Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-021-02298-x

ORIGINAL MANUSCRIPT

Effects of speech-rhythm disruption on selective listening with a single background talker

J. Devin McAuley[1] · Yi Shen[2] · Toni Smith[1] · Gary R. Kidd[3]

Accepted: 5 March 2021

Abstract

Recent work by McAuley et al. (*Attention, Perception, & Psychophysics, 82,* 3222–3233, 2020) using the Coordinate Response Measure (CRM) paradigm with a multitalker background revealed that altering the natural rhythm of target speech amidst background speech worsens target recognition (a target-rhythm effect), while altering background speech rhythm improves target recognition (a background-rhythm effect). Here, we used a single-talker background to examine the role of specific properties of target and background sound patterns on selective listening without the complexity of multiple background stimuli. Experiment 1 manipulated the sex of the background talker, presented with a male target talker, to assess target and background-rhythm effects with and without a strong pitch cue to aid perceptual segregation. Experiment 2 used a vocoded single-talker background to examine target and background-rhythm effects with envelope-based speech rhythms preserved, but without semantic content or temporal fine structure. While a target-rhythm effect was present with all backgrounds, the background-rhythm effect was only observed for the same-sex background condition. Results provide additional support for a selective entrainment hypothesis, while also showing that the background-rhythm effect is not driven by envelope-based speech rhythm alone, and may be reduced or eliminated when pitch or other acoustic differences provide a strong basis for selective listening.

Keywords Speech Perception · Selective Attention

Introduction

Understanding speech in noisy listening environments, such as in a crowded café, by a busy street, or in a large Zoom meeting, is a difficult perceptual problem that most of the hearing population faces regularly. Moreover, speech-in-noise (SIN) abilities vary greatly from person to person, and these individual differences are not fully explained by pure-tone hearing thresholds, cognitive ability, or age (Akroyd, 2008; Houtgast & Festen, 2008; Humes, Kidd, & Lentz, 2013b). One factor that has been considered to explain some individual differences in SIN perception is temporal processing ability. Much of this research has focused on temporal resolution—that is, the ability to detect brief or rapid temporal events (e.g., gap detection or amplitude-modulation detection). In this regard, despite the temporal characteristics of speech being important for speech perception (e.g., Darwin, 1975; Golumbic et al., 2012; Rosen, 1992), traditional psychoacoustic measures of temporal resolution have generally been found to be poor predictors of SIN ability (e.g., Humes & Dubno, 2010; Humes et al., 2013a, b; Kidd et al., 2007).

More recently, however, several studies have shown that a sensitivity to suprasegmental temporal patterns, or speech *rhythm,* may be more important for speech perception in noise than temporal resolving power (e.g., Aubanel et al., 2016; McAuley et al., 2020; Riecke et al., 2018; Wang et al., 2018). By *speech rhythm*, we mean the temporal patterning of speech sounds that leads to the *perception* of regularity and guides temporal expectations about when subsequent sounds in a speech stream are likely to occur. Importantly, fluctuations in the amplitude envelope of speech patterns give rise to a sense of regularity that can be used to guide expectations for the timing of future speech events. Speech timing at the syllabic level (within the range of 3–9 Hz) in particular has been shown to contribute to the rhythmic qualities of speech across

✉ J. Devin McAuley
dmcauley@msu.edu

[1] Department of Psychology, Michigan State University, East Lansing, MI 48824, USA

[2] Department of Speech and Hearing Sciences, University of Washington, Seattle, WA, USA

[3] Department of Speech, Language and Hearing Sciences, Indiana University, Bloomington, IN, USA

Published online: 29 March 2021

many languages (Dauer, 1983; Tilsen & Arvaniti, 2013) and is also likely to be important for early language acquisition (Goswami, 2019).

One theoretical framework that has been useful for exploring the role of rhythm in speech perception is dynamic attending theory (DAT), which posits the existence of attentional rhythms that are entrained by environmental stimuli (Jones, 1976; Jones & Boltz, 1989; Large & Jones, 1999; McAuley et al., 2006; Miller et al., 2013). According to DAT, periodic fluctuations in attention gradually adapt in phase and period in response to exogenous periodic (or quasiperiodic) stimulus rhythms, such that peaks in attention align with points in time where stimulus events are expected to occur. This alignment of maximal attentional energy to time points where relevant stimulus events are likely to be present is hypothesized to facilitate perception of those events (Jones, 1976; Jones & Boltz, 1989; Large & Jones, 1999; McAuley et al., 2006). Behavioral evidence from a variety of perceptual tasks has provided support for DAT (Barnes & Jones, 2000; Jones et al., 2002; McAuley & Jones, 2003; Miller et al., 2013).

Recent work has also provided evidence for dynamic attending in the domain of speech perception. For instance, the perception of ambiguities in syllable organization is influenced by earlier rhythmic context, suggesting that people are sensitive to speech rhythms and that these rhythms can set up temporal expectations that influence word segmentation (Baese-Berk et al., 2019; Dilley & McAuley, 2008; Morrill et al., 2014). Moreover, speech understanding in noise is adversely affected by a disruption of rhythmic regularities in spoken sentences. Isochronously retimed speech is more intelligible amid noise than anisochronously retimed speech (Aubanel et al., 2016). In multitalker babble, words occurring later in the target sentence are better recognized than are words occurring earlier in the same sentence, but not when the target is made artificially irregular (Wang et al., 2018), suggesting that temporal (rhythmic) expectations build up over time.

Successful recognition of target speech in a multitalker environment requires selective attention to the target, which may be facilitated by entrainment to the target rhythm. This is because rhythmic expectations, based on entrainment to familiar speech rhythms, help to predict the timing of upcoming information-carrying events in the target, which facilitates the alignment of attentional focus on those events. Conversely, when rhythmic expectations are violated by disrupted or unpredictable speech rhythms, there is a misalignment of attentional focus, which results in poorer speech recognition. However, disrupting background rhythm can facilitate target speech recognition by reducing the likelihood of entrainment to competing speech, thus reducing the chances for intrusion errors due to incidental entrainment to a competing speech stream. We will refer to this hypothesis about the effect of speech rhythm on dynamic attending and selective listening as the *selective entrainment* hypothesis. It is worth noting that selective entrainment generally depends on some type of detectable difference between target and background stimuli (e.g., spatial location, talker gender, talker identity, semantic context) to provide potentially discriminable carriers for the target and background rhythms.

The DAT-based selective entrainment hypothesis is consistent with recent findings from neurophysiological investigations. It is known that cortical neural oscillations that operate near the syllabic rate can exhibit phase-locked synchrony to the temporal envelope of speech, and it has been argued that this neural entrainment to the speech envelope is used as a mechanism for parsing connected speech into smaller units (Ding et al., 2016; Ghitza, 2011; Giraud & Poeppel, 2012; Riecke et al., 2018). Disruptions to synchronized neural activity to the ongoing speech envelope via brain stimulation modulates speech comprehension (Riecke et al., 2018). In situations of selective listening to target speech in the presence of competing speech, neural entrainment to the target envelope is enhanced when the listener selectively attends to the target (Ding & Simon, 2012, 2014; Golumbic et al., 2013).

