EXPLORING THE EFFECT OF RELATIVE TIMING OF TARGET AND BACKGROUND WORDS ON SPEECH UNDERSTANDING WITH AND WITHOUT A BACKGROUND RHYTHMIC CONTEXT

By

Toni Marie Smith

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Psychology—Master of Arts

2021

ABSTRACT

EXPLORING THE EFFECT OF RELATIVE TIMING OF TARGET AND BACKGROUND WORDS ON SPEECH UNDERSTANDING WITH AND WITHOUT A BACKGROUND RHYTHMIC CONTEXT

By

Toni Marie Smith

Using the Coordinate Response Measure (CRM) paradigm, recognition of target speech in the presence of competing speech has been shown to depend upon both the rhythmic context of target and background speech and fundamental frequency differences between the speakers (McAuley et al., 2021). In the present study, two experiments examined the effects of relative timing of target and background key words and the presence or absence of a background rhythmic context on target recognition using the same male talker for both target and background sentences. Exp. 1 varied the onset asynchrony between target and background key words when background rhythmic context was removed (i.e., the background consisted only of the competing key words) and Exp. 2 manipulated the rhythm of background speech leading up to key words, but left the key words intact with an onset asynchrony of ±50ms. Exp. 1 revealed an asymmetric U-shaped performance curve where (1) target recognition improved with increasing deviation of background key words from the expected onset timing of target keywords, and (2) target key words were better recognized when they began prior to the onset of background key words, compared to after. With the reintroduction of the background context in Exp. 2, performance was reduced to chance both when the background rhythm was intact and when it was rhythmically irregular, suggesting that listeners were unable to distinguish target and background sentences and could not develop expectations for target keyword timing

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

Understanding speech-in-noise is crucial for effective communication within the hearing population, since many of our social interactions occur in noisy environments such as family dinner tables, restaurants, busy street sides, or (more recently) lag-y video conference calls. Despite the importance of this ability in navigating the social world, the myriad of factors that underlie speech-in-noise perception are not well understood. There are a number of stimulus factors that have been shown to influence the recognition of speech-in-noise, including the type of background sounds (e.g., speech vs. non-speech) (Desjardins & Doherty, 2013), the number of background talkers for speech backgrounds (Rosen et al., 2013) and general cues to perceptual segregation, such as fundamental frequency differences between target and background talkers (Brokx & Nootboom, 1982; Assmann & Summerfield, 1989, 1990). A growing body of recent investigations has also implicated speech rhythm as having an influence on speech recognition in noise (e.g. Aubanel, Davis, & Kim, 2016; Wang et al., 2018; McAuley, Shen, Dec, & Kidd, 2020; McAuley, Shen, Smith & Kidd, 2021). The role of rhythm and timing in speech-in-noise perception is the focus of this thesis.

One mechanism by which speech rhythm has been hypothesized to contribute to successful perception of speech amidst noise is through Selective Entrainment, whereby temporal regularities in to-be-attended target speech guide attention in a manner that facilitates selective attention to the target (McAuley et al., 2020). The *Selective Entrainment Hypothesis* is based in Dynamic Attending Theory (DAT), which posits that there exist internal attentional oscillations that are entrained by external rhythms (Jones, 1976; Jones & Boltz, 1989, Large & Jones, 1999; McAuley et al., 2006). This entrainment leads to changes in the phase and period of internal (attentional) rhythms that align peaks in attentional energy to time points where relevant stimulus events are expected to occur. In turn, this alignment of attentional peaks with stimulus events is hypothesized to facilitate perception. Accordingly, behavioral evidence has shown a perceptual advantage for stimuli arriving at expected (compared to unexpected) times (Barnes & Jones, 2000; Jones et al., 2002; Miller, Carlson, & McAuley, 2013).

While the timing of natural speech is not strictly periodic, it is characterized by temporal patterns that give rise to the perception of regularity, which can be used to guide temporal expectations for the occurrence of upcoming speech sounds in the speech stream. Timing at the level of syllables in particular (about 3-9 Hz) has been shown to contribute to the quasi-rhythmic nature of speech across a number of languages (Dauer, 1983; Tilsen & Arvaniti, 2013), and is also likely to be important for early language acquisition (Goswami, 2019). Additionally, a body of behavioral evidence suggests not only that these regularities are present, but also that they facilitate the perception of speech. For instance, understanding speech in noise is adversely affected by the disruption of speech rhythm. Speech that has been isochronously re-timed is more intelligible in noise than speech that has been anisochronously re-timed (Aubanel, Davis, & Kim, 2016). Regularity in speech also appears to build temporal expectations over time; within the same sentence, later-occurring words are better recognized than earlier-occurring words in multi-talker babble, but not when the speech has been altered to be artificially irregular (Wang et al., 2018). Earlier rhythmic context within a sentence can also influence the perception of ambiguous syllable organization occurring later in the sentence, suggesting that such temporal expectations can influence word segmentation (Dilley & McAuley, 2008; Morrill et al., 2014; Baese-Berk et al., 2019).

