech rate variation at various scales. In comparing Experiments 1 and 2, we saw that expected proximal rate effects shift listeners' perception of vowel duration as a cue to voicing, when just one syllable precedes a target word. Experiment 2 however, showed that this effect could be reversed by the addition of distal context which conveyed rhythmic patterning. Taking these two experiments alone, we can conclude that effects related the rhythmic grouping appear to take precedent over proximal speech rate. However, Experiments 3 and 4 add nuance to this conclusion. Experiment 3, in which orthogonal overall speech rate variation was introduced, showed a reversal of the effect in Experiment 2, most notably for acoustically identical stimuli. These results suggest that global speech rate variation is an important factor in this equation, such that when global temporal structure is more variable, rhythmic grouping effects disappear. When this is the case, it appears that listeners revert to being influenced by rate, though Experiment 3 could not directly address the extent to which the observed rate effect was shaped by proximal or distal rate contexts. Experiment 4 suggested that the effects of rate in Experiment 3 were primarily shaped by proximal context, and moreover, that even with proximal context controlled, listeners shift categorization, when this proximal context is deviant from preceding patterns (both in terms of rhythm and rate).

In summary then, these results paint a complex picture: rhythmic grouping effects are robust when global rate is stable, and in that case they take precedence over proximal speech rate. However, when rate is variable within a trial (as in the fast condition in Experiment 3), and more strikingly across trials, they disappear.

These findings speak to previous results in suggesting that an additive interaction of speech rate and rhythmic grouping, as seen in Morrill et al. (2014), does not always obtain. As discussed above, this difference can likely be attributed to the way in which each of these influences was cued. In Morrill et al. (2014), F0 variation signaled grouping, where in the present case, alternating temporal patterns did. In this sense, speech rate variation was directly in conflict with rhythmic grouping (in terms of proximal context), which was exploited by listeners in the present study when global rate was variable. The same sort of directly competing influence was not present in Morrill et al. (2014). The present experiments thus offer an apparent constraint on additivity: when rhythmic variation is cued by duration, it is not additive with speech rate effects (as was seen in Experiment 3). Further testing of this claim could be carried out in cuing rhythm with F0 in a study with the same design as the present one, or manipulating rate to be blocked in presentation (or even between participants; note Morrill et al. 2014 manipulated *rhythm* as a between-participants factor such that each participant only heard one rhythmic pattern).

More generally, the present studies add further to our understanding how distal and proximal influences jointly shape listeners' perception of durational cues. As discussed in "Introduction", the literature offers a complex set of findings which show that in some cases proximal context is more heavily weighted than distal context (Newman & Sawusch, 1996; Reinisch et al., 2011), where in others distal patterns take precedence (Bosker, 2017), and are clearly robust when proximal context does not compete (Reinisch et al., 2011). The present results have shown that, under the right conditions, distal rhythmic context is prioritized over proximal rate. Nevertheless, these results generally affirm the importance of proximal context in the sense that when rate varies from trial to trial (globally) only proximal effects are robust. This is even the case when proximal context is invariant across conditions (Experiment 4), where proximal material is perceived as relatively long or short in relation to what precedes it. This latter finding underscores how temporal deviation from an established pattern can matter in rate-dependent perception: when the pre-target syllable in Experiment 4 differed from what would be expected based on preceding context, this proximal context influenced categorization instead of distal rhythmic information. This, in line with Reinisch et al. (2011) and Kidd (1989), offers support for the idea that when proximal context differs from an established (distal) pattern, it is focused upon by listeners. One notable outcome from Experiment 4 is that distal context did not matter at all in the face of deviating proximal context, as was also found by Kidd (1989).

Perhaps most importantly, the present study has shown the importance of global temporal structure as a mediating influence for the observed rhythmic effects. As noted previously, Kidd (1989) blocked the presentation of his stimuli by precursor type, such that trial-to-trial variation in rate (or any precursor characteristic) was not present, and in that design found clear effects of rhythmic expectancy. This study has shown that when such variation *is* present, rhythmic effects disappear, an important constraint on their when they can emerge. As noted above, Morrill et al. (2014) manipulated rhythmic grouping between subjects, such that a given participant heard only one rhythmic condition, at various rates. Their rhythmic effect was robust to rate variation (comparing across participants), however it remains unknown if the rhythmic effect would persist if a given participant heard variation in both rhythm and rate. Testing the limits of stimulus variation which participants can tolerate, while still evidencing an effect of rhythmic grouping will help further formulate constraints on this effect as it relates to global stimulus characteristics. In the present study, it seems safe to assume that we would observe a re-emergence of the rhythmic grouping effect in blocked presentation (essentially implementing Experiment 2, also using the same design as Kidd 1989), though future research may benefit from testing this directly.

