Attention, Perception, & Psychophysics (2020) 82:3222–3233
https://doi.org/10.3758/s13414-020-02064-5

Altering the rhythm of target and background talkers differentially affects speech understanding

J. Devin McAuley[1] · Yi Shen[2] · Sarah Dec[1] · Gary R. Kidd[2]

Published online: 26 May 2020

Abstract
Three experiments investigated listeners' ability to use speech rhythm to attend selectively to a single target talker presented in multi-talker babble (Experiments 1 and 2) and in speech-shaped noise (Experiment 3). Participants listened to spoken sentences of the form "Ready [Call sign] go to [Color] [Number] now" and reported the Color and Number spoken by a target talker (cued by the Call sign "Baron"). Experiment 1 altered the natural rhythm of the target talker and background talkers for two-talker and six-talker backgrounds. Experiment 2 considered parametric rhythm alterations over a wider range, altering the rhythm of either the target or the background talkers. Experiments 1 and 2 revealed that altering the rhythm of the target talker, while keeping the rhythm of the background intact, reduced listeners' ability to report the Color and Number spoken by the target talker. Conversely, altering the rhythm of the background talkers, while keeping the target rhythm intact, improved listeners ability to report the Color and Number spoken by the target talker. Experiment 3, which embedded the target talker in speech-shaped noise rather than multi-talker babble, similarly reduced recognition of the target sentence with increased alteration of the target rhythm. This pattern of results favors a dynamic-attending theory-based selective-entrainment hypothesis over a disparity-based segregation hypothesis and an increased salience hypothesis.

Keywords Speech perception · Attention: Selective · Temporal Processing

Introduction

Speech is a dynamic signal with temporal features that are essential for speech understanding. Speech timing has been shown to be crucial for speech perception in a wide variety of studies (for reviews, see Darwin, 1975; Golumbic, Poeppel, & Schroeder, 2012; Huggins, 1972; Jones & Boltz, 1989; Rosen, 1992). One of the important temporal features of speech is speech rhythm; 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. Many studies have shown that listeners' perception of speech is sensitive to speech rhythm context (Dilley & McAuley, 2008; Kidd, 1989; Peelle & Davis, 2012; Smith, Cutler, Butterfield, & Nimmo-Smith, 1989). From this perspective, there are many levels of temporal structure to consider, ranging from the relative timing of articulatory events within a syllable to the relative timing of phrases and larger groups of words or sentences. Primary contributors to speech rhythm are the timing of syllables, which are produced at rates between 3–9 Hz across a number of languages (see Dauer, 1983; Tilsen & Arvaniti, 2013), and the temporal patterns of syllabic stress (conveyed largely by amplitude, but also by duration and pitch change).

The theoretical basis for a consideration of the role of speech rhythm in spoken language processing is based on the broader theoretical framework of dynamic attending theory (DAT) developed by Jones and colleagues (Jones, 1976; Jones & Boltz, 1989; Large & Jones, 1999; McAuley Jones, Holub, Johnston & Miller, 2006). DAT postulates that temporal fluctuations in listeners' attention are entrained by periodic (or quasi-periodic) stimulus rhythms, such that attentional rhythms emerge that gradually align with stimulus rhythms, with attentional resources allocated at rhythmically expected (entrained) time points (Jones, 1976; Jones & Boltz, 1989; Large & Jones, 1999; McAuley et al., 2006). Thus, DAT predicts that stimulus events that occur at rhythmically expected

✉ 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, Indiana University, Bloomington, IN, USA

time points are better resolved than stimulus events that occur at unexpected time points.

Behavioral support for DAT has been found in a range of domains showing better detection and/or discrimination of stimulus events that occur at rhythmically expected times than unexpected times (Barnes & Jones, 2000; Jones et al., 2002; McAuley & Jones, 2003; Miller, Carlson & McAuley, 2013). In the domain of speech and language, ambiguously organized syllable sequences have been found to be perceived differently (e.g., in terms of phonemic categories or segmentation) depending on the rhythmic context presented earlier, supporting the view that listeners are sensitive to speech rhythm and that speech rhythms lead to the development of temporal expectations that influence the downstream processing of events in the speech stream (Baese-Berk et al., 2019; Dilley & McAuley, 2008; Kidd, 1989; Morrill et al., 2014). Neuroscientific support for DAT in the domain of speech and language has blossomed in recent years, providing increasing evidence that brains are entrained by speech rhythms, exhibiting neural oscillations at rates similar to the syllabic rate of speech (i.e., theta oscillations), which become phase-locked to the temporal envelope of speech. These studies have argued that, in line with DAT, neural entrainment to speech envelope is a fundamental neural mechanism for parsing running speech signals into smaller temporal units for comprehension (Ding et al., 2016; Ghitza, 2011; Giraud & Poeppel, 2012; Poeppel, 2003; Riecke et al., 2018). This research also reinforces the idea that speech understanding is not a passive, stimulus-driven process, but rather involves anticipation and active predictions of future events (Peelle & Davis, 2012).

In everyday listening situations, speech is not often heard in isolation but occurs amidst various sources of background noise. In many instances (e.g., a busy café or restaurant), background sounds consist of competing talkers. The study of listeners' ability to selectively attend to a target talker while ignoring background talkers, often termed the cocktail-party problem (after Cherry, 1953), has been extensively studied both behaviorally and neurophysiologically (see a collection of reviews in a volume edited by Middlebrooks et al., 2017). However, the specific role of talker rhythm in understanding speech in difficult listening situations has received relatively little attention. Towards this end, this article contrasts a selective-entrainment hypothesis, based on DAT, with two alternative hypotheses about how speech rhythm(s) of target and background talkers affect selective listening in a multi-talker environment.

In multi-talker listening environments, rhythmic information, carried by both the target and background speech, potentially affects selective attention to a target talker in several possible ways. First, differences in rhythm between target and background speech may be used to perceptually segregate the competing sound sources (Bregman, 1990). Previous research has shown that (in addition to spectral and spatial cues) differences in both amplitude and frequency modulation can be used to segregate co-occurring sounds and facilitate speech recognition in noise (e.g., Bregman et al., 1985; Marin & McAdams, 1991; McAdams, 1989; Zeng et al., 2005). A difference in tempo, or speaking rate, between talkers has also been shown to facilitate selective listening, with substantial improvements in speech recognition as the difference in tempo between the target and background speech increases (Kidd & Humes, 2014). In addition to differences in tempo or rate of modulation, differences in temporal structure or rhythm, can also facilitate the segregation of overlapping sounds from different sources, especially when the target stimulus has a predictable temporal pattern (e.g., Jones, Kidd, & Wetzel, 1981; Rimmele et al., 2011, 2012; Snyder et al., 2012). Similarly, a difference in rhythm in target and background speech (e.g., two talkers exhibiting different degrees of rhythmic regularities in their speech) may also promote the segregation of the target speech from the background, leading to improved intelligibility of the target speech. We will refer this hypothesis as the *disparity-based segregation hypothesis.* This hypothesis would predict improved intelligibility of target speech presented amidst background speech when the rhythms of the target or background speech are made increasingly dissimilar, regardless of the rhythmic properties of the individual speech streams.