As an alternative to the selective entrainment hypothesis, rhythmic differences between target and background speech may result in perceptual segregation without the involvement of attentional entrainment. That is, the rhythmic differences may serve as a "cue" for perceptual segregation of competing speech streams, much like stimulus properties such as fundamental frequency or spatial location. Past research has identified an array of acoustic cues that lead to obligatory segregation of target and background sound sources (see Bregman, 1990; Carlyon, 2004). For example, when two spoken utterances are presented simultaneously, increasing the fundamental frequency difference (ΔF0) between the voices of the target and background talkers increases intelligibility of the target (Assmann & Summerfield, 1989, 1990; Brokx & Nooteboom, 1982). Because male and female voices typically have fairly large ΔF0s (Lavan et al., 2019; Poon & Ng, 2015; Whiteside, 1998), the ΔF0 can act as a strong cue for the segregation of male and female voices in a multitalker context, and contribute to a release from speech-on-speech masking when the target and background talkers are of different sex (Brungart, 2001).

Differences in rhythmic regularities between the target and background speech may play a similar role to ΔF0. In this *disparity-based segregation* hypothesis, a substantial difference in rhythmic regularities between the target and background stimuli would facilitate segregation. This type of rhythm-based segregation has been demonstrated with pure-tone sequences, in which consecutive tones alternate between frequencies that tend to form two perceptual streams when the frequency difference is large and the intertone interval is small. Studies have shown that different rhythms within each stream contribute to stream segregation when adjacent frequency separation and intertone intervals are not sufficient for reliable segregation (see Bendixen, 2014; Jones et al.,

1981). However, it is unclear whether a difference in the temporal structure of two competing auditory patterns can facilitate selective listening in the absence of predictable rhythmic structure that affords entrainment.

The selective entrainment and disparity-based segregation hypotheses predict different outcomes for speech-on-speech masking when the rhythmic regularity of the *target* speech is disrupted, while that of the *background* speech is retained. The selective entrainment hypothesis emphasizes the rhythmic regularity of the target speech and its potential to lead to attentional entrainment. According to the selective entrainment hypothesis, decreasing the temporal regularity of the target rhythm reduces attentional entrainment to the target utterance, thus making it more difficult to form rhythmic expectations about the timing of to-be-reported target words, which results in poorer target speech-recognition performance. On the other hand, the disparity-based segregation hypothesis emphasizes the similarity between the target and background speech rhythms. According to the disparity-based segregation hypothesis, decreasing the regularity of the target rhythm enlarges the difference between the target and background speech rhythms, thus improving perceptual segregation of the target and background speech, which results in better target speech-recognition performance.

As an initial test of the selective entrainment and disparity-based segregation hypotheses, McAuley et al. (2020) measured speech recognition in a multitalker background and systematically altered the natural rhythms of either the target sentence or the background sentences. Consistent with the selective entrainment hypothesis, but not the disparity-based segregation hypothesis, increasing the level of rhythm alteration applied to the *target* speech made the recognition of keywords in the target sentence much more difficult. This was true even though the same alterations of the target rhythm had no effect on intelligibility of the target sentences presented in isolation without background sounds, supporting the conclusion that the effect of rhythm alteration was not due simply to reducing the intelligibility of the target words. In this work, it is possible, however, that disparity-based segregation was also involved, and the observed effect of target rhythm alteration was the net effect of the two processes.

The first goal of the present study focuses on the effect of target rhythm alteration (referred to here as the "target-rhythm effect") in order to investigate the potential involvement of disparity-based segregation and to further examine the robustness of the target-rhythm effect. To this end, Experiment 1 presents a target sentence with a single background sentence produced by a talker of either the same or different sex from the target talker. In addition to simplifying potential interactions between target and background rhythms, the use of a single-talker background will help to establish the generality of the rhythm-based effects initially observed in a multitalker context (McAuley et al., 2020).

Without changes to the background rhythm, selective entrainment predicts poorer target speech recognition as the level of target rhythm alteration increases with both same-sex and different-sex background talkers. On the other hand, if disparity-based segregation is involved, it will counter the negative effect of target-rhythm alteration caused by disruptions to selective entrainment. Furthermore, any disparity-based improvement should be smaller for the different-sex background talker than for the same-sex talker condition. This is because, in the presence of a large ΔF0 (in the different-sex talker condition), the contribution of rhythmic segregation to improvements in target recognition should be negligible (e.g., George & Bregman, 1989). Thus, without the benefits of rhythm-based segregation to counteract the decrease in entrainment, a larger target-rhythm effect would be expected for the different-sex than the same-sex talker condition.

A second major finding from McAuley et al. (2020) was that increasing the level of rhythm alteration applied to the *background* speech *improved* recognition of keywords in the target sentence (referred to here as the "background-rhythm effect"). That the same difference between target and background speech rhythm (intact vs. altered rhythm) leads to better performance when the background is altered, but worse performance when the target is altered, clearly shows that the difference between target and background rhythms is not the crucial factor. Rather, it appears that entrainment is the key factor: entrainment to the target speech is more difficult when natural speech rhythms in the target are disrupted, but entrainment to the target is enhanced when competing speech is not conducive to entrainment due to rhythmic alteration.

The second goal of the current study is to start to unpack the effect of background-rhythm alteration by examining the influence of interactions between the temporal envelopes of the target and background speech in the absence of semantic content in the background stimulus. Experiment 2 uses the same-sex single-talker background speech used in Experiment 1, but the background speech is tone vocoded. The rationale for using a tone-vocoded background is that the vocoding process maintains the broadband temporal envelopes of original background sentences but removes the temporal fine structure and renders the sentences unintelligible. It is known that the perception of the temporal envelope of one sound could be undermined by the presence of an amplitude-modulation imposed on the sound (i.e., modulation masking; see Bacon & Grantham, 1989; Houtgast, 1989) or by the temporal envelope of another simultaneous stimulus (i.e., modulation detection/discrimination interference; see Yost et al., 1989). Such low-level modulation masking/interference has been shown to play important roles for speech understanding in complex environments (e.g., Fogerty et al., 2016; Stone et al., 2012), but it is not clear whether the background-rhythm effect observed by McAuley et al.

(2020) can be fully explained by low-level envelope processing. Use of the vocoded background stimuli will provide a test of whether the envelope-based speech rhythm is sufficient to produce an improvement in the recognition of target speech when the background rhythm is altered. The lack of a background-rhythm effect under these conditions would indicate that a more speech-like stimulus, or perhaps semantic information, is necessary to obtain the background-rhythm effect. This would suggest that without speech fine structure or semantic information, a competing speech rhythm does not interfere with selective listening to a target sentence, either because of the lack of semantic interference or because of the acoustic disparity between vocoded speech and natural speech.

General methods

All stimuli consisted of two simultaneously presented spoken sentences amid a background of speech-shaped noise. Speech stimuli came from the CRM corpus (Bolia et al., 2000). Each sentence had the same structure: "Ready [call sign] go to [Color] [Number] now." In every sentence, one of eight call signs (e.g., "Baron," "Charlie," Eagle"), one of four Colors ("red," "green," "blue," or "white") and one of seven Numbers (1–8, excluding 7 from the target so that each possible number was monosyllabic) appeared. The target sentence always came from the same male talker (Talker #1) with a mean fundamental frequency of about 118.3 Hz (Allen et al., 2008), and contained the call sign "Baron." The call signs, Colors, and Numbers in the background were always different from those of the target. The background sentences always came from a different talker than the target. In Experiment 1 of the current study, the background talker was either a male (Talker #0) with a mean F0 of about 100.19Hz (Allen et al., 2008) or a female (Talker #4) with a mean F0 of about 211.06Hz (Allen et al., 2008) in two separate conditions. In Experiment 2, the background talker was always the male talker (Talker #0); the harmonic structure in the vocoded background sentence allowed setting the average F0 of the background sentence (100 Hz) close to that of the male background talker in Experiment 1, to provide similar F0-based segregation cues. During each experimental trial, participants were instructed to listen for a target sentence with the call sign "Baron" and report the Color and Number they heard in the target sentence by clicking on a square with that combination of Color and Number, presented via a custom MATLAB program.