What do these findings tell us about the mechanisms responsible for each of these influences? We have seen that proximal contrast effects are robust to global rate variation in the present study, while rhythmic grouping effects are not. This is consistent with the long-standing proposal that effects of contrasting local durations can be described by a general auditory mechanism (Diehl & Walsh, 1989), or more generally that adjustments for stimulus rate effects are immune to selective attention or variation in cognitive load (Bosker, 2017; Bosker et al., 2020). On the other hand, rhythmic grouping effects are rather fragile, something we can speculate is only apparent in designs with inter-mingling of temporal rhythmic variation and some other dimension (e.g., rate). Why might this be the case? Though rhythmic effects clearly have a robust and consistent influence in the domain of word segmentation, some earlier findings such as Dilley and McAuley (2008) employ between-participant comparisons when multiple cues are manipulated (e.g., F0 and duration), such that a participant doesn't hear orthogonal variation in said cues. The limits of rhythmic effects remain somewhat an open question, and the conditions in which they do and do not occur merits further exploration. The present results would suggest there are limits related to variability in global temporal structure, as described above. Indeed, it has been understood for some time that tracking temporal characteristics in speech is diminished by irregularity and variation (along the lines of Jones & McAuley 2005) and regularity has been shown to benefit speech processing (Quené & Port, 2005). The findings that rhythmic effects in rate-dependent perception are dependent on a certain amount of (temporal) regularity is generally consistent with this idea, and more generally with oscillator-based accounts of rhythmic expectancies in speech perception (see e.g., discussion in Dilley & McAuley 2008).

Of course, the strength and robustness of rhythmic effects, both in word segmentation and rate-dependent perception is a line of inquiry which requires more research. For example, Dilley and McAuley (2008) clearly show that distal rhythmic contexts (cued by F0) exert a strong influence in segmentation (in comparison to proximal context, and semantic context), which suggests they should rank high in a hierarchical model of segmentation (as in Mattys et al. 2005). Nevertheless, we've seen here that rhythmic grouping effects in rate-dependent perception are fragile, and can be overridden by proximal context. The extent to which the robustness of rhythmic effects can be linked to the task (e.g., segmentation versus segmental perception), the acoustic medium cuing rhythm (F0 versus duration), and global temporal traits to which participants are exposed must be teased apart to allow for a full accounting of the influence of rhythmic context in perception and processing. It is worth remarking here that the observed effect sizes are relatively small. With respect to the rhythmic effect in Experiment 2, this could be partially attributable to a competing influence of proximal context, in similar fashion to Reinisch et al. (2011), who found competing distal rate reduced the size of the effect of proximal context in their study. The fact that each of the present experiments included a competing pattern that predicted a different shift in categorization (except for Experiment 1) might be partly responsible for the relatively small effects seen here. This hypothesis could be tested by manipulating rhythmic context in the absence of contextual durational changes, as with F0 as outlined above, in which case we should expect to see a larger effect of rhythmic context. The present study also limited itself to testing perception of one single contrast, and controlled the precursor to be a single syllable. This offered a high level of temporal control, but came at the expense of more naturalistic stimulus designs which would have allowed us to confirm how the present effect generalizes across different segmental and metrical contexts.[9] Broadening the contrasts tested and the nature of the precursor (including more varied, or real-word precursors as in e.g., Dilley & McAuley 2008) will be another worthwhile empirical extension in this regard.

Another broader implication of these results is that effects of rhythmic timing patterns should be considered in light of a growing body of evidence that prosodic organization shapes perception of segmental contrasts in speech, including in the temporal domain (see e.g., Mitterer et al. 2019; Steffman & Katsuda 2020). If we consider rhythmic patterning to constitute part of a language's prosodic system, we should integrate these findings with what we know more generally about prosodic context in speech perception. Exploring the extent to which rhythmic effects are additive with other influences of prosodic context (e.g., temporal patterns associated with prosodic boundaries), and how listeners process rhythmic context in rate-dependent perception online (comparing to other online prosodic influences, as in Kim et al. 2018; Mitterer et al. 2019) will help us understand how the present findings relate more generally to listeners' processing of prosodic information in speech.