Parallel to these behavioral investigations into the role of rhythm in speech perception, evidence has also accrued that speech rhythms serve to entrain neural activity. Specifically, cortical neural oscillations have been shown to phase-lock 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 (Ghitza, 2011; Giraud & Poeppel, 2012; Ding et al., 2016; Riecke et al., 2018). Such neural synchronization is also modulated by attention; in a multi-talker context, selective attention to target speech enhances neural entrainment to the target speech envelope (Ding & Simon, 2012, 2014; Golumbic et al., 2013). Moreover, disruption of neural synchronization to target speech using trans-cranial alternating current stimulation modulates target speech recognition, suggesting that the entrainment of neural activity to the speech envelope plays a causal role in the understanding of target speech presented with a competing talker (Riecke et al., 2018).

A recent investigation by McAuley and colleagues (2020) found that alterations to the natural rhythm of both target and background speech influences target speech recognition, but in opposite ways. Their experiments used the Coordinate Response Measure (CRM) paradigm, where speech stimuli all have the same form: "Ready [Call Sign] go to [Color] [Number] now" (Bolia et al., 2000). Each sentence has one of eight Call Signs (e.g. "Baron," "Charlie," "Eagle"), one of four Colors ("Red," "Blue," "White," or "Green") and one of eight numbers (1-8, excluding "seven" in order to maintain a constant number of syllables from trial to trial). Participants are told to attend to the target sentence, which always contains the call sign "Baron," and report the Color and Number that appear within that sentence. When the target sentence is presented amidst other sentences, the Call signs, Colors, and Numbers that appear in the background sentences are always different than that of the target. The natural rhythm of target speech and the natural rhythm of background speech were independently altered to make the speech increasingly rhythmically irregular.

The authors found that increasing alterations to the natural rhythm of target speech led to *poorer* recognition of target Color and Number (a target rhythm effect). Conversely, increasing alteration of background speech rhythm led to *better* target Color and Number recognition (a background rhythm affect) as well as a reduction in intrusion errors (responses coming from the background) (McAuley, Shen, Dec, & Kidd, 2020). These results are consistent with the *Selective Entrainment Hypothesis,* which predicts both the target and background rhythm effects. If selective entrainment to target speech plays a role in the perception of target speech in noise, disruption of the natural rhythm of target speech should impede target speech recognition. Without a stable speech rhythm to guide temporal expectations, there is predicted to be a misalignment of attentional focus and information-carrying events in the target speech. In contrast, a disruption of the natural rhythm of background speech should enhance target recognition. This is hypothesized to occur because of a reduction of competing entrainment to the background rhythm, thereby strengthening entrainment by the target speech rhythm. In essence, without a regular background rhythm it is less likely that attention would be accidentally entrained by the background rhythm at the expense of entrainment to the target. It would also reduce the likelihood of intrusions from the background due to inadvertent entrainment to the background speech rhythm.

In a set of follow-up studies using backgrounds that varied in their similarity to the target, the target rhythm effect was found to be robust to disparity-based segregation (McAuley, Shen, Smith, & Kidd, 2021). The background rhythm effect, however, was only observed with a background talker of the same sex as the target. When the background was of the opposite sex, or when it contained amplitude envelope information but was removed of semantic content or temporal fine structure, the background rhythm effect was absent. This suggests that the background rhythm effect is not driven by amplitude envelope-based rhythm alone and may be reduced when there are strong cues for perceptual segregation or when the background is unintelligible (McAuley, Shen, Smith, & Kidd, 2021).

Although McAuley and colleagues provide evidence for both target and background rhythm effects that are in line with the Selective Entrainment Hypothesis, the work raises some outstanding questions. A key assumption in prior experiments is that the rhythmic context leading up to the Color and Number words is what guides expectations for the timing of those words; moreover, it is the disruption of that rhythmic context (and, in effect, disruption of temporal expectations) that drives changes in performance. However, McAuley and colleagues applied the rhythm alteration to the entirety of the sentences- not just the rhythmic context leading up to Color and Number alone. As a result, the rhythm alteration additionally alters the relative timing of the onsets of background and target Color and Number words. To account for this, McAuley and colleagues applied the rhythm alteration using a range of different phases, so the relative timing of background and target Color and Number words varied from trial to trial – thus averaging out any systematic effect of relative onset timing. Nonetheless, it is possible that separate from an effect of rhythmic context (entrainment), onset asynchrony between background and target Color and Number words differentially affects target recognition.