A second, related, hypothesis is an *increased salience hypothesis* whereby a rhythmic difference between target and background speech may lead to increased *salience* of the rhythm that is atypical, thereby attracting greater attention. This hypothesis would predict improved intelligibility of target speech presented amidst background speech when the natural rhythm of the target speech is altered to make it atypical (unnatural), and *reduced* intelligibility of the target speech when the natural rhythm of the background speech is made more salient by rhythmic alteration.

In contrast to the above two hypotheses, a *selective entrainment hypothesis* (based on DAT) makes very different predictions. Since selective attentional entrainment relies on a quasi-regular rhythmic structure, there should be more effective attentional entrainment with target speech that has a predictable natural (quasi-periodic) rhythm than with speech with a less natural (i.e., arrhythmic or disordered) rhythm. Thus, from a DAT perspective, altering the natural rhythm of the target speech in a multi-talker environment should undermine selective entrainment to the target and lead to poor recognition of the target speech. Conversely, altering the natural rhythm of the background should *facilitate* selective attention to the target (and thereby enhance recognition of target speech) by reducing potential interference from entrainment to the background. According to this hypothesis, rhythmic disparity is not helpful unless it facilitates entrainment to the target speech.

The selective entrainment, disparity-based segregation, and increased-salience hypotheses were tested in a series of three experiments that introduced alterations to the natural rhythms of

target and background speech in multi-talker listening environments. Participants listened to spoken sentences of the form "Ready [Call sign] go to [Color] [Number] now" and reported the Color and Number spoken by a target talker (cued by the Call sign "Baron"). Experiment 1 independently altered the natural speech rhythms of the target and the background talkers, comparing the effects of the rhythm alterations for two versus six asynchronous background talkers. We compared two versus six background talker conditions because the number of background talkers has been previously shown to affect overall performance in intact rhythm conditions (Eddins & Liu, 2012). Eddins and Liu (2012) considered the psychometric properties of the CRM paradigm for various types of background sounds, including conditions that varied the number of background talkers. Two-talker and four-talker backgrounds showed similar psychometric properties, with the four-talker background condition more difficult than the two-talker background conditions at the same SNR levels. Here, we were interested in examining whether any effect of altering the rhythm of the background talkers would depend on the number of overlapping speech rhythms in the background. Experiment 2 considered parametric rhythm alterations over a wider range of magnitudes, altering the rhythm of either the target or the background talkers. Experiment 3 altered the natural rhythm of the target speech embedded in speech-shaped noise rather than a multi-talker background in order to examine the dependence of an effect of target rhythm on the type of background.

According to the disparity-based segregation hypothesis, rhythm alterations to either the target speech or background speech would increase the difference between the target and background rhythms, which should facilitate perceptual segregation and lead to improved recognition of the target speech. On the other hand, according to the selective entrainment hypothesis, rhythm alterations to the target speech would undermine the selective entrainment to the target speech, leading to degraded recognition (Aubanel, Davis, & Kim, 2016; Wang et al., 2018). Moreover, both the disparity-based segregation and selective entrainment hypotheses predict that rhythm alterations to the background speech should improve the recognition of the target speech, either through decreased similarity in rhythms of the target and background talkers or through reduced competition for entrainment to the target speech. The increased salience hypothesis, in contrast, predicts degraded target recognition with rhythm alterations to the background speech.

General methods

Speech stimuli used in the current experiment were sentences from the Coordinate Response Measure (CRM) corpus (Bolia et al. 2000). Each CRM sentence was in the format of "Ready [Call sign] go to [Color] [Number] now." Sentences used in this study included eight different Call Signs ("Baron," "Charlie," "Eagle," etc.), four different Colors ("Blue," "Red," "Green," and "White"), and seven different numbers (1–8, excluding 7 to limit all numbers to a single syllable). Across the three experiments, recordings of the sentences from four male talkers (one used as target and three used as background) and three female talkers (for background), were used, for a total of 1,568 sentences. The Call sign for the target was always "Baron" and the target talker was the same male talker for all experiments. The participants' task was to listen for the sentence with the Call sign "Baron" and identify the Color and Number in that sentence. Participants provided responses by using a mouse to select the correct Color and Number combination on a computer running a custom MATLAB program. With four Colors and seven Numbers, chance performance on each trial was 1/28 (~ 3.6%).

The target was presented with a simultaneous competing background. The background was either a multi-talker babble consisting of multiple CRM sentences (in Experiments 1 and 2), or a broadband noise spectrally shaped to match the long-term spectrum of the target speech (Experiment 3). When present, the background sentences had Call signs, Colors, and Numbers that were different from those in the target sentence. For the two- and six-talker backgrounds, each background sentence was produced by a unique talker from the CRM corpus, always different from the target talker, and the background always consisted of an equal number of male and female talkers. For the two-talker background, onset asynchronies of -50, and 50 ms, relative to the onset of the target sentence, were assigned to the two randomly selected background CRM sentences. For the six-talker background, onset asynchronies of -150, -100, -50, 50, 100, and 150 ms were used for the six randomly selected background sentences. The level of the target was fixed at 65 dB SPL and the overall level of the background, including all background sentences, was set to signal-to-noise ratios (SNRs) of -6, 0, and -2 dB for the noise, two-talker, and six-talker backgrounds, respectively.[1]

[1] The 0 and -2 dB SNRs for the two- and six-talker backgrounds were selected to equate the performance in recognizing both Color and Number at a level that was approximately 50% correct for the unaltered rhythm conditions. This performance level agreed well with the psychometric properties of the CRM-sentence recognition in noise reported in previous studies. Eddins and Liu (2012) measured the recognition of Color and Number in CRM sentences in two- and four-talker babble noises. In the two- and four-talker background conditions, the SNRs required to reach 50% correct for Color and Number recognition were -1.3 dB and -5.4 dB, respectively. Similar to the current study, a lower SNR was needed to equate the task performance for the greater number of background talkers. This also generally agrees with various other studies that measured open-set speech recognition in multi-talker backgrounds. For example, Rosen et al. (2013) measured sentence recognition in multi-talker babble for fixed SNRs of –6 and –2 dB, as the number of background talkers increased from 2 to 16; here, recognition performance gradually improved about 2% for every doubling of the number of background talkers. Freyman et al. (2004) showed that the SNR required to reach a sentence-recognition performance of 50% decreased from approximately -0.5 dB for two-talker backgrounds to approximately -3 dB for six-talker backgrounds.

These SNRs were chosen so that similar overall task performance was expected from these background conditions.