Extending the present results along these lines, and those outlined above will accordingly better our understanding of the constraints on rhythmic influences in rate-dependent perception, and how these effects relate to other contextual influences in language comprehension.

Open practices statement

The data for these experiments and the code used to analyze the data are available on the Open Science Foundation website and can be found at https://osf.io/tuh7g/.

Appendix 1: Combined analysis of Experiment 2 and 3

To allow for more concrete comparison of Experiments 2 and 3, a model was fit to combined data from both experiments, described in this section. Given the idea the global rate regularity is beneficial for rhythmic effects we might expect these influences to change over the course of an experiment, as listeners accumulate more exposure to a global rate, with consistent rate in Experiment 2 potentially strengthening the rhythmic effect over time. Accordingly, testing how listeners' responses shift over the course of an experiment might be insightful, which motivated inclusion of trial number as a variable in the model. The model predicted listeners' responses as a function of the continuum, rhythm condition, and experiment (contrast-coded with Experiment 2 mapped to -0.5, and Experiment 3 mapped to 0.5), as well as trial number (scaled and centered within each experiment).[10] Note that rate was not included as a predictor in the model because only Experiment 3 varied rate. Random effects in the model included all fixed effects and interactions as by-participant random slopes, save for experiment. This model will allow us to more thoroughly compare the differences observed across these two experiments, while additionally testing how they vary over time. If it is the case that more exposure to regularity helps strengthen rhythmic grouping effects in Experiment 2, we should expect a three-way interaction in the model between experiment, rhythm and trial, which would show increasing strength of rhythmic effects over the course of

[9] This is also worth considering in light of the finding that repeating speech can sometimes be perceived as sung, i.e., the speech-to-song illusion (Deutsch, Henthorn, and Lapidis, 2011), which might impact possessive listeners' perception of rhythmic timing patterns.

[10] Though Experiment 3 contained twice as many trials as Experiment 2, scaling and centering trial as a predictor for each experiment individually allows scaled values to occupy the same range, making trial as a variable more comparable across experiments.

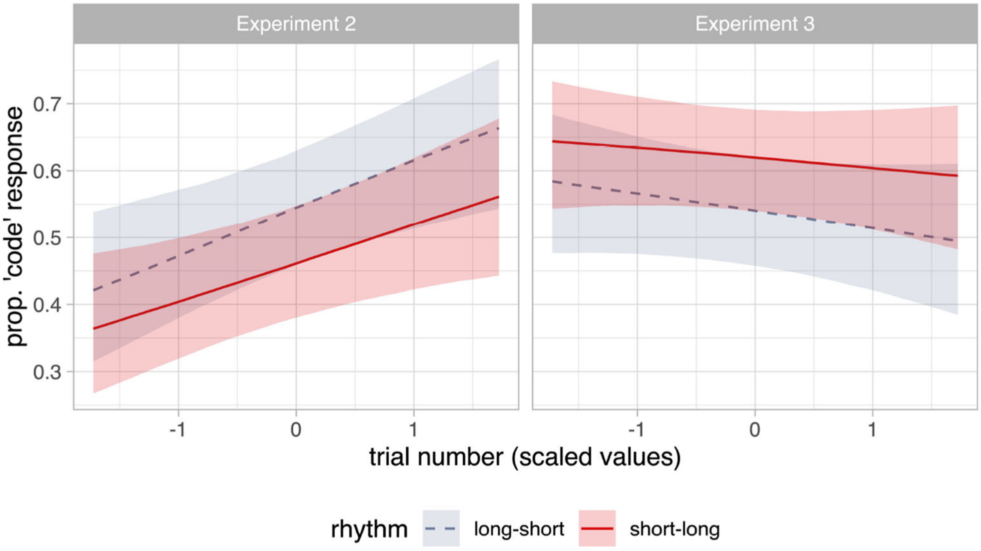

Fig. 7 Model fit for the effect of trial number on overall responses from the combined analysis. Fits are split by rhythm, indicated by line type and color within a panel, and by Experiment, across panels

the trials in Experiment 2 (where global rate is invariant), but not in Experiment 3. We might also expect to see an interaction between experiment and trial if a different global rate in each impacted overall responses.