In some conditions, the natural rhythm of either the background or the target utterances were disrupted by temporally expanding and contracting the speech in a sinusoidal fashion (Fig. 1). Alterations to the original CRM sentences were made using Praat's Pitch Synchronous Overlap and Add (PSOLA) algorithm, according to a compression ratio (CR) given by $CR(t) = 1 + m \sin(2\pi f_m t + \phi)$, where CR > 1 indicates a slowing of speech and CR < 1 indicates compressed, or faster speech. The rhythm alteration rate, f_m, was set to 1 Hz. based on McAuley et al. (2020), who showed that this value preserved speech intelligibility while still providing a strong percept of timing variation. Additionally, the low rhythm alteration rate of 1 Hz ensured that during each 1-second period within the speech stimuli the duration of the period was maintained, with temporal expansion applied to half of the period and temporal compression applied to the remaining half (see Fig. 1b–c). Consequently, the rhythm alteration manipulation did not change the syllabic rate (i.e., the number of syllables per second) of the speech stimuli. This also means that the effect of the rhythm alterations on the modulation spectrum was quite limited. The rhythm alteration does not affect amplitude modulation (other than the expansion and contraction

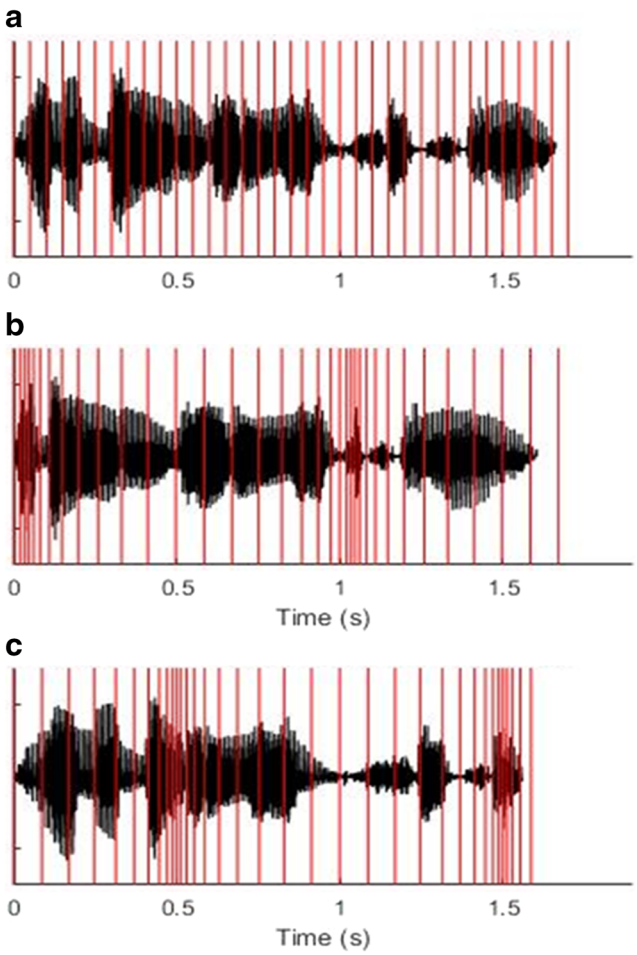

Fig. 1 Examples of rhythm unaltered and altered versions of a spoken CRM sentence of the form "Ready [call sign] go to [color] [number] now.' The top panel (**a**) shows the sample sentence where the rhythm is unaltered ($m = 0$), as represented by the bars equally spaced in time. The middle and bottom panels show how the same time points in the speech signal are shifted by the rhythm transformation ($m = 0.75$, maximally altered condition) for two different phases (**b**, $\phi = 5\pi/4$; **c**, $\phi = \pi/2$)

of the rate of modulation) and the slow (1 Hz) rate and modest amount of expansion/compression is unlikely to have affected perception of the segments (based on prior work with time-compressed or expanded speech: e.g., Gordon-Salant et al., 2007). The degree of rhythm alteration is determined by the modulation depth, m, which took on values of either 0.0, 0.25, 0.50, or 0.75 depending on condition. The initial phase of rhythm alteration, ϕ, was randomly assigned for each trial within a block from a set of equally probable values (0, $\pi/4$, $2\pi/4$, $3\pi/4$, $4\pi/4$, $5\pi/4$, $6\pi/4$, and $7\pi/4$). Onset asynchronies of either +50 ms or −50 ms, with equal probability, were introduced to the background sentences relative to the target before the rhythm alteration took place in order to make certain that rhythm alteration effects are not simply due to misalignment between the target and background Color and Number.

In all conditions, the presentation levels of both the target and background sentences were set to 65 dB sound pressure level (SPL). Recognizing target speech with only one equal-level talker in the background was expected to be relatively easy (Rosen et al., 2013). In order to control for the overall difficulty of the task and avoid ceiling effects, the target and background sentences were presented in an additional speech-shaped noise. Stimuli were presented diotically using Sennheiser HD 280 Pro over-the-ear headphones at a sampling rate of 22050 Hz.

Experiment 1

Methods

Participants and design Thirty-six participants (eight males, 28 females) were recruited from the Michigan State University Department of Psychology participant pool and received course credit as compensation for participating. All were native speakers of American English and were screened for normal hearing (pure tone average, or PTA <20 dB HL, in both ears). The experiment had a 2 (background-talker sex: male or female) × 2 (type of rhythm alteration: target or background) × 4 (level of rhythm alteration: m = 0, 0.25, 0.50, 0.75) mixed-factorial design. Background talker sex and type of rhythm alteration were between-subjects factors, leading to four participant groups: same-sex background/background rhythm alteration (n = 9), different-sex background/background rhythm alteration (n = 9), same-sex background/target rhythm alteration (n = 10), different-sex background/target rhythm alteration (n = 8). The level of rhythm alteration was manipulated within subjects. Additional speech-shaped noise was added with an SNR (target relative to speech-shaped noise) of −6 dB for the different-sex background, compared with 0 dB for the same-sex background. The 6-dB difference in SNR was used to yield roughly equivalent performance for the same-sex and different-sex backgrounds in the unaltered rhythm condition (m = 0), based on pilot testing.

Procedure

The experiment was conducted in a single test session of 16 experimental blocks. Each block consisted of 40 trials with the same level of rhythm alteration. Each of the four levels of rhythm alteration occurred four times total, once within each set of four blocks; the order of rhythm alteration levels was counterbalanced across sets. Additionally, the entire sequence of 16 blocks was presented in one of two orders; one order was the reverse of the other, with order counterbalanced across subjects. A mandatory break was provided after eight blocks, and participants were encouraged to take breaks as needed between blocks. Afterward, participants completed surveys about their personal and musical background and about the strategies they used during the experiment. The entire session lasted approximately 1.5 hours.