For the experiments reported here, rhythm alterations were imposed on target and background sentences. The alteration consisted of temporal expansion and contraction of portions of the sentences in a sinusoidal pattern (see Fig. 1). The rhythm alterations were realized using the Pitch Synchronous Overlap and Add (PSOLA) algorithm as implemented in Praat (e.g., Moulines & Charpentier, 1990). When applying rhythm alteration, the original sentence from the CRM corpus was temporally compressed or expanded according to a compression ratio (CR) which is a sinusoidal function of time (t): $CR(t) = 1 + m\sin(2\pi f_m t + \phi)$, where f_m is the rhythm alteration rate (set to 1 Hz) and m is the degree of temporal expansion/contraction. The value of m was thus the amount of rhythm alteration, (i.e., depth of the modulation), which was set to 0, 0.25, 0.50, and 0.75 in different conditions. The initial phase of the rhythm alteration, ϕ, was randomly drawn from the set: 0, $\pi/4$, $2\pi/4$, $3\pi/4$, $4\pi/4$, $5\pi/4$, $6\pi/4$, and $7\pi/4$, with equal probability. Pilot experiments confirmed that an f_m value of 1 Hz gives rise to a relatively strong impression of timing variation (with $m > 0.0$) that clearly disrupts the natural timing of the sentence while having a negligible effect on intelligibility (when presented in quiet).

One consequence of the rhythm alteration was that it moved the key words in the target sentence (i.e., Color and Number) to a new temporal location either earlier or later in the sentence, depending on ϕ. To ensure that the effect of rhythm alteration was not merely to create misalignment in Color and Number between the target and background, hence reducing the amount of energetic masking, onset asynchronies were introduced to misalign the key words (i.e., Color and Number) among the target and background sentences before rhythm alteration. Moreover, due to the randomized phases used for the target and background rhythm alteration, the overall degree of temporal overlap for Color and Number between the target and background did not depend on the amount of rhythm alteration (i.e., the value of m).

The intelligibility of the rhythm-altered speech presented in isolation without a competing background was confirmed with a group of 11 listeners who listened to CRM sentences at four levels of rhythm alteration: $m = 0$ (unaltered rhythms), $m = 0.25$ and $m = 0.50$ (intermediate levels of rhythm alteration), and $m = 0.75$ (the maximal level of rhythm alteration examined in the experiments reported below). For all values of m, listeners correctly identified the Color and Number at or near 100% ($M = 0.99$, $SD = 0.01$); performance was > 97% for all listeners in all conditions, except for one listener who scored 93% in one condition ($m = 0.75$). Thus, although the rhythm alteration affects the naturalness of the speech rhythm, the manipulation does not affect target speech intelligibility for the four values of m examined here.

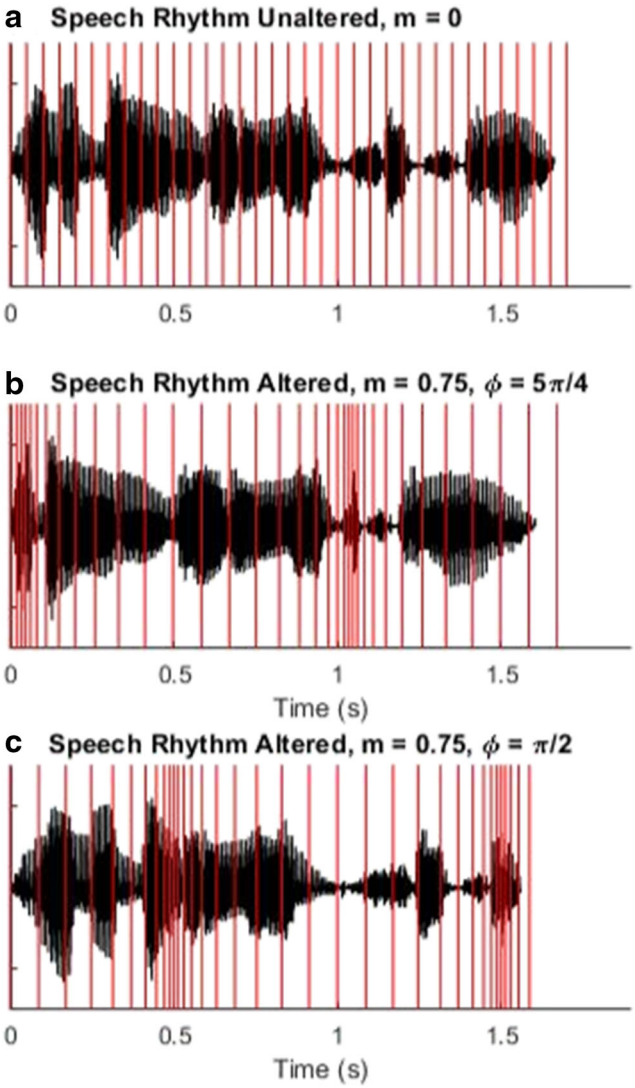

Fig. 1 Examples of rhythm unaltered and altered versions of a spoken CRM sentence of the form "Ready [Call sign] go to [Color] [Number] now.' The top panel (**Panel A**) shows the sample sentence where the rhythm is unaltered ($m = 0$), as represented by placing grid lines equally spaced in time. The middle and bottom panels shows 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 (**Panel B,** $\phi = 5\pi/4$; **Panel C,** $\phi = \pi/2$)

Analysis of variance (ANOVA) was used to assess the statistical significance ($\alpha = 0.05$) for main effects and interactions. ANOVAs were supplemented by additional trend analyses and post-hoc tests when warranted. Effect sizes (Cohen's d) are reported for all comparisons. For each of the three experiments reported below, we considered potential effects of music training on CRM performance. No significant relationships between number of years of formal music training and CRM performance were found for the intact or altered rhythm conditions (Pearson r values < 0.35, with ps > 0.1). Music training information for participants is included in each of the experiments below for completeness.

Experiment 1

Methods

Participants and design Nineteen native speakers of American English (n = 11, female), aged 18–25 years (M = 19.8, SD = 1.8), completed the experiment in return for course credit in an undergraduate psychology course. All participants were screened for normal hearing (PTA < 20 dB HL) and varied in number of years of formal music training (0–15 years, M = 4.8, SD = 4.5). The experiment implemented a 2 (number of background talkers: 2 vs. 6) × 2 (target rhythm alteration: m = 0.0 or 0.5) × 2 (background rhythm alteration: m = 0.0 or 0.5) within-subjects design. The value of m = 0.5 was selected because it represented an easily detectable change in speech rhythm that did not affect target speech intelligibility (see *General methods*).

Procedure On each trial, participants listened to CRM sentences and reported the Color and Number spoken in the target sentence, which was cued by the Call Sign "Baron." The target sentence was presented with either two- or six-talker backgrounds with 50-ms between the onsets of each sentence, with the target sentence always presented in the temporal middle. The rhythm of the background or target sentence was either intact (m = 0) or altered (m = 0.5) using all four target/background combinations to create four rhythm alteration conditions (target/background m values = 0/0, 0/0.5, 0.5/0, and 0.5/0.5). The number of background talkers and rhythm alteration condition was held constant within a block of trials. Each participant completed 16 blocks of trials with 40 trials per block for a total of 640 trials. There were two 40-trial blocks in each condition and the order of blocks was counterbalanced between participants, such that across participants all conditions were experienced in the same average position within the sequence of blocks in order to address the possibility of practice effects. After completion of the experimental trials, participants completed a survey that consisted of a series of demographic and background questions that included questions about musical experience and any strategies they may have used during the experiment. The entire experiment lasted ~1.5 h.