Support for a role of onset asynchrony between to-be-attended and to-be-ignored speech material comes from a number of sources. First, asynchrony of onsets is a strong cue for segregation in both speech and non-speech sounds (Bregman, 1990). Second, performance on the CRM paradigm used by McAuley et al. (2020, 2021) has been shown to be better for sentences presented asynchronously than for sentences presented synchronously (Humes, Kidd, & Fogerty, 2017). The latter finding, however, relates to the asynchrony of sentence onsets rather than for the onsets of the Color/Number words within the sentences that participants are specifically listening for. Other research on the effects of onset asynchrony has tended to focus on the fusion of tones or individual speech sounds, rather than whole words or phrases embedded within sentences. For example, onset-based grouping effects on the number of sounds heard have been observed when vowel formants that have a common or uncommon onset time are presented (Darwin, 1981). Onset asynchronies can also prevent a mistuned frequency component from influencing the perceived pitch of a harmonic complex (Darwin & Ciocca, 1992), and can abolish the fusion of short noise bursts across frequency and location (Turgeon, Bregman, & Roberts, 2005).

To directly address the issue of onset asynchrony in the present investigation, Experiment 1 used a modified version of the CRM paradigm to examine how differing amounts of onset asynchrony between target and background Color and Number words impacts target recognition in the absence of background rhythmic context. Here, listeners were presented with a single target sentence and an isolated set of background Color/Number words spoken by the same talker. The onset of background Color and Number relative to the target Color and Number was varied. In this scenario, it is expected that the listener will form temporal expectations for the onset of the target Color/Number word pair based on the natural spoken rhythm of the target sentence. If this is the case, correct Color/Number recognition should improve with increasing asynchrony of Color/Number onsets, regardless of the direction of asynchrony. That is, according to a DAT-based *Temporal Expectation Hypothesis*, performance would be worst when target and background Color and Number are synchronous, because in this condition the timing of both the target and background Color/Number pairs will be consistent with the temporal expectations set by the target speech rhythm and so there will be less information to distinguish between the two. As onset asynchronies become larger, it is expected that target speech recognition will improve because the to-be-ignored background color and number will not align with the temporal expectations established by the target speech rhythm.

An alternative possibility is that there may simply be a bias to attend to whichever Color/Number pair appears first, regardless of temporal expectations (i.e. a *Temporal Order Hypothesis*). Distractors that are early or late with respect to a target stimulus have been shown to interfere differently with task performance. For example, when synchronizing with a moving target dot that has a sinusoidal trajectory, having additional moving dots on the screen only interferes with synchronization when the distractors lead in phase (Booth & Elliot, 2015). There is also evidence from memory research that distractor-response binding effects in retrieval-based probe responding appear when the distractor occurs before the target stimulus that must be responded to, but not when the distractor occurs after the target (Frings & Moeller, 2012). Generally, it seems that early distractors are in a sense more distracting than late distractors. In the absence of an influence of temporal expectations coming from the target speech, this would produce a linear effect where the more the background leads the target, the more it interferes and the worse performance becomes; conversely, the more the background lags the target, the less it interferes and the better performance becomes.

It is also possible that the data support both the Temporal Expectation and Temporal Order hypotheses. If a temporal order bias interacts with temporal expectations, this would result in the beneficial effect of increasing onset asynchrony on performance being attenuated when the background Color/Number onset leads the target (compared to when it lags the target). A final possibility is that temporal expectations for Color/Number onset and the temporal order of target and background color and number play no role in correct target recognition, resulting in no difference in performance as a function of onset asynchrony.

A second factor of interest in the present investigation concerns the effects of the rhythm alteration on the intelligibility of the Color and Number words. McAuley et al. (2020) established that the rhythm alteration does not make individual target sentences any less intelligible when presented in isolation in quiet listening conditions (McAuley, Shen, Dec, & Kidd, 2020). However, it is possible that the rhythm alteration degrades the intelligibility of the Color and Number words in more difficult listening situations (i.e., in the presence of noise or other competing sounds). The purpose of Experiment 2 was thus to investigate the effects of rhythm alteration, while controlling for Color/Number intelligibility as well as Color/Number onset asynchrony. Toward this end, Experiment 2 focused on the background rhythm effect, namely the *improvement* in target color and number recognition found when applying the rhythm alteration to a to-be-ignored background talker. One motivation for focusing on the background rhythm effect, rather than the target rhythm effect, is that previous work has found that the background rhythm effect is notably absent when the background is unintelligible vocoded speech; this suggests that the background rhythm effect might depend in part on background speech intelligibility (McAuley, Shen, Smith, & Kidd, 2021).

In Experiment 2, the rhythm alteration was applied only to the beginning of the background sentence in order to manipulate temporal expectations, while Color and Number words remained unaltered (i.e., intact). In addition, the onset of background relative to target Color and Number was controlled across rhythm alteration conditions. If the background rhythm effect is indeed due primarily to reduced inadvertent entrainment to the background speech (and thus a facilitation of selective entrainment to the target speech), the manipulation of the background rhythmic context alone should be sufficient to produce the effect when the intelligibility and timing of background Color and Number words is held constant.