Results

Figure 2 shows mean proportion of correct responses (reporting both the correct target Color and Number) for the same-sex and different-sex background conditions for alterations of the target rhythm (Fig. 2a) and alterations of the background rhythm (Fig. 2b). Results revealed a clear target-rhythm effect for both the male (same sex) background talker and the female (different sex) background talker that was consistent across listeners. Linear trend analyses showed that increasing levels of target rhythm alteration reliably reduced target recognition with both same sex, $F(1, 9) = 131.97$, $p < .001$, $\eta^2 = 0.94$, and different sex background talkers, $F(1, 7) = 190.49$, $p < .001$, $\eta^2 = 0.97$. There was also a robust background-rhythm effect for the same-sex background condition, but not for the different-sex background condition. Increasing levels of background rhythm alteration dramatically improved target recognition for the same-sex background, $F(1, 8) = 63.58$, $p < .001$, $\eta^2 = 0.89$, but to a much lesser degree for the different-sex background, $F(1, 8) = 5.91$, $p = .041$, $\eta^2 = 0.43$. For the observed target-rhythm effect, the slope was slightly less negative for the same-sex background ($b = -0.17$) than for the different-sex background ($b = -0.24$). Conversely, for the background-rhythm effect, the slope was close to zero for the different-sex background ($b = 0.04$), but very positive for the same-sex background ($b = 0.27$).

Next, we considered how types of errors were modulated by alteration of the target and background rhythms. Of particular interest were errors based on listeners' reporting of target words from the background sentence instead of words in the target sentence. These background intrusions provide a measure of inappropriate (for this task) attention to the background

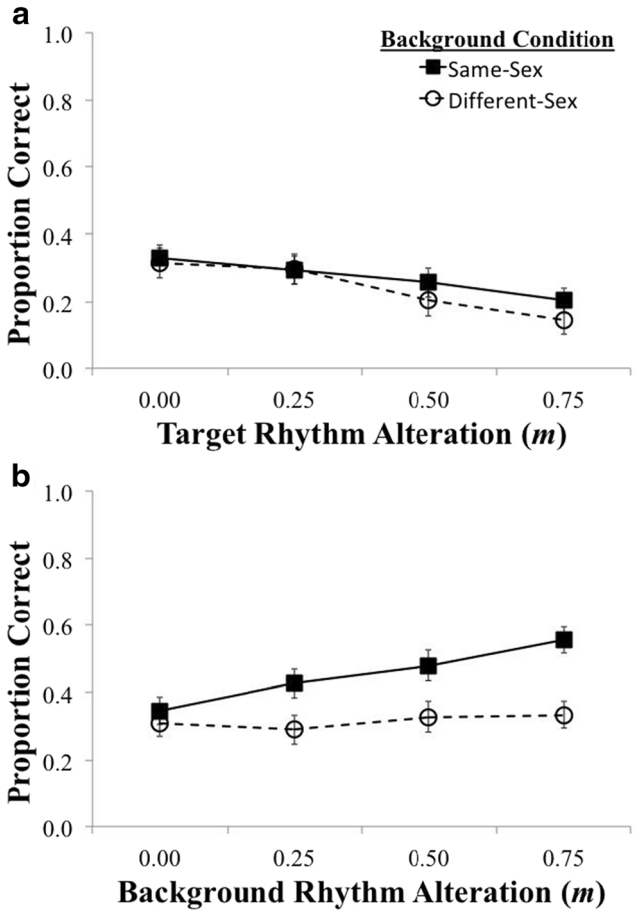

Fig. 2 Experiment 1: Proportion correct target Color and Number recognition for each level of rhythm alteration (m = 0.0, 0.50, 0.25, 0.75). Dotted lines (with open circles) represent the different-sex background talker condition and solid lines (with squares) represent the same-sex background talker condition. **a** shows proportion correct when the target rhythm was altered (same-sex: n = 10, different-sex: n = 8), while **b** shows proportion correct when the background rhythm was altered (same-sex: n = 9, different-sex: n = 9). Error bars represent standard error of the mean

sentences. This allows us to determine whether errors are due to misdirected attention, or to a more general distraction effect (in which the background sentences are acting merely as a masking noise) that would tend to cause random errors (based on pure guessing). Thus, an analysis of intrusions errors can provide a clearer picture of listeners' attentional focus under various conditions, which helps to evaluate hypotheses about attentional entrainment.

Intrusion results are shown in Fig. 3. Overall, the same-sex condition tended to produce more Color intrusions (M = 0.42, SD = 0.14) than the different-sex condition (M = 0.32, SD = 0.064), $t(25.47) = 2.86$, $p = .008$, 95% CI [0.029, 0.18]. The same-sex condition also tended to produce more Number intrusions (M = 0.39, SD = 0.20) than the different-sex condition (M = 0.22, SD = 0.058), $t(21.34) = 3.74$, $p = .001$, 95% CI [0.079, 0.28]. Levene's test indicated unequal variances for both Color ($F = 7.38$, $p = .01$) and Number ($F = 14.27$, $p = .001$) intrusions, so degrees of freedom for both t tests have been adjusted. This difference indicates that errors in the different-sex condition tended to include more random errors (unrelated to the background Color and Number) than in the same-sex condition.

Moreover, on top of the overall differences in the proportion of intrusions in the same-sex and different-sex background conditions, alteration of the target rhythm produced similar linear trends in the positive direction (more intrusions) for both background talkers (see Fig. 3a–b). For the same-sex background condition, greater alteration of the target rhythm led to more intrusions for both Color, $F(1, 9) = 5.80$, $p = .038$, $\eta^2 = 0.39$, and Number, $F(1, 9) = 8.67$, $p = .016$, $\eta^2 = 0.49$. The same was true for the different-sex background condition, Color, $F(1, 7) = 14.47$, $p = .007$, $\eta^2 = 0.67$; Number, $F(1, 7) = 7.21$, $p = .031$, $\eta^2 = 0.51$.

Conversely (as shown in Fig. 3c–d), increasing alteration of the background rhythm led to fewer Color and Number intrusions for both the same-sex background, Color intrusions, $F(1, 8) = 79.11$, $p < 0.001$, $\eta^2 = 0.91$; Number intrusions, $F(1, 8) = 59.31$, $p < .001$, $\eta^2 = 0.88$, and the different-sex background, Color intrusions, $F(1, 8) = 13.69$, $p = .006$, $\eta^2 = 0.63$; Number intrusions, $F(1, 8) = 33.61$, $p < .001$, $\eta^2 = 0.81$. There was also an interaction between background rhythm alteration and background talker sex for both Color, $F(3, 48) = 9.49$, $p < .001$, $\eta^2 = 0.37$, and Number, $F(3, 48) = 7.667$, $p < .001$, $\eta^2 = 0.324$, intrusions. Slopes for the same-sex background condition ($b = -0.28$ for Color, $b = -0.31$ for Number) were much more negative than those for the different-sex background condition ($b = -0.022$ for Color, $b = -0.026$ for Number).

Discussion

The finding of a target-rhythm effect for both same-sex and different-sex background conditions is consistent with the selective entrainment hypothesis and the results of McAuley et al., (2020). With increasing alteration of the target rhythm, proportion of correct Color and Number responses decreased, and proportion of Color and Number intrusion errors increased for both the same-sex and different-sex background conditions. Thus, the effect of target rhythm alteration on target understanding and intrusion errors is the same regardless of the ease of segregation (based on an F0 difference) of the background speech from the target speech. Results also show evidence of a contribution of disparity-based segregation to the target-rhythm effect in the same-sex condition. The increase in rhythmic disparity with increases in target rhythm alteration appears to have helped performance in the same-sex condition by reducing the negative effect of target rhythm alteration on entrainment, relative to the different-sex condition. Thus, it appears that any positive effect of rhythm disparity is less effective when a strong cue for perceptual segregation (an F0 difference in this case) is present.