The combined analysis (model output shown in Table 7 in Appendix 2) finds that, in addition to the expected effect of continuum, only two predictors were credible. The first was the interaction of precursor rhythm and experiment, in line with the reversal of the rhythmic effect observed across experiments. Unsurprisingly, comparing model contrasts using *emmeans* showed that each experiment evidenced a different credible effect of rhythm (Experiment 2: β=0.33, 95% CI = [0.12, 0.56]; Experiment 3: β=-0.32, 95% CI = [-0.50, -0.15]), as was established by the main effects of rhythm in the individual analysis of these experiments (see also Fig. 5).

The other model estimate that was observed to be credible was the interaction between trial and Experiment, suggesting that responses changed over the course of each experiment in a different fashion. Change over the course of each Experiment is shown in Fig. 7, plotting scaled trial number by overall "code", responses, also split by condition. The *emtrends* function of *emmeans* was used to test for the influence of changing trial number in each experiment. This assessment finds a credible effect of trial in Experiment 2, whereby "code" responses increase over the course of the experiment (β=0.26, 95% CI = [0.04, 0.45]). In comparison, no credible change across trials was found in Experiment 3, though the estimated effect is negative unlike that in Experiment 2 (β=-0.08, 95% CI = [-0.27, 0.08]), providing weak evidence for a *decrease* in "code" responses over the course of Experiment 3. The three way interaction between Experiment, rhythm, and trial was not observed to be credible, and indeed post-hoc inspection of the influence of rhythm over the course of each experiment found that the effect did not change reliably in either, though the estimate was positive in Experiment 2, and negative in Experiment 3.[11] We thus do not have clear evidence for an effect of rhythm that changes in a different way across experiments, though we *do* have a clear evidence for a difference in overall responses, such that listeners reliably increased their "code" responses over the course of Experiment 2. What might explain this effect? One possibility is that regular rhythmic patterns influenced listeners' perception of speech rate in the stimuli, given that previous links between rhythmicity and perceived duration have been suggested in the literature.[12] Horr and Di Luca (2015) found that stimuli which presented an isochronous pulse train of tones were perceived to last longer than an interval of the same duration which was anisochronous. If it is the case that listeners' perception of global rate could incorporate this effect, listeners should develop perceived slower global pace with increased exposure to isochronous rhythmic patterning in the stimuli. This in turn would make a given stimulus sound relatively fast, increasing "code" responses over the course of the experiment, and would, by

[11] In Experiment 2: β=0.05, 95% CI = [-0.16, 0.27]; in Experiment 3, β=-0.04, 95% CI = [-0.21, 0.12]. This provides, at best, very weak evidence for an asymmetrical change over trials for the effect of rhythm in each experiment.

[12] In an exploratory analysis, trial number was included in all other models reported in this paper. The only experiment in which trial, or any of its interactions showed a credible effect was Experiment 2, also the only experiment which evidenced the rhythmic grouping effect.

hypothesis, be disrupted by variation in global rate, as in Experiment 3, though this explanation is speculative.

The *lack* of an effect of trial in Experiment 3 can also be compared to previous studies examining global rate effects. For example, it might be expected that the neutral rate condition would be perceived as slower in relation to fast rate (in line with Maslowski et al. 2020, discussed in "Introduction"), and therefore we would see relatively decreased "code" responses in Experiment 3's neutral rate condition as compared to Experiment 2, however this is not the case. Baese-Berk et al. (2014) showed effects like these grow in strength over the course of an experiment as listeners accumulate exposure to global rate patterns, further suggesting we might have expected to see this change occurring over the course of the trials in Experiment 3. A likely explanation for the present lack of an effect is the relatively short duration of Experiment 3, which lasted approximately 20 min. In both Maslowski, Meyer, and Bosker (2019) and Baese-Berk et al. (2014), the experiment lasted longer than 50 min, giving listeners longer exposure to global rate patterns. In fact, Baese-Berk et al. (2014), who report their experiment took approximately 1 h to complete, analyzed the effect of global rate over the course of the experiment in three blocks. In the first block of the experiment (corresponding to about 20 min) there was no observable effect of global rate (see Baese-Berk et al. 2014 Figure 2), which only emerged in the second block and strengthened in the third. This data suggests 20 min of exposure to a global rate pattern may not be enough time to generate previously documented effects, consistent with the lack of an effect seen here in Experiment 3. We can also note that the model estimate for trial in Experiment 3, though it is not credible, is in the direction that we would expect given a faster global rate in the experiment (i.e., decreasing "code" responses over the course of the experiment), suggesting a longer experiment might have allowed for the expected effect to appear.