GENERAL METHODS

Speech stimuli were taken from the Coordinate Response Measure (CRM) Corpus (Bolia et al., 2000). Sentences from this corpus all have the form "Ready [Call Sign] go to [Color] [Number] now." Each sentence contains one of eight Call Signs (e.g. "Baron," "Charlie," "Eagle"), one of four Colors ("Red," "Blue," "White," or "Green") and one of eight numbers (1-8, excluding "seven" in order to maintain a constant number of syllables from trial to trial). The Call Signs, Colors, and Numbers that appear in the target and background were always different. Both target and background sentences came from the same male talker (talker #1). The target sentence always contained the Call Sign "Baron," which acted as a cue for which sentence to attend to. The target sentence was always a complete sentence, while the background consisted of just the portion beginning with the Color and Number in Experiment 1 and the complete background sentence in Experiment 2. Both target and background sentences were presented binaurally at 65 dB SPL, using Senheiser HD 280 Pro over-the-ear headphones at a sampling rate of 22050 Hz. On each trial, participants reported the Color and Number they heard in the target sentence by selecting a square on a computer screen with the corresponding combination of Color and Number, presented via a custom MATLAB program.

The study took place over two sessions. Experiment 1 and Experiment 2 were administered in separate sessions on different days. The order of the experiments was counterbalanced and randomly assigned for each participant, in order to control for carryover or practice effects from one experiment to the next. Each session lasted approximately 1.5 hours.

At the beginning of Session 1, participants were given a brief familiarization task. In one block of 32 trials, participants were presented with intact CRM target sentences in quiet (with no background) and were instructed to report the Color and Number that appeared within each sentence. This familiarization task acted as a means to screen participants for obvious task-relevant hearing difficulties or a failure to understand instructions. Additionally, the task served to acclimate participants to the procedure prior to experiencing the more difficult experimental conditions. The exclusion criterion for use of a participant’s experimental data was performance below 90% on the familiarization task. No participant’s scores fell below this level and none were excluded on this basis.

At the end of both sessions, participants completed surveys about the strategies that they used during the experiment as well as any factors that might have influenced their performance. Additionally, at the end of Session 1 participants completed a survey about their personal and musical background. At the end of Session 2 participants completed a short form of the Speech and Spatial Qualities of Hearing (SSQ) questionnaire (Noble et al., 2013) and the Noise Exposure Questionnaire (NEQ) (Johnson et al., 2017). The SSQ includes questions about one’s subjective ease of sound segregation, listening to speech in noise, and locating sounds (e.g. “You are talking with one other person and there is a TV on in the same room. Without turning the TV down, can you follow what the other person you’re talking to says?”). Participants respond using a 0 to 10 scale where 0 means “Not at all” and 10 means “Perfectly” (with the exception of two questions which use different anchors) (Noble et al., 2013). The NEQ indexes the frequency and length of exposure of individuals to both occupational and non-occupational noise (e.g. use of power tools, attending loud events such as concerts, driving loud vehicles), which can be used to calculate a measure of annual noise exposure that is indicative of risk for noise-induced hearing loss.

EXPERIMENT 1

Methods

Participants and Design. 18 participants (15 female; age range: 19-26, *M* = 21.6, *SD* = 2.0) were recruited from the Michigan State University community and were compensated at a rate of $10/h in the form of digital Amazon gift cards. All participants were native speakers of American English and had self-reported normal hearing. Onset asynchrony of CRM key words (Color and Number) was manipulated within subjects (OA = 0ms, ±25ms, ±50ms, ±100ms, ±200ms).

Stimuli. In Experiment 1, full target sentences were presented with a single partial background sentence. The beginning of each background sentence was removed and replaced with silence so that only the phrase "[Color] [Number] now" was heard. The onset of the background key word pair relative to the target key word pair was manipulated. Specifically, the onset asynchrony was defined as the timing of the onset of the background Color word relative to the onset of the target Color word.

Procedure. The experiment was conducted in a single test session of 15 experimental blocks. Each block consisted of 36 trials. Each of the nine OA conditions (±0ms, ±25ms, ±50ms, ±100ms, ±200ms) was presented 4 times per block for a total of 60 presentations over the 15 blocks. For each consecutive subset of 9 trials within a single block, each OA condition occurred once in randomized order. A mandatory break was provided about halfway through the experiment (after 8 blocks), and participants were encouraged to take breaks as needed after each block.

Results

For each participant, the proportion of correct responses (trials where both the correct Color and correct Number were reported) was calculated separately at each level of onset asynchrony (Figure 1). Consistent with the target rhythm guiding temporal expectations about the onset timing of the target Color and Number, there was a significant quadratic trend as a function of OA, $F(1, 17) = 158.09$, $p < 0.001$, $\eta^2 = 0.90$, where performance was worst for an onset asynchrony of -50ms, and improved with increasing deviations from this value in either direction.. There was additionally a significant linear trend as a function of OA, $F(1, 17) = 76.89$, $p < 0.001$, $\eta^2 = 0.82$, where performance was overall better when the background was lagging compared to when the background was leading.