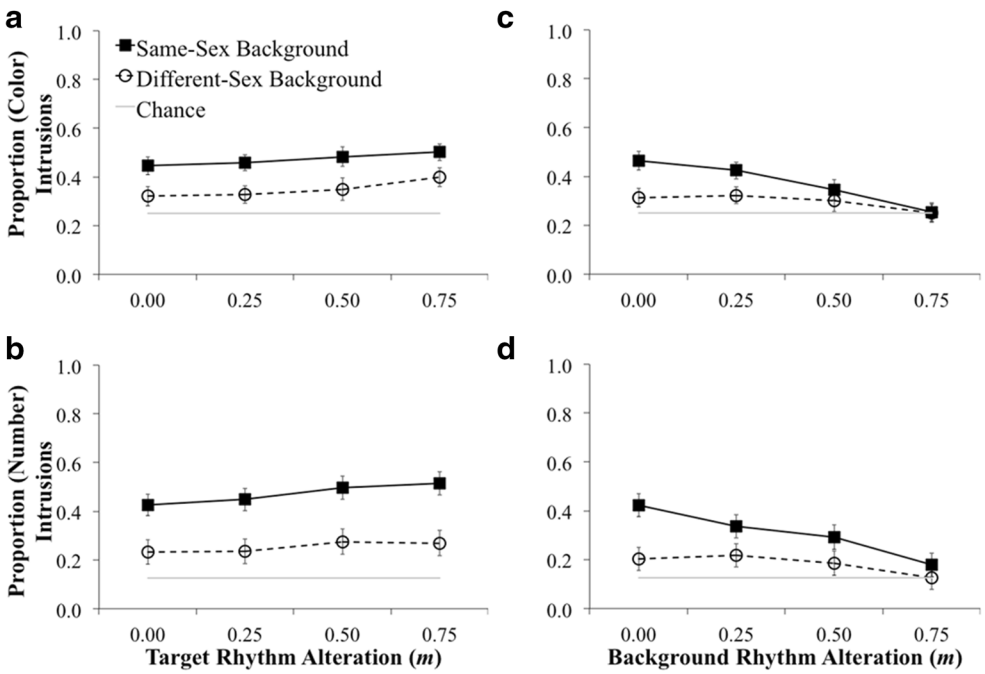

Fig. 3 Proportion of intrusions for each level of rhythm alteration (m = 0.0, 0.50, 0.25, 0.75). Intrusions with increasing target rhythm alteration (same-sex condition: n = 10, different-sex condition: n = 8) are shown on the left in **a** for Color and **b** for Number. Intrusions with increasing background rhythm alteration (same-sex condition: n = 9, different-sex condition: n = 9) are shown on the right in **c** for Color and **d** for Number. Dashed lines (with open circles) represent the different-sex background-talker condition and solid lines (with filled squares) represent the same-sex background-talker condition. Grey lines represent the chance of selecting the Color or Number in the background at random. Error bars represent standard error of the mean

Consistent with McAuley et al. (2020), there was also a background-rhythm effect, whereby alteration of the background rhythm enhanced target speech understanding and decreased intrusions. However, for both correct responses and intrusions, the changes in performance with increasing background rhythm alteration were much smaller for the different-sex background condition than the same-sex background condition. Thus, it appears that when the background is easily segregated via F0 cues (different-sex background condition), introducing rhythmic irregularity to background speech provides little to no advantage for target speech understanding. That is, listeners' attention is less likely to be entrained by the to-be-ignored background when it is rhythmically irregular, but a rhythmic disparity between background and target can also provide a segregation cue. However, the results further suggest that both mechanisms have less influence when stimuli are easily segregated based on other factors.

One additional result that warrants discussion is that, overall, there was a marked difference in the proportion of intrusion errors between the same-sex and different-sex background conditions. These differences were present even when the rhythm was unaltered, at which point proportion correct scores did not differ between the two backgrounds. In the different-sex background condition, the overall performance on Color/Number recognition was dominated by energetic masking from the speech-shaped noise: without it, performance would have been substantially higher. (Recall that the SNR for the different-sex condition was 6 dB lower than in the same-sex condition, in order to equate performance in the m = 0 condition.) The lower number of intrusions for the different-sex background compared to the same-sex background supports the idea that the different-sex background talker formed a separate auditory stream.

Experiment 2

Experiment 2 addresses the question of whether the envelope-based rhythm of the background has an effect on selective listening to the target when temporal fine structure is removed, rendering the background unintelligible. To investigate this question, Experiment 2 applies a tone-vocoding process to the same-sex single-talker background CRM sentences from Experiment 1 and then uses the same rhythm alteration manipulation in a similar experimental design to investigate the background-rhythm effect. A test of the target-rhythm effect is also included in the design to further investigate whether that effect is dependent upon the nature of the competing background pattern.

Methods

Participants and design Twenty participants (three males, 17 females), recruited from the Michigan State University Department of Psychology subject pool, participated in the

experiment. Participants were native speakers of American English and were screened for normal hearing (PTA <20 dB HL, in both ears). The experiment implemented a 2 (type of rhythm alteration: target or background) × 4 (level of rhythm alteration: m = 0, 0.25, 0.50, 0.75) mixed factorial design. Type of alteration was a between-subjects factor, leading to two participant groups (target rhythm altered, n = 10; background rhythm altered, n = 10). The level of rhythm alteration was manipulated within subjects.

Stimuli In Experiment 2, the same single-talker background sentences as in Experiment 1 were tone vocoded following the rhythm alteration to make them unintelligible while maintaining their broadband temporal envelopes. The tone vocoding was applied using the following processing steps. First, the long-term spectrum and the broadband temporal envelope were computed for each background sentence. The envelope extraction was done by first half-wave rectifying the waveform of the sentence, followed by low-pass filtering using a sixth-order Butterworth filter at 32 Hz. Second, a harmonic tone complex was then generated with a fundamental frequency (F0) of 100 Hz. The tone complex consisted of all harmonics between 100 and 8000 Hz summed in random phase. Finally, each complex was spectrally filtered to match the long-term spectrum of the selected background sentence and amplitude-modulated to match the broadband envelope of the sentence. The signal processing applied here is similar to a single-channel noise-vocoder (Shannon et al., 1995), except that the temporal fine structure of speech was replaced by that of a harmonic complex rather than a broadband noise. In order to match performance at the unaltered (m = 0) condition to the same-sex background talker condition from Experiment 1, speech-shaped noise was added to the background to produce an SNR (target relative to speech-shaped noise) of −6 dB. This value was selected through pilot testing.

Procedure The experiment was conducted in 16 blocks of 40 trials. The level of rhythm alteration remained constant within a block. Each level of rhythm alteration occurred four times, once within each set of four blocks. Across each set of four blocks, the order of rhythm alteration conditions was counterbalanced. Trial blocks were presented in one of four orders. Participants were encouraged to take breaks with a mandatory break after eight blocks (halfway through the experiment). Afterward, participants completed surveys on their personal and musical background, as well as on any strategies they used while performing the task. The entire session took approximately 1.5 hours.

Results

Figure 4 shows mean proportion of correct responses (reporting both the correct target Color and Number) for the target-rhythm and background-rhythm conditions, at each of the four levels of rhythm alteration with tone-vocoded background. A 2 × 4 mixed factorial ANOVA on proportion correct revealed a main effect of type of rhythm alteration (target or background), $F(1, 18) = 5.68$, $p = .028$, $\eta^2 = 0.24$, a main effect of level of rhythm alteration (m = 0.0, 0.25, 0.50, 0.75), $F(3, 54) = 12.83$, $p < .001$, $\eta^2 = 0.42$, and a significant interaction between type of rhythm alteration and level of rhythm alteration, $F(3, 54) = 10.26$, $p < .001$, $\eta^2 = 0.36$. Altering the natural rhythm of target speech produced the expected negative linear trend, $F(1, 9) = 70.53$, $p < .001$, $\eta^2 = 0.89$; increasing alteration of the target rhythm reduced target recognition. The slope of target-rhythm effect for the tone-vocoded background (b = −0.19) was very similar to the slope for the male (same-sex) background condition in Experiment 1 (b = −0.17).