Results and discussion

Figure 2 shows proportion correct for identifying both Color and Number for the four rhythm alteration conditions (i.e., combinations of m values for target and background) for two-talker background (Panel A) and six-talker background (Panel B). A repeated-measures analysis of variance (ANOVA) with a 2 (number of background talkers) × 2 (target rhythm alteration depth: 0.0 vs. 0.5) × 2 (background rhythm alteration: 0.0 vs. 0.5) design revealed a main effect of number of background talkers, $F(1,18) = 6.67$, $p = 0.019$, $\eta^2 = 0.27$, a main effect of target rhythm, $F(1,18) = 122.2$, $p < 0.001$, $\eta^2 = 0.87$, a main effect of background rhythm, $F(1,18) = 216.6$, $p < 0.001$, $\eta^2 = 0.92$, and an interaction between number of background talkers and background rhythm, $F(1,18) = 9.94$, $p = 0.006$, $\eta^2 = 0.27$. There were no other reliable interactions (all ps > 0.16).

Overall performance was better for the two-talker background condition (M = 0.53, SD = 0.17) than for the six-talker background condition (M = 0.49, SD = 0.13), $t(18)$ = 2.58, $p < 0.001$, Cohen's d = 0.59. Altering the target rhythm produced an 11 percentage-point reduction in performance (altered target, M = 0.46, SD = 0.14; unaltered target, M = 0.57, SD = 0.15; Cohen's d = -2.54), while altering background rhythm resulted in a nearly identical *improvement* in performance compared to the unaltered background (altered

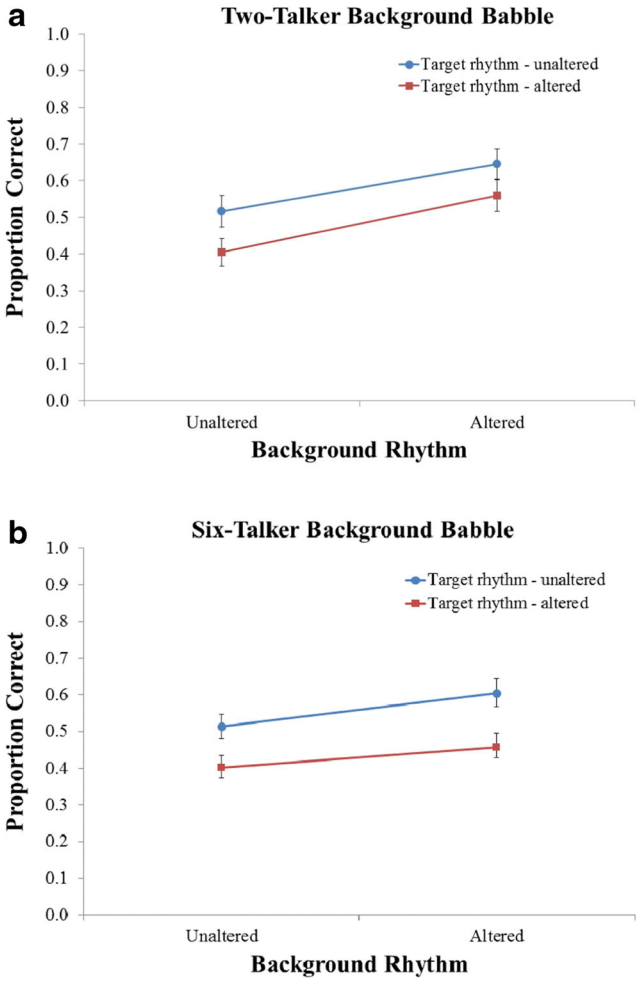

Fig. 2 Mean proportion correct target Color and Number responses for two background talkers (**Panel A**) and six background talkers (**Panel B**) for the four rhythm conditions (target unaltered – background unaltered, target unaltered – background altered, target altered – background unaltered, target altered – background altered); error bars correspond to standard error. Altering the target rhythm reduced correct reporting of the target Color and Number, while altering the background rhythm, increased correct reports of the target Color and Number

background, $M = 0.57$, $SD = 0.15$; unaltered background, $M = 0.46$, $SD = 0.14$; Cohen's $d = 3.38$). The interaction between number of background talkers and background rhythm suggested that the overall better performance with two talkers in background compared with six talkers in background was due to a larger enhancing effect of alterations in background rhythm with two talkers in the background (unaltered, $M = 0.46$, $SD = 0.16$; altered, $M = 0.60$, $SD = 0.17$), $t(18) = 9.86$, $p < 0.001$, Cohen's d = 2.26, than with six talkers (unaltered, M = 0.46, $SD = 0.16$; altered, $M = 0.53$, $SD = 0.14$), $t(18) = 6.50$, $p < 0.001$, Cohen's d = 1.49.

In sum, results of Experiment 1 favor the selective-entrainment hypothesis over the disparity-based segregation and increased-salience hypotheses. Altering the rhythm of a target talker presented amongst background talkers makes it more difficult to listen selectively to the target talker, while altering the natural rhythm of the background talkers makes it easier to focus on the target sentence and ignore the background talkers.

Although the results of Experiment 1 show that listeners' speech recognition performance is affected by rhythm alteration to the target and background talkers in the current selective listening paradigm, the single level of rhythm alteration (i.e., a single value of $m = 0.5$) does not allow for an assessment of the effect of different amounts of rhythm alteration on entrainment and selective listening. To address this issue, a second experiment was conducted to assess how the effect of rhythm alteration changes as the rhythm alteration varies over a larger range. This was done by parametrically varying either the target rhythm or the background rhythm over four levels ($m = 0.0$, 0.25, 0.5, and 0.75) while keeping the background or the target rhythm unaltered, respectively.

Experiment 2

Methods

Participants and design Twenty native speakers of American English ($n = 16$, female), aged 20–30 years ($M = 22.6$, $SD = 2.6$) completed the experiment in return for monetary compensation. All participants were screened for normal hearing (PTA < 20 dB HL) and varied in number of years of formal music training (0–14 years, $M = 4.0$, $SD = 4.6$). The experiment implemented a 2 (stimulus condition: target vs. background) × 4 (rhythm alteration: 0.0, 0.25, 0.5, 0.75) mixed-factorial design. Participants were randomly assigned to one of two conditions ($n = 10$ in each) that held either the rhythm of the target or background constant ($m = 0.0$), while varying the rhythm of the other. Within each stimulus condition, participants heard the four levels of rhythm alteration ($m = 0.0$, 0.25, 0.5, 0.75).