Next, we considered the types of errors that were made by participants. For each participant the proportion of Color intrusions (trials where the background Color was reported instead of the target Color) (Figure 2A) and Number intrusions (trials where the background Number was reported instead of the target Number) (Figure 2B) were calculated separately at each level of onset asynchrony. There was a significant quadratic trend for Color intrusions, $F(1, 17) = 43.81$, $p < 0.001$, $\eta^2 = 0.72$), and Number intrusions, $F(1, 17) = 90.38$, $p < 0.001$, $\eta^2 = 0.84$. In each case, the pattern of results is the *opposite* that of proportion correct scores. That is to say, intrusion errors were most frequent for an onset asynchrony of -50ms, and were reduced with increasing deviations from this value in either direction.There was again a significant linear trend in the opposite direction of proportion correct scores for both Color intrusions, $F(1, 17) = 78.18$, $p < 0.001$, $\eta^2 = 0.82$, and Number intrusions, $F(1, 17) = 71.67$, $p < 0.001$, $\eta^2 = 0.81$. The background onset occurring first leads to *more* intrusions compared to when the background onset occurs second.

We additionally examined the relationship between several individual difference characteristics and performance. To get one overall measure of performance, proportion correct scores were averaged across all OA conditions individually for each participant. Pearson correlations were run between these overall scores and self-reported years of formal music training, average SSQ scores (measuring self-reported hearing abilities), and annual noise exposure (ANE). SSQ and ANE scores were unavailable for 2 participants because they did not complete the second session of the study where the corresponding surveys were administered, leaving n = 16 for those analyses. SSQ scores were calculated by averaging across responses to questions (*M* = 7.15, *SD* = 1.15). Two of the sixteen questions were excluded from the average because the response scale anchors were presented incorrectly for those two questions for the majority of participants. ANE was calculated based on the procedure outlined in Johnson et al. (2017) where the minimum possible value was 64 and higher values mean greater noise exposure (*M* = 70.94, *SD* = 3.19). Out of 18 participants, 12 reported having formal music training. Including those who did not receive any formal music training, the mean number of years of formal training for this group was 4.67 (*SD* = 4.85). No correlation was found between task performance and years of formal music training, *r* = -0.001, *p* = 0.99. There was also no relationship between performance and SSQ scores, *r* = 0.09, *p* = 0.73, or between performance and ANE, *r* = 0.19, *p* = 0.48.

Discussion

The purpose of Experiment 1 was to determine how the relative timing of background and target Color and Number (absent the preceding background rhythmic context) impacts target speech recognition. Consistent with the Temporal Expectation Hypothesis, results show that increasing asynchrony of background Color/Number onset with respect to the expected temporal onset of the target Color/Number generally leads to improved performance and a reduction of intrusion errors. This U-shaped curve, however, was slightly left-shifted such that performance was worst (i.e. the background Color/Number pair was more intrusive) when the background Color/Number onset led the target by a small amount.

Separately, there was a tendency to select whichever Color/Number begins first, thus also providing support for the Temporal Order Hypothesis. This interaction between temporal expectations and temporal order effects leads to a pattern of performance such that at larger asynchronies where the background leads, there is an improvement in target recognition attributable to a violation of temporal expectations by the background, but the improvement is attenuated by the detrimental effect of the background occurring first.

EXPERIMENT 2

Previous work has demonstrated a background rhythm effect such that increasing alteration of the natural rhythm of background speech enhances target speech recognition (McAuley et al., 2020). If selective entrainment is driving the background rhythm effect it is expected that the rhythmic context leading up to the Color and Number is what builds temporal expectations, which are unaffected by the timing of the Color and Number words themselves. Toward this end, Experiment 2 applied the rhythm alteration only to the beginning of the background sentence leading up to the Color and Number (which will be referred to as the "precursor"). This manipulation should interfere with temporal expectations, without differentially interfering with the intelligibility or timing of background key words between rhythm conditions. This will ensure that any effects of background rhythm alteration are not due to reductions in the intelligibility of the background Color and Number associated with the rhythm manipulation.

Methods

Participants and Design. 16 participants that took part in Experiment 1 (14 female; age range: 19-26, M = 21.50, SD = 1.90), also participated in Experiment 2 and were compensated for their participation at a rate of $10/h in the form of digital Amazon gift cards. Background speech rhythm alteration was manipulated within subjects (m = 0, 0.25, 0.50, 0.75), as was onset asynchrony (OA = +50ms, -50ms).