In contrast to the same-sex background condition in Experiment 1, there was not a background-rhythm effect with the toned-vocoded background $F(1, 9) = 0.21$, $p = .66$, $\eta^2 = 0.023$. Thus, the rhythmic pattern associated with the broadband envelope of the background speech was not sufficient to produce the background-rhythm effect in the absence of temporal fine structure and semantic information. Proportion of correct responses in the condition with no rhythm alteration (m = 0) did not differ between the target rhythm (M = 0.36, SD = 0.046) and background rhythm (M = 0.38, SD = 0.12) groups, $t(18) = -0.50$, $p = .63$, 95% CI [−0.11, 0.070] (equal variances not assumed), confirming that for the two groups of participants, baseline performance in the unaltered rhythm condition was the same.

Discussion

The tone-vocoding applied in Experiment 2 resulted in stimuli with an intact broadband envelope, but which were

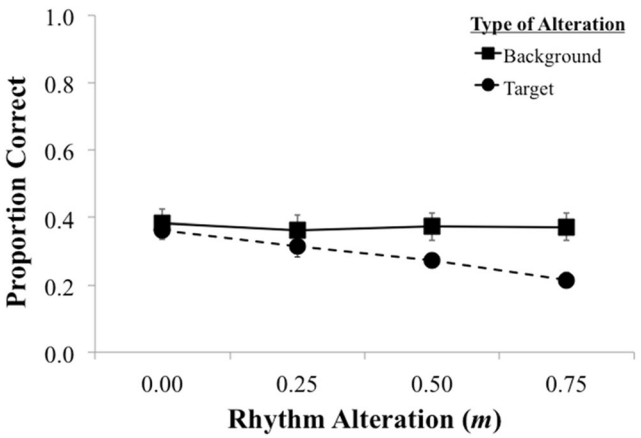

Fig. 4 Experiment 2: Proportion correct target Color and Number recognition for each level of rhythm alteration (m = 0.0, 0.50, 0.25, 0.75) with a tone-vocoded background. Dashed lines (with circles) show performance when the target rhythm was altered (n = 10), while solid lines (with squares) show performance with the altered background rhythm (n = 10). Error bars represent standard error of the mean

unintelligible and lacked the dynamic variations in spectral and temporal fine structure over time that exist in natural speech. The lack of a background-rhythm effect with a tone-vocoded background suggests that the background-rhythm effect depends on more than just the amplitude envelope of the background and its relation to the target rhythm. Although the lack of potentially interfering semantic information may account for the insensitivity to changes in the background rhythm, it is also possible that the lack of speech fine structure may have made it easier to segregate target and background, based on acoustic differences unrelated to the envelope-based rhythm. The matching of the background F0 to the target was intended to match the difficulty of perceptual segregation in the same-sex condition of Experiment 1, but the other acoustic differences resulting from the vocoding may have made it easier to segregate target and background. Regardless of the roles of acoustic differences or the lack of semantic information, it is clear that the envelope-based speech rhythm alone was not sufficient to produce the background-rhythm effect under these conditions.

The target-rhythm effect, in contrast, is robust to different types of backgrounds. The observation of a target-rhythm effect amid a background of vocoded speech in Experiment 2 extends the results of Experiment 1, where we observed a target-rhythm effect in both same-sex and different-sex background conditions. This provides further evidence that the target-rhythm effect is driven by a weakening of selective entrainment to target speech that occurs in difficult listening situations regardless of type of background.

General discussion

Taken together, the results of Experiments 1 and 2 investigating the target and background-rhythm effects are most consistent with the selective entrainment hypothesis, but also reveal contributions of disparity-based segregation to speech understanding in multitalker contexts. When listening to speech in the presence of competing sounds, speech recognition performance decreases with increasing deviations from the natural rhythm of the target speech. This target-rhythm effect, which is consistent with the selective entrainment hypothesis, but not the disparity-based segregation hypothesis, does not depend on the spectral and temporal characteristics of the competing background sounds. However, the finding of a smaller effect of target-rhythm alteration when target and background talkers had the same F0 (eliminating pitch as a segregation cue) suggests that there may be a contribution of disparity-based segregation that is evident (weakening the target-rhythm effect) only when segregation is difficult. Overall, these results support the view that the target-rhythm effect is primarily due to poorer entrainment to the target when its rhythm is altered. This is consistent with the earlier findings of McAuley et al. (2020) who found a target-rhythm effect and other support for the selective entrainment hypothesis using a similar paradigm with backgrounds consisting of multiple talkers (two or six), as well as with speech-shaped noise.

Facilitation of target word recognition by altering the natural rhythm of background speech (the background-rhythm effect) was also replicated in the present study. However, in contrast to the target-rhythm effect, the background-rhythm effect was only observed when listening to a male talker in a background consisting of a single male talker (same-sex background condition in Experiment 1). The background rhythm had little or no effect on target word recognition when listening to a male talker with a background consisting of a single female talker (different-sex background condition) or tone-vocoded speech.

For the different-sex background, the large F0 difference between the target and background talker causes the background speech to be perceptually segregated from the target. Consequently, the background rhythm may have less influence on entrainment to the target rhythm, even when the target and background rhythms are fairly similar. The background-rhythm effect appears to occur only when patterns are difficult to segregate, and an effortful listening strategy is required to track the target speech pattern and avoid intrusions from the background speech that follows a similar rhythm to the target speech. In this case, disrupting the natural rhythm of the background speech reduces the likelihood of entrainment to the background speech and facilitates attentional tracking of the target speech.

One question that emerges is whether the lack of a background-rhythm effect for the different-sex condition could be because performance was matched across the same-sex and different-sex background conditions in the intact rhythm condition ($m = 0$) by having more speech-shaped noise added in the different-sex background condition. This was done so that when examining the effects of both target and background rhythm alteration, we were starting at the same performance level for both the same-sex and different-sex conditions. However, this leaves open the possibility that the lack of a background-rhythm effect for the different-sex condition could be due to the disparity in the amount of speech-shaped noise across the two conditions. That is, more speech-shaped noise in the different-sex background resulted in greater energetic masking of the background speech and less potential for speech rhythm-based masking release. This seems unlikely for two reasons. First, the amount of added noise is the same for the background-rhythm and target-rhythm manipulations, yet, although the background-rhythm effect is not observed for the different-sex background condition, the target-rhythm effect is observed with a different-sex background. Second, in pilot work, we did match the SNR of the additional speech-shaped noise for the two background conditions and showed a similar lack of a background rhythm effect.

For the vocoded background, the lack of a background rhythm effect may have been due to a number of factors.

First, the vocoded background speech implemented in the current study only carried the broadband envelope and long-term spectrum of the original speech and was thus unintelligible. The background rhythm effect, demonstrated for multitalker backgrounds by McAuley et al. (2020) and for the same-sex single-talker background in Experiment 1 was largely driven by the reduction in intrusions from the keywords in the background sentences as the rhythm of the background was made increasingly irregular. It is possible that the vocoded background, with no intelligible Color and Number word, was not able to generate intrusions and hence no background rhythm effect was observed.