Procedure On each trial, participants listened to CRM sentences and reported the Color and Number of the target sentence (cued by the Call sign "Baron"). The target sentence was presented amidst a two-talker background, with the same 50-ms sentence-onset asynchrony as in Experiment 1, and the target sentence always in the middle position. Each participant completed eight blocks of trials with 40 trials per block for a total of 320 trials. The rhythm alteration was held constant within a block and the order of rhythm alternation conditions was counterbalanced across participants in order to minimize the potential for practice effects. Following completion of the CRM task, participants completed a short survey that included questions about musical background and training. The entire experiment lasted ~1 h.

Results and discussion

Figure 3 shows the proportion correct in identifying both Color and Number for the two listener groups with either the target rhythm altered or the background rhythm altered, respectively. Consistent with the selective-entrainment hypothesis, increasing rhythm alteration had opposite effects depending on whether the rhythm manipulation was applied to the target talker or to the background talkers. Both manipulations resulted in significant linear trends. Altering the rhythm of the target talker, while keeping the background rhythm unaltered (intact), reduced listeners ability to report the correct Color and Number of the target talker, $F(1,9) = 51.49$, $p < 0.001$, $\eta^2 = 0.85$. Conversely, altering background rhythm, while keeping the target rhythm unaltered (intact), improved listeners ability to report the correct Color and Number of the target talker, $F(1,9) = 17.52$, $p = 0.002$, $\eta^2 = 0.66$. The slopes for the two conditions were very similar, but with opposite signs (target rhythm effect, $b = -0.27$; background rhythm effect, $b = 0.25$). Overall performance for the unaltered ($m = 0.0$) condition did not differ between the two listener groups (target rhythm alteration group: $M = 0.49$, $SD = 0.13$; background rhythm alteration group: $M = 0.46$, $SD = 0.20$), $t(18) = 0.42$, $p = 0.68$, 95% CI [-0.19,0.13].

Two questions that emerge from these results are (1) to what extent (when listeners make errors) do they report Colors and Numbers that are spoken by one of the two background talkers and (2) how does the proportion of errors that are intrusions from the background change with increasing rhythm alteration of the target and the background. Similar to the analyses of proportion correct recognition of Color and Number, analyses of error responses (intrusions from the background talkers) as a function of rhythm alteration of the target and background revealed two linear trends. Consistent with the selective-entrainment hypothesis, Fig. 4 shows that the tendency to make Color and Number errors that were intrusions from one of the background talkers (but misattributed to the target) increased with increasing amounts

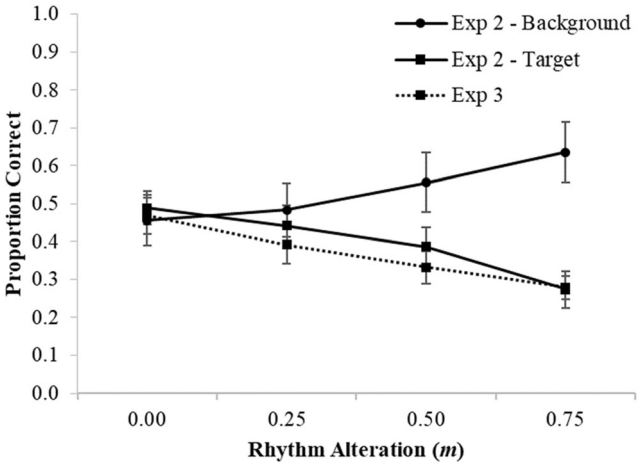

Fig. 3 Mean proportion correct target Color and word responses for the four levels of rhythm alteration (m = 0.0, 0.25, 0.50, 0.75) in Experiment 2 applied to either target speech (squares) or background speech (circles); error bars correspond to standard error. Also shown (dotted line) is the effect of target rhythm alteration with speech-shaped noise in the background (Experiment 3) Increasing alteration of the target talker rhythm, while keeping the background rhythm unaltered (intact), similarly reduces performance for both multi-talker background babble and speech-shaped noise. Increasing alteration of the background rhythm, while keeping the target rhythm unaltered (intact), improves performance

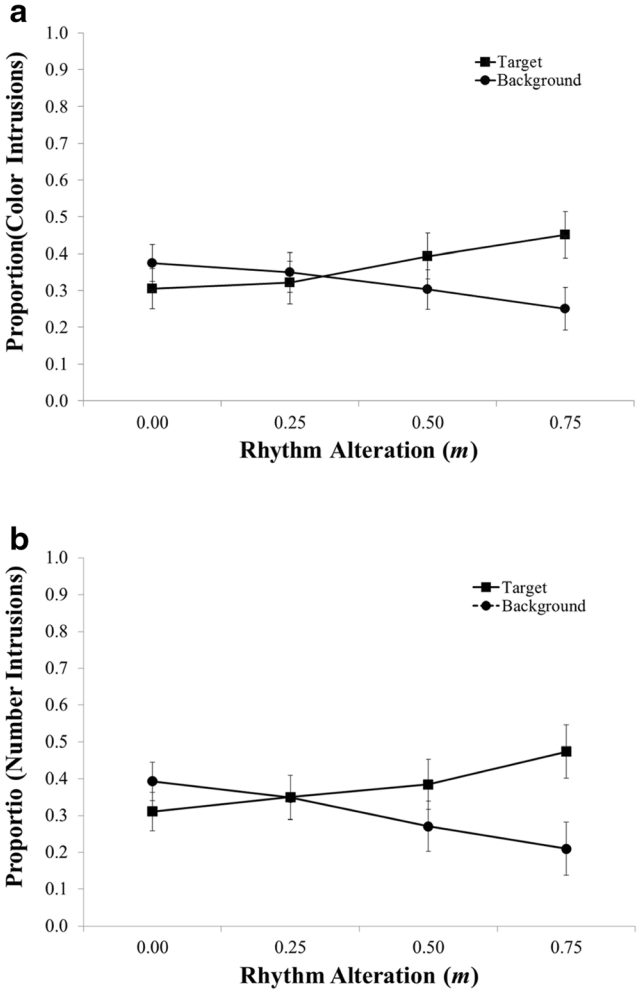

Fig. 4 Mean proportions of Color errors (**Panel A**) and Number errors (**Panel B**) in Experiment 2 that were Colors and Numbers, respectively, spoken by the background talkers, but misattributed to the target (i.e., intrusions). The two lines in each panel are a function of the four levels of rhythm alteration. Squares represent manipulations of the target rhythm and circles represent manipulation of the background rhythm; error bars correspond to standard error

of alteration of the target rhythms (Color intrusions, $F(1,9) = 26.79$, $p = 0.001$, η^2=0.75; Number intrusions, $F(1,9) = 17.03$, $p = 0.003$; η^2=0.65, but *decreased* with increasing amounts of alteration of the background rhythms (Color intrusions, $F(1,9) = 13.35$, $p = 0.005$, η^2=0.60; Number intrusions, $F(1,9) = 17.86$, $p = 0.002$, η^2=0.67). As with the analysis of correct responses, the slopes for the analyses of errors were similar, but with opposite signs (target rhythm effect: Color, $b = 0.20$, Number, $b = 0.21$; background rhythm effect: Color, $b = -0.17$, Number, $b = -0.25$).