Stimuli. On each trial, target CRM sentences were presented with complete background CRM sentences spoken by the same talker (talker #1). In some conditions, the natural rhythm of background speech was disrupted. This disruption was achieved by temporally expanding and contracting the speech in a sinusoidal fashion. In order to preserve the intelligiblity of the background Color and Number, only the precursor (“Ready [Call Sign] go to”) was altered while the key words (“[Color] [Number] now.”) remained intact. 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 $\mathrm{CR}(t) = 1 + m \sin(2\pi f_m t + \phi)$ (Fig. 3). The rate of rhythm alteration, f_m, was set to 1Hz, based on McAuley et al (2020), who showed that this value preserved speech intelligibility in quiet while still providing a strong percept of timing variation. 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. The initial phase of 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 $7\pi/4$) so that different parts of each sentence were expanded or contracted. Onset asynchronies (background color word onset relative to target color word onset) were set to +50ms or -50ms with equal probability. Both target leading (+50ms) and background leading (-50ms) conditions were included and randomly varied from trial to trial so that participants could not simply distinguish between target and background Color and Number words based on which pair appeared first. This also provided a test of how the order of Color/Number onsets influences target recognition in the presence of rhythmic contexts in both the target and the background.

Procedure. The experiment was conducted in a single test session of 16 experimental blocks. Each block consisted of 32 trials with the same level of rhythm alteration. Each of the four levels of rhythm alteration occured four times total, once within each set of 4 blocks; the order of rhythm alteration levels was counterbalanced across sets. Additionally, the entire sequence of 16 blocks was presented in one of four orders which were counterbalanced across participants. A mandatory break was provided after 8 blocks, and participants were encouraged to take breaks as needed.

Results

For each participant, the proportion of correct responses (trials where both the correct Color and correct Number were reported) were calculated separately at each level of background rhythm alteration (m = 0.0, 0.25, 0.50, 0.75) and onset asynchrony (OA = +50ms, -50ms) (Figure 4). Compared to the equivalent OA conditions from Experiment 1 where the background contained only the Color and Number words, performance was overall much worse in Experiment 2 where the full background sentence was present. This suggests that the presence of a background rhythm may have disrupted temporal expectations for the target rhythm, leaving participants at a disadvantage for identifying (based on timing) which Color/Number pair came from the target and which came from the background. A 4 x 2 repeated measures ANOVA revealed no significant main effect of background rhythm alteration, $F(3, 45) = 2.10$, $p = 0.11$, $\eta^2 = 0.12$, or onset asynchrony, $F(1, 15) = 1.53$, $p = 0.235$, $\eta^2 = 0.093$, and no significant interaction, $F(3, 45) = 0.80$, $p = 0.50$, $\eta^2 = 0.051$.

Similar to Experiment 1, for each participant the proportion of Color intrusions (Fig. 5A) and Number intrusions (Fig. 5B) were calculated separately at each level of background rhythm alteration (m = 0.0, 0.25, 0.50, 0.75) and onset asynchrony (OA = +50ms, -50ms). There was a main effect of onset asynchrony for both Color, $F(1, 15) = 4.56$, $p = 0.05$, $\eta^2 = 0.23$, and Number, $F(1, 15) = 8.35$, $p = 0.01$, $\eta^2 = 0.36$, intrusions, but the direction of the effects were reversed: there were more Color intrusions when the target was leading (compared to when the background was leading) and more Number intrusions when the background was leading (compared to when the target was leading). There was no main effect of background rhythm alteration, nor was there an interaction between onset asynchrony and background rhythm alteration.

Intrusion errors accounted for nearly every incorrect trial, and averaged across background rhythm alteration and onset asynchrony participants selected the target word in their response (Color: *M* = 0.50, *SD* = 0.13; Number: *M* = 0.48, *SD* = 0.14) as much of the time as they selected the word coming from the background (Color: *M* = 0.49, *SD* = 0.13; Number: *M* = 0.52, *SD* = 0.14) for both Color, *t*(15) = 0.42, *p* = 0.68, 95% CI [-0.04, 0.05], and Number, *t*(15) = -1.43, *p* = 0.17, 95% CI [-0.09, 0.02]. If we assume that participants heard *both* Colors and *both* Numbers on each trial, chance performance would be 50% for either Color or Number. Notably, the proportion of correct Color responses and the proportion of correct Number responses were approximately 0.50 in all conditions, as were the proportions of Color intrusions and Number intrusions.

As in Experiment 1, we examined the relationship between overall performance and formal music training (*M* = 4.81, *SD* = 5.06), SSQ scores (*M* = 7.15, *SD* = 1.15), and ANE (*M* = 70.94, *SD* = 3.19). Proportion correct scores were averaged across both rhythm alteration and OA conditions individually for each participant. No correlation was found between performance and years of formal music training, *r* = 0.05, *p* = 0.86. There was also no relationship between performance and SSQ scores, *r* = -0.22, *p* = 0.41, or between performance and ANE, *r* = 0.13, *p* = 0.62.

Discussion

It was expected that with increasing alteration of the natural speech rhythm of the background precursor, recognition of target speech would improve. This would have replicated the previously observed background rhythm effect, while controlling the timing and intelligbility of Color and Number words. Instead, however, a background rhythm effect was not observed. Rhythm alteration of the background precursor had no effect on the proportion of correct responses or on the proportion of intrusion errors. One interpretation of this result is that the previous observations of the background rhythm effect were not attributable to background speech *rhythm* specifically, but instead were a result of incidental changes in the relative timing of target/background key words or in background key word intelligibility.