Second, the lack of intelligibility and other acoustic properties of intelligible speech in the vocoded stimuli may have made entrainment less likely in the context of a speech-recognition task. Peelle et al. (2013) showed that the strength of neural entrainment by speech is weaker when speech is less intelligible, suggesting that entrainment by speech is not entirely driven by a response to envelope-based rhythms. It may be that despite the acoustic similarity to the target speech (e.g., similarity in F0 and temporal envelope), a background sound does not effectively compete for entrainment in a speech-recognition task if it is not sufficiently speech-like. Consequently, alterations to the background rhythm in non-speech sounds may have limited influence on entrainment and selective listening to the target sentences.

A third factor that may have played a role in the lack of a background rhythm effect in Experiment 2 is that the acoustic changes resulting from the vocoding process may have facilitated perceptual segregation simply by decreasing the acoustic similarity between the target and background. Despite of the fact that the F0 of the vocoded background speech was chosen to match closely to that of the target, the observed effect of background rhythm alteration was more similar to the different-sex than the same-sex background condition in Experiment 1. Of note, both the different-sex and vocoded backgrounds required the same 6 dB of additional speech-shaped noise (introduced to equate performance in the condition with no rhythm alteration) to achieve performance similar to that in the same-sex condition. And as the degree of background rhythm alteration increased, performance did not improve for either the different-sex or vocoded background, while performance improved significantly for the same-sex condition. Thus, it may be that the vocoded background was perceptually segregated from the target, despite the similarity in F0. By removing the variations in F0 and short-term spectrum in the background sentences while preserving the long-term spectrum and broadband envelope of the background speech, it appears that the vocoding process used in the current study may have resulted in a background that was perceptually distinct from the target.

These results show that the target-rhythm and background-rhythm effects are not simply due to interference between target and background speech rhythms defined by the temporal envelopes. Rather, they reflect complex interactions between perceptual segregation and selective entrainment. When the target and background speech are easily segregated based either on acoustic cues such as F0 difference, the listener may be able to selectively listen to the target speech among many already formed perceptual streams and the recognition of key words in the target speech depends on the entrainment to the target rhythm. In this case, introducing irregularity to the target rhythm would undermine the efficiency in entrainment and hence adversely impact recognition performance. In contrast, altering the rhythm of perceptually segregated background speech does not affect performance. This lack of rhythm interaction across perceptual streams is consistent with listeners' insensitivity to between-stream pitch and temporal relations observed in many studies of auditory stream segregation (see Bregman, 1990).

When the target speech cannot be easily perceptually segregated from the background (perhaps by automatic or early perceptual mechanisms), selective listening to the target speech is affected by competition for entrainment among co-occurring speech rhythms. Under these more difficult listening conditions, selective listening may also depend on schema-based segregation mechanisms that use knowledge of the spectrotemporal constraints of speech to guide the selection of the target speech (see Bey & McAdams, 2002; Bregman, 1990; Moore & Gockel, 2012; Szalárdy et al., 2020, for related discussions of primitive vs. schema-based mechanisms). These mechanisms may involve speech intelligibility, separating the auditory scene into an intelligible target stream and an unintelligible background stream. Selective entrainment may play a crucial role here by facilitating attentional alignment to the rhythm of target speech that conforms to the expected temporal regularity of naturally produced speech (e.g., Schröger et al., 2014).

In summary, while the target rhythm effect does not depend on the difficulty of perceptually segregating competing sounds, the background rhythm effect does. Degrading the rhythmic regularity of the target disrupts entrainment to the target rhythm and increases the likelihood of accidental entrainment to the to-be-ignored competing background speech rhythm, regardless of the acoustic properties of the background speech. In contrast, degrading the regularity of the background rhythm *facilitates* entrainment to the target rhythm by removing competition for entrainment from co-occurring speech rhythms; but only when the competing speech is difficult to perceptually segregate from the target speech.

Although the background rhythm effect was not observed with vocoded background speech in the current study, it would be premature to conclude that this effect only occurs with intelligible background speech. The vocoding procedure implemented in Experiment 2 removed both intelligibility and semantic information, but it also removed acoustic properties

that may be important for the perception of speech rhythm and selective entrainment. It may be that F0 contour or some aspects of the temporal fine structure of speech affect perceived similarity and/or perceived rhythm in ways make a competing sound more difficult to segregate and more likely to entrain attentional rhythms when listening to speech, even in the absence of semantic information. Future studies will examine the background-rhythm effect with different types and degrees of similarity to target speech sounds to determine the degree to which the background-rhythm effect can occur in the absence of semantic interference.

Acknowledgments The authors thank Anusha Mamidipaka, Alison Eberle, Becca Vroegop, Yan Cong, and Nicole Boog for their assistance with data collection and helpful insights, Dylan V. Pearson at Indiana University for assistance with stimulus generation and members of the Timing, Attention and Perception Lab at Michigan State University for their helpful suggestions and comments at various stages of this project. NIH Grant R01DC013538 (PIs: Gary R. Kidd and J. Devin McAuley) supported this research.

References

Akroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology, 47*, 53–71. https://doi.org/10.1080/14992020802301142

Allen, K., Carlile, S., & Alais, D. (2008). Contributions of talker characteristics and spatial location to auditory streaming. *The Journal of the Acoustical Society of America*, *123*(3), 1562–1570. https://doi.org/10.1121/1.2831774

Assmann, P. F., & Summerfield, Q. (1989). Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency. *The Journal of the Acoustical Society of America*, *85*(1), 327–338. https://doi.org/10.1121/1.397684

Assmann, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, *88*(2), 680–697. https://doi.org/10.1121/1.399772

Aubanel, V., Davis, C., & Kim, J. (2016). Exploring the role of brain oscillations in speech perception in noise: intelligibility of isochronously retimed speech. *Frontiers in Human Neuroscience, 10*, 430. https://doi.org/10.3389/fnhum.2016.00430

Bacon, S. P., & Grantham, D. W. (1989). Modulation masking: Effects of modulation frequency, depth, and phase. *The Journal of the Acoustical Society of America, 85*(6), 2575–2580. https://doi.org/10.1121/1.397751

Baese-Berk, M. M., Dilley, L. C., Henry, M. J., Vinke, L., & Banzina, E. (2019). Not just a function of function words: Distal speech rate influences perception of prosodically weak syllables. *Attention, Perception, & Psychophysics*, *81*(2), 571–589. https://doi.org/10.3758/s13414-018-1626-4

Barnes, R., & Jones, M. R. (2000). Expectancy, attention, and time. *Cognitive Psychology, 41*, 254–311. https://doi.org/10.1006/cogp.2000.0738

Bendixen, A. (2014). Predictability effects in auditory scene analysis: a review. *Frontiers in Human Neuroscience, 8*, 60. https://doi.org/10.3389/fnins.2014.00060

Bey, C., & McAdams, S. (2002). Schema-based processing in auditory scene analysis. *Perception & Psychophysics*, *64*(5), 844–854. https://doi.org/10.3758/BF03194750

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *Journal of the Acoustical Society of America, 107*(2), 1065–1066. https://doi.org/10.1121/1.428288

Bregman, A. S. (1990). *Auditory scene analysis*. MIT Press.

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics, 10*, 23–36. https://doi.org/10.1016/S0095-4470(19)30909-X

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109. https://doi.org/10.1121/1.1345696

Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Sciences, 8,* 465-471. https://doi.org/10.1016/j.tics.2004.08.008

Darwin, C. J. (1975). On the dynamic use of prosody in speech perception. *Haskins Laboratories Status Report on Speech Research 42–43*, 103–115.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics, 11,* 51–62. https://doi.org/10.1016/S0095-4470(19)30776-4

Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language, 59,* 294–311. https://doi.org/10.1016/j.jml.2008.06.006

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences,* 109(29), 11854–11859. https://doi.org/10.1073/pnas.1205381109

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in Human Neuroscience*, 8, 311. https://doi.org/10.3389/fnhum.2014.00311

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience, 19,* 158.