Overall, the results of Experiment 2 show a pattern of results that favor the selective-entrainment hypothesis over both the disparity-based segregation and increased salience hypotheses. Increasing alteration of the target rhythm, predicted to weaken entrainment to the target talker, both reduces correct Color and Number reports and increases Color and Number intrusions from the background. Conversely, increasing alteration of the background rhythm, predicted to reduce interference from the background by decreasing the likelihood of inadvertent entrainment (and thereby attention) to the background, both improves correct Color and Number reports and decreases Color and Number intrusions from the background.

To further clarify the role of the background rhythm in selective attention to the target, a third experiment was conducted in which listeners performed the same speech recognition experiment using CRM sentences, but with the target sentence embedded in speech-shaped noise, rather than multi-talker babble. The speech-shaped noise was presented at a SNR that matched performance with the unaltered target sentences presented in two-talker babble in Experiment 2. Of interest was performance for alterations in the target talker rhythm presented with speech-shaped noise in the background compared to the two-talker background condition in Experiment 2. Because the rhythmic alteration had a substantial effect on performance in the presence of competing talkers, but little or no effect when sentences were presented in isolation, the question arises as to whether the reduced recognition of the target speech with increased alteration of the target rhythm is dependent on having multi-talker babble in the background. That is, the target-rhythm effect may be dependent upon interference from the background speech to such a degree that the effect is not present (or greatly diminished) when listening is made more challenging by the presence of a non-speech background sound. If so, then Experiment 3 should reveal no target rhythm effect when the target is embedded in speech-shaped noise. However, if the

target rhythm effect is present whenever difficult listening conditions make it more difficult to follow the temporal patterns of speech and predict the timing of speech events, then the same target rhythm effect should be found in Experiment 3 as in Experiment 2: Weaker selective entrainment to the target speech should result in poorer performance as the target rhythm is made more variable, even without the competition from competing speech.

Experiment 3

Methods

Participants and design Eleven native speakers of American English (n = 10, female), aged 18–22 years (M = 19.0, SD = 1.2) completed the experiment in return for monetary compensation or course credit. All participants were screened for normal hearing (PTA ≤ 20 dB HL) and varied in number of years of formal music training (0–9 years, M = 3.6, SD = 3.2). Participants heard four levels of rhythm alteration applied to the target talker (m = 0.0, 0.25, 0.5, 0.75) while the background was speech-shaped noise. The SNR of the speech-shaped noise was set to -6 dB, based on a series of pilot studies, which were used to determine the SNR that matched performance for the intact rhythm conditions for the two-talker background in Experiment 2. For these pilot studies, we tested small numbers of participants (n = 3–4) on the intact target presented in speech-shaped noise for a range of SNR values and then used the resulting performance curve to estimate the SNR predicted to yield a performance score (~47%) that matched performance for the corresponding intact conditions in Experiment 2.

Procedure On each trial, participants listened to CRM sentences and reported the Color and Number of the target sentence. Each participant completed eight blocks of trials with 40 trials per block for a total of 320 trials. The level of rhythm alteration was held constant within a block and randomized between blocks, such that each level of rhythm alteration was presented in two blocks (80 total trials). Following completion of the CRM task, participants completed a short survey that included questions about musical background and training. The entire experiment lasted ~1 h.

Results and discussion

Figure 3 (dashed line) shows proportion correct in identifying both Color and Number as a function of target rhythm alteration for the target talker embedded within speech-shaped noise with a -6 dB SNR, in comparison to the findings from Experiment 2 in which a two-talker background was used with SNR = 0. Similar to Experiment 2, there was a robust linear drop in performance with increasing target rhythm alteration, $F(1,10) = 45.6$, $p < 0.001$, $\eta^2 = 0.85$. Moreover, visual inspection of the data from the two experiments reveals two overlapping lines with almost identical slopes (Experiment 2, target rhythm effect, b = -0.27; Experiment 3, target rhythm effect, b = -0.25). Thus, when the difficulty of recognition of the unaltered target speech is equated by adjusting the SNR for the two types of background sounds, the detrimental effect of altering the target rhythm is the same for both backgrounds. These results provide support for the view that the target-talker rhythm effect is due to less effective entrainment to the target speech rather than attention being drawn to (or entrained by) the more natural rhythm of the background speech.

General discussion

Three experiments investigated listeners' use of rhythm to selectively attend to a target speech stream in the presence of competing speech or speech-shaped noise, contrasting a selective entrainment hypothesis with a disparity-based segregation hypothesis and an increased salience hypothesis. All three experiments used the Coordinate Response Measure (CRM) paradigm in which listeners report the Color and Number spoken by a target talker, presented amidst multi-talker babble (Experiments 1 and 2) or speech-shaped noise (Experiment 3). Consistent with the selective entrainment hypothesis, Experiments 1 and 2 revealed that rhythmic alteration of the target speech leads to poorer recognition of the words in the target sentences (a target rhythm effect), while rhythmic alteration of the competing background speech results in less interference with recognition of the rhythmically-intact target speech (a background rhythm effect). Providing further support for the selective entrainment hypothesis, error analyses showed that the proportion of intrusion errors (misreporting Colors and Numbers that were present in the background, but not the target) increased with rhythm alterations of the target, but decreased with rhythm alterations of the background.

While these findings provide support for the selective entrainment hypothesis, the pattern of results is not consistent with either the disparity-based segregation hypothesis or the increased salience hypothesis, which both incorrectly predict that altering the speech rhythm of the target should improve (not worsen) performance by making the target speech rhythm more distinct from the competing speech. The increased salience hypothesis further incorrectly predicts that recognition of the target speech should be worse when the speech rhythm of the background is altered; we found the opposite. Thus, while rhythmic differences can facilitate segregation of co-occurring speech streams, attentional focus on a particular stream is dependent on the presence of predictable speech rhythms that provide a basis for attentional entrainment.

Listening to a spoken message is more difficult when the natural speech rhythm of the target is disrupted, even when that disruption results in an increase in the to-be-attended target's rhythmic distinctiveness from competing speech patterns.

Experiment 3 extended the present results by examining the influence of the type of background sounds on the target rhythm effect. Here, speech-shaped noise was used as the background, rather than multi-talker babble, so that the effect of different amounts of target rhythm alteration (i.e., the target rhythm effect) could be examined without the presence of competing speech rhythms, but for listening environments that were more difficult than presenting the target in isolation (i.e., in quiet). For this experiment, the SNR for the speech-shaped noise had to be set 6 dB lower than for the two-talker babble in order to roughly equate performance for the unaltered (intact) target rhythm across the two background conditions. The fact that a lower SNR ratio was required to equate performance in the speech-shaped noise and two-talker background conditions (-6 dB vs. 0 dB) indicates that the rhythmic or linguistic content in the background speech produces more interference than steady-state noise at the same SNR. Nonetheless, increasing amounts of rhythm alteration to the target speech degraded recognition of the target speech in a graded manner that was nearly identical to the two-talker background condition in Experiment 2 across the different levels of rhythm alteration (m = 0.25, 0.50, and 0.75).