However, another possiblity that seems more likely in the present context is that aside from leaving Color and Number words intact and controlling onset asynchrony across rhythm conditions, the stimuli in this experiment differed in another critical way from the prior work of McAuley et al (2020, 2021). Namely, using the same CRM talker as both the target and the background talker eliminated fundamental frequency (F0) cues or other cues of speech quality that could have been used to initially segregate the target and background into separate auditory streams. Without this initial segregation that could be used to differentiate between and selectively track the speech rhythm of one sentence over the other, the combined target and background sentences might have been percieved as one auditory object with a jumbled, irregular rhythm. Such a jumbling of rhythms would be reminiscient of how a familiar melody interleaved with a rhythmically irregular tone sequence is not recognized until the interleaving tones are in a sufficiently different pitch range to be percieved as a separate auditory stream (Dowling, 1973). This interpretation is suggested by the result that both proportion correct scores and the proportion of intrusions computed separately for Color and Number were close to chance level, potentially indicating that participants could not differentiate between the target and background and had to guess which Color and Number came from which sentence.

GENERAL DISCUSSION

Prior experiments have established both a target rhythm effect and a background rhythm effect using the CRM paradigm that are consistent with a selective entrainment hypothesis. With the target rhythm effect, increasing alteration of the natural speech rhythm of a to-be-attended target sentence worsens recognition of target speech. With the background rhythm effect, increasing alteration of the natural rhythm of a distracting background talker (or talkers) improves recognition of target speech (McAuley et al., 2020, 2021). The observation of these effects supported a DAT-based *Selective Entrainment Hypothesis* which proposed that listeners' attention is selectively entrained by the natural rhythm of to-be-attended target speech, which facillitates the tracking of that speech over time in difficult listening situations. The *Selective Entrainment Hypothesis* would predict that background rhythm alteration improves target recognition because inadvertent entrainment to the background (at the expense of entrainment to the target speech rhythm) would be reduced. The background rhythm effect has proved fickle, however, and does not occur either when the background is unintelligible or can easily be segregated into a separate auditory stream based on strong fundamental frequency cues, suggesting that there might be more to the effect than a disrupted background rhythm alone (McAuley, Shen, Smith, & Kidd, 2021). The experiments reported here were designed to examine two stimulus characteristics that might have contributed to target speech recognition independent of the background rhythm itself: (1) the timing of background Color and Number words relative to the timing of target Color and Number words and (2) the intelligibility of background Color and Number words due to rhythm alteration. Either of these factors might have in part produced changes in performance with increasing alteration of the background rhythm. Experiment 1 additionally explored how the deviation or conformity of a distracting background Color and Number pair to the expected timing (based on the target rhythm) of the target Color and Number influences target recognition.

Experiment 1 demonstrates that when there are no temporal expectations coming from the background that could disrupt the buildup of temporal expectations for the target, the relative timing of backgound Color and Number words with respect to target Color and Number words alone is sufficient to influence performance. The results support both a *Temporal Expectation Hypothesis* and a *Temporal Order Hypothesis,* such that a distracting background Color and Number pair becomes less intrusive as the onset increasingly violates temporal expectations for the target Color and Number words, and are more intrusive when the background onset leads the target (compared to when the background onset lags).

The improvement of performance with large deviations from synchrony is compatible with the idea that the natural rhythm of target speech sets up temporal expectations for the occurance of future speech events, and that these expectations have an influence on speech perception. This is consistent with the broader literature on speech rhythm, which has shown that the temporal patterning of speech can influence how later speech events within the same stream are percieved in a way that is congruent with the expected continuation of the pattern (e.g. Dilley & McAuley, 2008; Baese-Berk et al., 2019). It is also consistent with the perspective of DAT, which would predict that a temporally predictable stimulus such as speech can entrain attentional rhythms in order to concentrate attentional energy near the expected time of future information-carrying stimulus events in order to better percieve those events and better ignore irrelevant ones (Jones, 1976; Jones & Boltz, 1989, Large & Jones, 1999).

The asymmetric effect of asynchrony on performance suggests that the distracting background Color and Number words are more intrusive when they appear prior to the expected onset of the target Color and Number words. In contrast, in Experiment 2 where the background precursor was present and the background Color and Number words appeared either 50ms before or after the target Color and Number words, there was no effect of onset order on performance. Additionally, Experiment 2 produced much worse performance overall than either the +50ms or -50ms OA conditions from Experiment 1. Since the presence or absence of the background precursor was the sole difference between the stimuli in Experiment 2 and the ±50ms OA stimulus conditions in Experiment 1, the contrast in performance and in the effect of onset order (or lack thereof) can be attributed to the background context.