In summarizing the comparison across experiments, we have reaffirmed that global speech rate variation disrupts the rhythmic grouping effects seen in Experiment 2, as discussed in "Results and discussion". We also have evidence that temporally regular rhythmic patterns induce a change over time, potentially indicating listeners' increasing sensitivity to the pattern.

Appendix 2: Model summaries for all Experiments

Fixed effect estimates, and upper and lower 95% CI are given in each table. A credible fixed effect, for which the CI exclude zero, is bolded. Model estimates are given for (by-participant) random intercepts, and for random slopes.

Table 2 Model results for Experiment 1

Fixed effects	Estimate	Est. Error	L95% CI	U95%CI
Intercept	0.19	0.14	−0.09	0.47
Precursor	**−0.36**	**0.19**	**−0.73**	**−0.004**
Continuum	**1.14**	**0.18**	**0.79**	**1.50**
precursor:cont	−0.17	0.13	−0.43	0.07
Random effects	Estimate	Est. Error		
sd(intercept)	0.72	0.11		
sd(precursor)	0.91	0.17		
sd(continuum)	0.95	0.15		
sd(precursor:cont)	0.31	0.17		

Table 3 Model results for Experiment 2

Fixed effects	Estimate	Est. Error	L95% CI	U95%CI
Intercept	0.02	0.22	−0.41	0.45
Precursor	**0.33**	**0.12**	**0.09**	**0.57**
Continuum	**1.79**	**0.16**	**1.48**	**2.12**
precursor:cont	0.15	0.13	−0.11	0.39
Random effects	Estimate	Est. Error		
sd(intercept)	1.16	0.17		
sd(precursor)	0.35	0.18		
sd(continuum)	0.80	0.14		
sd(precursor:cont)	0.15	0.12		

Table 4 Model results for Experiment 3

Fixed effects	Estimate	Est. Error	L95% CI	U95%CI
Intercept	0.31	0.16	0.0	0.64
Precursor rate	**−0.19**	**0.08**	**−0.34**	**−0.04**
Precursor rhythm	**−0.32**	**0.09**	**−0.50**	**−0.14**
Continuum	**2.24**	**0.20**	**1.86**	**2.64**
Precursor rate:cont	0.11	0.10	−0.09	0.31
Precursor rhythm:cont	0.04	0.11	−0.16	0.27
Precursor rate:precursor rhythm	**−0.38**	**0.19**	**−0.77**	**−0.03**
Precursor rate:precursor rhythm:cont	0.36	0.21	−0.04	0.77
Random effects	Estimate	Est. Error		
sd(intercept)	0.86	0.13		
sd(precursor rate)	0.11	0.08		
sd(precursor rhythm)	0.26	0.11		
sd(continuum)	1.08	0.16		
sd(precursor rate:cont)	0.16	0.10		
sd(precursor rhythm:cont)	0.18	0.12		
sd(precursor rate :precursor rhythm)	0.59	0.24		
sd(precursor rate :precursor rhythm:cont)	0.29	0.22		

Table 5 Pairwise comparison of contrasts for all rhythm and rate combinations in Experiment 3

Conditions compared	estimate	L95% CI	U95% CI
Long-short fast vs. short-long fast	−0.13	−0.37	0.12
Long-short + neutral vs. long-short + fast	**−0.38**	**−0.63**	**−0.15**
Long-short + fast vs. short-long + neutral	−0.12	−0.37	0.09
Long-short + neutral vs. short-long + fast	**−0.50**	**−0.73**	**−0.26**
Short-long + neutral vs. short-long + fast	0	−0.23	0.25
Long-short + neutral vs. short-long + neutral	**−0.50**	**−0.77**	**−0.24**