Fogerty, D., Xu, J., & Gibbs, B. E. (2016). Modulation masking and glimpsing of natural and vocoded speech during single-talker modulated noise: Effect of the modulation spectrum. *The Journal of the Acoustical Society of America, 140*(3), 1800–1816. https://doi.org/10.1121/1.4962494

George, M. F. S., & Bregman, A. S. (1989). Role of predictability of sequence in auditory stream segregation. *Perception & Psychophysics*, *46*, 384–386.

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology, 2*, 130. https://doi.org/10.3389/fpsyg.2011.00130

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

Golumbic, E. M. Z., Poeppel, D., & Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. *Brain and Language, 122*, 151–161. https://doi.org/10.1016/j.bandl.2011.12.010

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Simon, J. Z., Poeppel, D., & Schroeder, C. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron, 77*, 980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Gordon-Salant, S., Fitzgibbons, P. J., & Friedman, S. A. (2007). Recognition of time-compressed and natural speech with selective temporal enhancements by young and elderly listeners. *Journal of Speech, Language, and Hearing Research, 50*(5), 1181–1193. https://doi.org/10.1044/1092-4388(2007/082)

Goswami, U. (2019). Speech rhythm and language acquisition: an amplitude modulation phase hierarchy perspective. *Annals of the New York Academy of Sciences, 1453,* 67–78. https://doi.org/10.1111/nyas.14137

Houtgast, T. (1989). Frequency selectivity in amplitude-modulation detection. *The Journal of the Acoustical Society of America, 85*(4), 1676–1680. https://doi.org/10.1121/1.397956

Houtgast, T., & Festen, J. M. (2008). On the auditory and cognitive functions that may explain an individual's elevation of the speech reception threshold in noise. *International Journal of Audiology, 47*(6), 287–295. https://doi.org/10.1080/14992020802127109

Humes, L. E., & Dubno, J. R. (2010). Factors affecting speech understanding in older adults. In S. Gordon-Salant, R. D. Frisina, A. N. Popper, & R. R. Fay (Eds.), The *aging auditory system* (pp. 211–257). Springer.

Humes, L. E., Busey, T. A., Craig, J., & Kewley-Port, D. (2013a). Are age-related changes in cognitive function driven by age-related changes in sensory processing? *Attention, Perception, & Psychophysics*, *75*(3), 508–524. https://doi.org/10.3758/s13414-012-0406-9

Humes, L. E., Kidd, G. R., & Lentz, J. J. (2013b). Auditory and cognitive factors underlying individual differences in aided speech-understanding among older adults. *Frontiers in Systems Neuroscience, 7,* 55. https://doi.org/10.3389/fnsys.2013.00055

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review, 83*, 323–355. https://doi.org/10.1037/0033-295X.83.5.323

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review, 96,* 459–491. https://doi.org/10.1037/0033-295X.96.3.459

Jones, M. R., Kidd, G., & Wetzel, R. (1981). Evidence for rhythmic attention. *Journal of Experimental Psychology: Human Perception and Performance, 7*(5), 1059–1073. https://doi.org/10.1037/0096-1523.7.5.1059

Jones, M. R., Moynihan, H., MacKenzie, N., & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science, 13*, 313–319. https://doi.org/10.1111/1467-9280.00458

Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America*, *122*(1), 418–435. https://doi.org/10.1121/1.2743154

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review, 106*, 119–159. https://doi.org/10.1037/0033-295X.106.1.119

Lavan, N., Domone, A., Fisher, B., Kenigzstein, N., Scott, S. K., & McGettigan, C. (2019). Speaker sex perception from spontaneous and volitional nonverbal Vocalizations. *Journal of Nonverbal Behavior, 43*(1), 1–22. https://doi.org/10.1007/s10919-018-0289-0

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*, 1102–1125. https://doi.org/10.1037/0096-1523.29.6.1102

McAuley, J. D., Jones, M. R., Holub, S., Johnston, H. M., & Miller, N. S. (2006). The time of our lives: Life span development of timing and event tracking. *Journal of Experimental Psychology: General, 135*, 348–367. https://doi.org/10.1037/0096-3445.135.3.348

McAuley, J. D., Shen, Y., Dec, S., Kidd, G. (2020). Altering the rhythm of target and background talkers differentially affects speech understanding: Support for a selective-entrainment hypothesis. *Attention, Perception, & Psychophysics, 82,* 3222–3233. https://doi.org/10.3758/s13414-020-02064-5

Miller, J. E., Carlson, L. A., & McAuley, J. D. (2013). When what you hear influences when you see: Listening to an auditory rhythm influences the temporal allocation of visual attention. *Psychological Science, 24*(1), 11–18. https://doi.org/10.1177/0956797612446707

Moore, B. C., & Gockel, H. E. (2012). Properties of auditory stream formation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1591), 919–931. https://doi.org/10.1098/rstb.2011.0355

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*, 69–74. https://doi.org/10.1016/j.cognition.2013.12.006

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, *23*(6), 1378–1387. https://doi.org/10.1093/cercor/bhs118

Poon, M. S., & Ng, M. L. (2015). The role of fundamental frequency and formants in voice gender identification. *Speech, Language and Hearing, 18*(3), 161–165. https://doi.org/10.1179/2050572814Y.0000000058

Riecke, L., Formisano, E., Sorger, B., Baskent, D., & Gaudrain, E. (2018). Neural entrainment to speech modulates speech intelligibility. *Current Biology, 28*, 161–169. https://doi.org/10.1016/j.cub.2017.11.033

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *336*(1278), 367–373. https://doi.org/10.1098/rstb.1992.0070

Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America, 133*(4), 2431–2443.

Schröger, E., Bendixen, A., Denham, S. L., Mill, R. W., Bőhm, T. M., & Winkler, I. (2014). Predictive regularity representations in violation detection and auditory stream segregation: from conceptual to computational models. *Brain Topography*, *27*(4), 565–577. https://doi.org/10.1007/s10548-013-0334-6

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304. https://doi.org/10.1126/science.270.5234.303

Stone, M. A., Füllgrabe, C., & Moore, B. C. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *The Journal of the Acoustical Society of America, 132(1)*, 317-326.

Szalárdy, O., Tóth, B., Farkas, D., Orosz, G., Honbolygó, F., & Winkler, I. (2020). Linguistic predictability influences auditory stimulus classification within two concurrent speech streams. *Psychophysiology, 57*(5), e13547. https://doi.org/10.1111/psyp.13547

Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America, 134*(1), 628–639. https://doi.org/10.1121/1.4807565

Wang, M., Kong, L., Zhang, C., Wu, X., & Li, L. (2018). Speaking rhythmically improves speech recognition under "cocktail-party" conditions. *The Journal of the Acoustical Society of America, 143*, EL255–EL259.

Whiteside, S. P. (1998). The identification of a speaker's sex from synthesized vowels. *Perceptual and Motor Skills, 87*(2), 595–600. https://doi.org/10.2466/pms.1998.87.2.595

Yost, W. A., Sheft, S., & Opie, J. (1989). Modulation interference in detection and discrimination of amplitude modulation. *The Journal of the Acoustical Society of America*, *86*(6), 2138–2147. https://doi.org/10.1121/1.398474