These results indicate that the target rhythm effect is not simply due to a tendency to entrain to the more regular (natural) speech rhythms in the presence of competing speech. Rather, entrainment aids selective listening to speech in difficult listening situations regardless of the type of background interference – at least for the speech-shaped noise and multi-talker backgrounds investigated here.

One potential alternative account of the target rhythm effect that warrants consideration is based on the observation that there may be a more stable auditory template for identifying the word "Baron" in the intact (unaltered) target rhythm condition than in the altered target rhythm conditions. Although a possibility, which we cannot rule out entirely to account for at least part of the target rhythm effect, we think that an auditory template matching strategy for explaining the target rhythm effect is unlikely for several reasons. First, the target is the same male talker in all conditions – so listeners have the target voice to use to identify the target talker in addition to the code name "Baron." Second, the Call sign Baron is the only Call sign that begins with B and thus it seems likely that Baron is readily discriminable form the other Call signs for all levels of the target rhythm manipulations. Moreover, none of the participants reporting difficulty identifying the target talker or Baron, and our subjective impression is that identifying the speaker that says "Baron" is not where the difficulty arises in performance of the target rhythm altered conditions, but rather occurs upstream in predicting the timing of the Color and Number. Finally, Experiment 2 further shows that the target rhythm effect is graded, with increasingly worse performance with increasing alteration of the target rhythm. Given that the selective entrainment hypothesis explains both the target rhythm and background rhythm effects, which are in opposite directions, we find this account to be the more parsimonious explanation.

An additional aspect of the background rhythm effect that warrants discussion is why the magnitude of the background rhythm effect in Experiment 1 was found to be larger with two talkers in the background compared with six. There are several possible reasons why this may have been the case. First, acoustically, the two-talker background exhibits a greater degree of envelope fluctuations and a more evident peak in its modulation spectrum near the syllabic rate than the six-talker background (Humes et al., 2017). Imposing rhythm alteration to a two-talker background reduces the total modulation power and broadens the modulation spectrum, causing a greater difference between the envelopes of the target and background. On the other hand, the six-talker background exhibits a broad modulation spectrum, which is already distinct from that of the target speech even before applying rhythm alteration. Therefore, the same degree of rhythm alteration may have a larger effect for the two-talker than six-talker background. Second, as the number of background talkers increases from two to six, individual background talkers become less likely to be perceived as perceptually segregated sound sources (e.g., Kashino & Hirahara, 1996; Zhong & Yost, 2017). Since the rhythm alteration was imposed on individual background talkers, its effect may be stronger when each of the talkers is perceptually distinct from other talkers. Third, the SNR in the current study is expressed relative to the overall level of the multi-talker background. This means that the target levels relative to each individual background sentence were 3.01 and 5.78 dB for the two- and six-talker backgrounds, respectively. Since a higher target intensity may draw greater attention to the target (e.g., Richards et al., 2013), the reduced effect of background rhythm alteration observed for the six-talker background may be caused by the higher relative target level in that condition.

An examination of response strategies reported by participants in a survey given at the end of each experiment revealed that while some subjects (~10%) commented on following the timing or rhythm of the sentences, there were several different strategies, with "closing the eyes" being the most common reported strategy by far. There was no evidence of a shift in strategies across the different listening conditions in the three experiments in the current study. Our examination of response strategies further revealed, however, that no particular strategy appeared to be associated with either better or worse performance on the CRM task, or the tendency for participants to show an effect of rhythm alteration. Indeed, the consistency of

the pattern of results across participants was particularly striking with most participants showing an effect of rhythm alteration in line with the selective-entrainment hypothesis, regardless of any strategy they may have reported using in the task. Thus, it appears that most listeners are either unaware of using rhythm as a listening strategy, or it is so second nature, that they don't consider it to be a "strategy" for performing the task, any more than "listening carefully" would be considered a strategy.

The current findings in support the selective entrainment hypothesis add to a growing body of evidence demonstrating the importance of talker rhythm in understanding speech in difficult listening situations. Behavioral and neurophysiological evidence (cited in the introduction) has shown behavioral and neural entrainment to speech rhythms plays an important role in speech perception. Much of this work suggests that entrainment to speech rhythms is not a simple, passive, stimulus-driven entrainment process, but rather involves anticipation and active predictions of future events and incorporates hierarchical levels of temporal structure. Along these lines, a number of studies have shown that neural oscillations can be selectively entrained by the amplitude envelope of the target speech in multi-talker listening situations (Ding & Simon, 2012; Golumbic et al., 2013; Horton et al., 2013; Rimmele et al., 2015).

Studies similar to the current one have examined the effects of different types of rhythm manipulations in an attempt to understand how listeners use temporal structure to guide selective listening. For example, Wang et al. (2018) found that natural rhythmic speech was recognized better than speech that was rhythmically altered (by mixing portions of fast, normal, and slow speech within a sentence) in a two-talker babble background. They also found that the advantage of rhythmically intact speech was more pronounced for words later in the sentence, suggesting that attentional entrainment improves over time throughout a sentence. Aubanel et al. (2016) found that altering natural speech rhythm harms speech intelligibility in speech-shaped noise, even when speech is made artificially isochronous (although some isochronous schemes, such as ones based on vowel onset rather than envelope peaks, were less disruptive than others). These findings indicate that the entrainment underlying speech perception is an active process of tracking and predicting quasi-periodic speech rhythms, rather than a more passive response to fluctuations in the amplitude envelope.

The lack of support for the disparity-based segregation or the increased salience hypothesis does not mean that entrainment is the only factor involved in the segregation of spoken sentences. Strong pitch or spatial differences between talkers may lead to good segregation (and selective listening) even when temporal irregularities are introduced. This may mean that some stimulus differences, like pitch and spatial location, are more primitive or obligatory, while segregation based on rhythmic differences involves higher-level mechanisms (see Bregman's (1990) discussion of primitive vs. schema-based segregation). However, the influence of supposedly more primitive cues, like pitch, on segregation is dependent on temporal relations (see Bregman, 1990; Jones et al., 1981) and primitive or obligatory segregation may reflect the limits of attentional mechanisms, rather than a more peripheral mechanism. Further research examining the types of rhythmic differences that can be utilized for selective listening when different pitch or spatial (or other) cues are present may help to clarify how listeners use spectral-temporal structure to guide selective listening, and in doing so, it may help to clarify the relation between stream segregation and selective listening.