Table 6 Model results for Experiment 4

Fixed effects	Estimate	Est. Error	L95% CI	U95%CI
Intercept	0.10	0.12	−0.13	0.34
Precursor rate	−0.06	0.08	−0.31	0.19
Precursor rhythm	**−0.15**	**0.08**	**−0.31**	**−0.01**
Continuum	**1.71**	**0.18**	**1.35**	**2.08**
Precursor rate:cont	0.15	0.10	−0.06	0.34
Precursor rhythm:cont	0.05	0.10	−0.14	0.26
Precursor rate:precursor rhythm	−0.08	0.14	−0.36	0.19
Precursor rate:precursor rhythm:cont	0.36	0.21	−0.04	0.77
Random effects	Estimate	Est. Error		
sd(intercept)	0.86	0.13		
sd(precursor rate)	0.11	0.08		
sd(precursor rhythm)	0.26	0.11		
sd(continuum)	1.08	0.16		
sd(precursor rate:cont)	0.16	0.10		
sd(precursor rhythm:cont)	0.18	0.12		
sd(precursor rate :precursor rhythm)	0.59	0.24		
sd(precursor rate :precursor rhythm:cont)	0.29	0.22		

Table 7 Model results for the combined analysis of Experiment 2 and Experiment 3

Fixed effects	Estimate	Est. Error	L95% CI	U95%CI
Intercept	0.17	0.14	−0.10	0.45
Precursor rhythm	−0.00	0.07	−0.15	0.15
Continuum	**2.13**	**0.13**	**1.87**	**2.40**
Experiment	0.31	0.28	−0.24	0.86
Trial	0.09	0.07	−0.05	0.23
Precursor rhythm:cont	−0.11	0.08	−0.27	0.05
Precursor rhythm:exp	**0.66**	**0.14**	**0.38**	**0.93**
Cont:exp	0.48	0.27	−0.06	1.00
Precursor rhythm:trial	−0.01	0.07	−0.14	0.13
Cont:trial	0.22	0.06	0.09	0.34
Exp:trial	**−0.34**	**0.14**	**−0.62**	**−0.07**
Precursor rhythm:cont:exp	0.14	0.15	−0.17	0.44
Precursor rhythm:cont:trial	−0.02	0.09	−0.20	0.17
Precursor rhythm:exp:trial	0.09	0.14	−0.17	0.36
Cont:exp:trial	0.06	0.12	−0.18	0.30
Precursor rhythm:cont:exp:trial	0.06	0.17	−0.28	0.39
Random effects	Estimate	Est. Error		
sd(intercept)	1.08	0.11		
sd(precursor rhythm)	0.27	0.10		
sd(continuum)	0.99	0.11		
sd(trial)	0.49	0.06		
sd(precursor rhythm:cont)	0.13	0.09		
sd(precursor rhythm:trial)	0.14	0.10		
sd(cont:trial)	0.33	0.06		
sd(precursor rhythm:cont:trial)	0.28	0.14		

Acknowledgements Many thanks are due to Adam Royer for recording speech materials for the stimuli, and to Yang Wang, Danielle Bagnas, and Qinxia Guo for help with data collection. Additional thanks are due to four anonymous reviewers, and to members of the UCLA Phonetics lab, for insightful feedback and discussion.

Declarations

Conflict of Interests The author declares that they have no conflict of interest.

References

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.

Barnes, R., & Jones, M. R. (2000). Expectancy, attention, and time. *Cognitive Psychology*, *41*(3), 254–311.

Barry, W., Andreeva, B., & Koreman, J. (2009). Do rhythm measures reflect perceived rhythm?. *Phonetica*, *66*(1-2), 78–94.

Boersma, P., & Weenink, D. (2020). Praat: doing phonetics by computer (version 6.1.09). http://www.praat.org.

Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. *Attention, Perception, & Psychophysics*, *79*(1), 333–343.

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.

Bosker, H. R., Sjerps, M. J., & Reinisch, E. (2020). Temporal contrast effects in human speech perception are immune to selective attention. *Scientific Reports*, *10*(1), 1–11.

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2015). Metrical expectations from preceding prosody influence perception of lexical stress. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 306–323.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, *22*(3), 129–159.

Crochiere, R. (1980). A weighted overlap-add method of short-time Fourier analysis/synthesis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(1), 99–102.

Crowhurst, M. J., & Olivares, A. T. (2014). Beyond the iambic-trochaic law: The joint influence of duration and intensity on the perception of rhythmic speech. *Phonology*, *31*(1), 51–94.

Cutler, A., & Darwin, C. J. (1981). Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. *Perception & Psychophysics*, *29*(3), 217–224.

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *The Journal of the Acoustical Society of America*, *129*(4), 2245–2252.

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.

Dilley, L. C., Mattys, S. L., & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language*, *63*(3), 274–294.

Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, *59*(3), 294–311.

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

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85*, 761–768.

Handel, S. (1993). *Listening: An introduction to the perception of auditory events*. Cambridge: The MIT Press.

Hawkins, S., & Smith, R. (2001). Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistics*, *13*, 99–188.

Hay, J. S. F., & Diehl, R. L. (2007). Perception of rhythmic grouping: Testing the iambic/trochaic law. *Perception & Psychophysics*, *69*(1), 113–122.

Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. Chicago: University of Chicago Press.

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.

Hoequist, C. E., & Kohler, K. J. (1986). Further experiments on speech rate perception with logatomes. *Arbeitsberichte des Instituts fur Phonetik der Universitit Kiel*, *22*, 29–136.

Horr, N. K., & Di Luca, M. (2015). Taking a long look at isochrony: Perceived duration increases with temporal, but not stimulus regularity. *Attention, Perception, & Psychophysics*, *77*(2), 592–602.

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review*, *83*(5), 323–355.

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

Jun, S.-A. (2012). Prosodic typology revisited: Adding macro-rhythm. In *Proceedings of speech prosody*, Vol. 6.

Jungers, M. K., Palmer, C., & Speer, S. R. (2002). Time after time: The coordinating influence of tempo in music and speech. *Cognitive Processing*, *1*(2), 21–35.

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

Kim, S., Mitterer, H., & Cho, T. (2018). A time course of prosodic modulation in phonological inferencing: The case of Korean post-obstruent tensing. *Plos one*, *13*(8), e0202912.

Kösem, A., Bosker, H. R., Jensen, O., Hagoort, P., & Riecke, L. (2020). Biasing the perception of spoken words with transcranial alternating current stimulation. *Journal of Cognitive Neuroscience*, *32*(8), 1428–1437.

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.

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review*, *106*(1), 119–159.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, *5*(3), 253–263.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means. https://CRAN.R-project.org/package=emmeans.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*(6), 1001–1010.

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2020). Eye-tracking the time course of distal and global speech rate effects. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(10), 1148–1163.

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.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology: General*, *134*(4), 477–500.

McAuley, J. D., & Jones, M. R. (2003). Modeling effects of rhythmic context on perceived duration: A comparison of interval and entrainment approaches to short-interval timing. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(6), 1102–1125.

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.

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.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, *46*(6), 505–512.

Mitterer, H., Kim, S., & Cho, T. (2019). The glottal stop between segmental and suprasegmental processing: The case of Maltese. *Journal of Memory and Language*, *108*, 104034.

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.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5-6), 453–467.

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

Oppenheim, A. V., & Schafer, R. W. (1975). *Digital signal processing*. Upper Saddle River: Prentice-Hall.

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, *3*.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, 539–558.

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.

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.

Quené, H. (2013). Longitudinal trends in speech tempo: The case of Queen Beatrix. *The Journal of the Acoustical Society of America*, *133*(6), EL452–EL457.

Quené, H., & Port, R. F. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica*, *62*(1), 1–13.

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of the Acoustical Society of America*, *51*(4B), 1296–1303.

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.

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.

Steffman, J. (2019). Intonational structure mediates speech rate normalization in the perception of segmental categories. *Journal of Phonetics*, *74*, 114–129.

Steffman, J., & Jun, S.-A. (2019). Perceptual integration of pitch and duration: Prosodic and psychoacoustic influences in speech perception. *The Journal of the Acoustical Society of America*, *146*(3), EL251–EL257.

Steffman, J., & Katsuda, H. (2020). Intonational structure influences perception of contrastive vowel length: The case of phrase-final lengthening in Tokyo Japanese. *Language and Speech*, 0023830920971842.

Stilp, C. (2018). Short-term, not long-term, average spectra of preceding sentences bias consonant categorization. *The Journal of the Acoustical Society of America*, *144*(3), 1797–1797.

Stilp, C. (2020). Acoustic context effects in speech perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(1), e1517.

Tehrani, H. (2020). Appsobabble: Online applications platform. https://www.appsobabble.com.

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.

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161.

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

Warren, R. M. (1985). Criterion shift rule and perceptual homeostasis. *Psychological Review*, *92*(4), 574–584.

Woodrow, H. (1909). *A quantitative study of rhythm: The effect of variations in intensity, rate and duration*. San Francisco: Science Press.

Woodrow, H. (1911). The role of pitch in rhythm. *Psychological Review*, *18*(1), 54–77.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.