In summary, there are two main findings that emerge from the present set of experiments. First, the experiments reveal a *target rhythm effect* in understanding speech in difficult listening conditions. Alteration of the natural speech rhythm of a to-be-attended target utterance that is embedded in either a multi-talker background, or speech-shaped noise, *degrades* recognition of target speech. The target rhythm effect does not appear, however, when the to-be-attended target appears in quiet listening conditions (i.e., in isolation). Thus, natural speech rhythms, with their quasi-periodic structure, appear to take on an increasingly important role in speech understanding in more challenging listening environments, independent of the type of interfering background sounds.

Second, the experiments are the first (as far as we are aware) to reveal a *background rhythm effect* whereby alteration of the natural speech rhythms of interfering background speech *improves* recognition of target speech. However, the rhythm sensitivity demonstrated in the current study suggests that other studies of speech-on-speech masking that have shown better performance when background speech consists of a different language or has a different "accent" than the target speech (e.g., Calandruccio et al., 2010, 2014; Van Engen & Bradlow, 2007), may reflect (at least in part) a closely related background-rhythm effect based on the difference between target and background speech rhythms associated with different languages and accents. One outstanding question that warrants future investigation is whether the background rhythm effect reported here may depend on the linguistic content of the background. Some support for this possibility comes from the work of Peele and colleagues who have provided evidence that linguistic content strengthens neural entrainment (Peelle, Gross, & Davis, 2013).

From dynamic attending theory (DAT) perspective, the alterations of the background speech rhythm reduce the likelihood that listeners' attention to the background speech will be inadvertently entrained by the background speech (as greater irregularity in the background speech rhythms supports weaker entrainment than the more natural speech rhythms of the target). This interpretation is supported by the observed decrease in Color and Number intrusion errors with increased

rhythm alteration of the background speech. Taken together, the DAT-based selective entrainment hypothesis provides a unified account of both the target rhythm effect and the background rhythm effect, highlighting the importance of speech rhythm in understanding speech in difficult listening conditions.

Acknowledgements The authors thank Audrey Drotos, Anusha Mamidipaka, and Paul Clancy for their assistance with data collection and their insights and many helpful comments over the course of the project, Dylan V. Pearson at Indiana University for assisting 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.

Open Practices Statement The data and materials for all experiments will be made available at http://taplab.psy.msu.edu. None of the experiments were pre-registered.

References

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.

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*, 571-589.

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

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*, 1065-1066.

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

Bregman, A. S., Abramson, J., Doehring, P., & Darwin, C. J. (1985). Spectral integration based on common amplitude modulation. *Perception & Psychophysics, 37*, 483-493.

Calandruccio, L., Dhar, S., & Bradlow, A.R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America, 128,* 860-869.

Calandruccio, L., Bradlow, A.R., & Dhar, S. (2014). Speech-on-speech masking with variable access to the linguistic content of the masker speech for native and non-native speakers of English. *Journal of the American Academy of Audiology, 25,* 355-366.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*, 975-979.

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.

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

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*, 11854-11859.

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

Eddins, D. A., & Liu, C. (2012). Psychometric properties of the coordinate response measure corpus with various types of background interference. *The Journal of the Acoustical Society of America, 131,* EL177-EL183.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America, 115*, 2246-2256.

Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology, 2*, 130.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience, 15*, 511.

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.

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.

Horton, C., D'Zmura, M., & Srinivasan, R. (2013). Suppression of competing speech through entrainment of cortical oscillations. *Journal of Neurophysiology, 109*, 3082-3093.

Huggins, A. W. (1972). On the perception of temporal phenomena in speech. *Journal of the Acoustical Society of America, 51*, 1279-90.

Humes, L. E., Kidd, G. R., & Fogerty, D. (2017). Exploring use of the coordinate response measure in a multitalker babble paradigm. *Journal of Speech, Language, and Hearing Research, 60*, 741-754.

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

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review, 96*, 459-491.

Jones, M. R., Kidd, G., & Wetzel, R. (1981). Evidence for rhythmic attention. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 1059-1073

Jones, M.R, Moynihan, H., MacKenzie, N., & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science, 13,* 313-319.

Kashino, M., & Hirahara, T. (1996). One, two, many—Judging the number of concurrent talkers. *The Journal of the Acoustical Society of America, 99*, 2596-2603.

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

Kidd, G. R., & Humes, L. E. (2014). Tempo-based segregation of spoken sentences. *The Journal of the Acoustical Society of America, 136*, 2311-2311.

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

Marin, C., & McAdams, S. (1991). Segregation of concurrent sounds: II. effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width. *Journal of the Acoustical Society of America, 89*, 341-351.

McAdams, S. (1989). Segregation of concurrent sounds, I: Effects of frequency modulation coherence. *Journal of the Acoustical Society of America, 86*, 2148-2159.

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.

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.

Middlebrooks, J. C., Simon, J. Z., Popper, A. N., & Fay, R. R. (Eds.). (2017). *The auditory system at the cocktail party* (Vol. 60). Cham, Switzerland: Springer.

Miller, J., Carlson, L., & 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,* 11-18.

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.

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

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

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*, 1378-1387.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time.' *Speech Communication. 41,* 245-255.

Richards, V. M., Shen, Y., & Chubb, C. (2013). Level dominance for the detection of changes in level distribution in sound streams. *The Journal of the Acoustical Society of America, 134,* EL237-EL243.

Riecke, L., Formisano, E., Sorger, B., Baskent, D., & Gaudrain, E. (2018). Neural entrainment to speech modulates speech intelligibility. *Current Biology, 28,* 161-169.

Rimmele, J. M., Golumbic, E. Z., Schröger, E., & Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex, 68*, 144-154.

Rimmele, J. M., Jolsvai, H., & Sussman, E. (2011). Auditory target detection is affected by implicit temporal and spatial expectations. *Journal of Cognitive Neuroscience, 23*, 1136-1147.

Rimmele, J. M., Schröger, E., & Bendixen, A. (2012). Age-related changes in the use of regular patterns for auditory scene analysis. *Hearing Research, 289*, 98-107.

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,* 2431-2443.

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.

Smith, M.R., Cutler, A., Butterfield, S., and Nimmo-Smith, I. (1989). Perception of rhythm and word boundaries in noise-masked speech. *Journal of Speech, Language and Hearing Research, 32*, 912-920.

Snyder, J. S., Gregg, M. K., Weintraub, D. M., & Alain, C. (2012). Attention, awareness, and the perception of auditory scenes. *Frontiers in Psychology, 3*, 17.

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*, 628-639.

Van Engen, K.J., & Bradlow, A.R. (2007). Sentence recognition in native- and foreighn-language multi-talker background noise. *The Journal of the Acoustical Society of America, 121,* 519-526.

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.

Zeng, F., Nie, K., Stickney, G. S., Kong, Y., Vongphoe, M., Bhargave, A., … Cao, K. (2005). Speech recognition with amplitude and frequency modulations. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 2293-2298.

Zhong, X., & Yost, W. A. (2017). How many images are in an auditory scene? *The Journal of the Acoustical Society of America, 141*, 2882-2892.