The question then is what is the background precursor doing? If the presence of a *rhythmic* background context that can disrupt the development of precise temporal expectations for the target is what makes the task of Experiment 2 comparatively more difficult than Experiment 1, then the *selective entrainment hypothesis* would predict that as the background rhythm becomes increasingly irregular inadvertent entrainment to the background would be reduced and performance would improve. This was not the case in Experiment 2. Despite rhythmic alteration of the background precursor, participants were equally likely to select the Colors and Numbers from the background as they were to select the Colors and Numbers from the target no matter the level of alteration. This is in contrast to a previous experiment with a single male background talker that was different from the male target talker where the entirety of the background sentence (including the Color and Number) had the rhythm alteration applied to it. With the same levels of rhythm alteration (m = 0.0, 0.25, 0.50, 0.75) applied, there was a clear background rhythm effect such that an increasingly altered background rhythm led to improved performance (McAuley et al., 2021).

The discrepancy between the prior observation of the background rhythm effect and the results of the current Experiment 2 might still be explained by disrupted temporal expectations for the target. If participants were unable to distinguish between the target and background sentences toward the beginning of the stimulus, there would not be two distinct speech streams with rhythms that could be selectively entrained to but instead one single auditory stream with a jumbled and unpredictable rhythm. The background rhythm effect had previously been observed for a single-talker male background with a different male target talker, where the average F0 of the two talkers was similar but not identical and other vocal qualities might also have differed between them (McAuley et al., 2021). In Experiment 2 of the present study we instead used recordings from the same talker for both target and background, thus eliminating any characteristic differences that could be used to initially segregate target from background. While performance at baseline (with no background rhythm alteration) is comparable between Experiment 2 and this prior two-talker experiment, speech shaped noise had been added to the stimuli in the prior experiment in order to make the task more difficult. No such noise was added in the present study, suggesting that the lack of segregation cues available when the target and background talker were identical did indeed make it harder to distinguish the two sentences from each other. Not having a temporally predictable target rhythm that is perceptually distinct would also result in small deviations from synchrony not being as useful a cue for which Color/Number pair is correct, which would explain the reduction in performance with the reintroduction of the background precursor from Experiment 1 to Experiment 2.

At first, this explanation might seem to contradict an earlier interpretation of the lack of a background rhythm effect when the background talker was of a different sex than the target talker. We had suggested that the lack of an effect was due to the *presence* of a *strong* F0-based segregation cue, which rendered selective entrainment superfluous (McAuley et al., 2021). However, the joint evidence that the background rhythm effect does not occur *either* in situations where the target and background are easily segregated into separate auditory streams *or* when there is a lack or absence of primary segregation cues might suggest a sort of "Goldilocks" zone where the background rhythm becomes a deleterious presence. Such a Goldilocks principle would predict that the background rhythm effect will only occur if the following conditions are satisfied: (1) The listening situation is difficult enough to require the use of secondary perceptual processes for attending to the target in addition to the use of primary segregation cues, (2) There are sufficient segregation cues to facilitate the initial selection of one rhythm over the other as a source of temporal expectations, and (3) The source of the competing background rhythm can be mistaken for the target.

However, based on the experiments reported here we cannot eliminate the less interesting possibilities that the background rhythm effect is not an entrainment effect at all, but is instead an effect of either the background Colors and Numbers becoming less intrusive due to changes in intelligibility or systematic differences in the relative timing of the Color and Number words. To investigate the first possibility further, it will be important to compare the intelligibility of isolated Color and Number pairs amidst noise with different amounts of rhythm alteration. It has been established with pilot testing that the rhythm alteration does not impact intelligibility in quiet (McAuley et al., 2020), but if the intelligibility of the words that participants are meant to report is reduced when presented in more difficult listening situations, this could lead to both a background rhythm effect (by virtue of a reduction in intrusions) in situations where intrusions are likely and a target rhythm effect regardless of the possibility of intrusions.

The second possibility that the background rhythm effect is one of relative timing of Color and Number words seems somewhat less plausible. The relative onset of background color and number words varied from trial to trial in past work since the background rhythm alteration was applied with different phases, which should have prevented any systematic differences in onset asynchrony between rhythm conditions. Additionally, there was no difference in performance in the present study between the +50ms and -50ms OA conditions in the presence of the background precursor and any incidental changes in onset asynchrony due to rhythm alteration in prior experiments were likely small. The fact that there was such a large effect of onset asynchrony in Experiment 1, however, means that this possibility cannot be entirely dismissed and warrants further scrutiny.

Overall, the present pair of experiments demonstrates that expectations for target speech timing can help listeners to distinguish between a target and an intruding background, but also that such temporal expectations are weakened or no longer useful when target and background speech are not distinct enough (e.g. based on fundamental freuqency differences) to trigger initial perceptual segregation and selection. Future experiments will further investigate the effect of onset asynchrony when target speech timing is irregular and rhythm-based temporal expectations are weakened, and a second line of experiments will systematically vary the F0 difference between target and background when the background context is intact or rhythmically irregular.

APPENDIX

APPENDIX

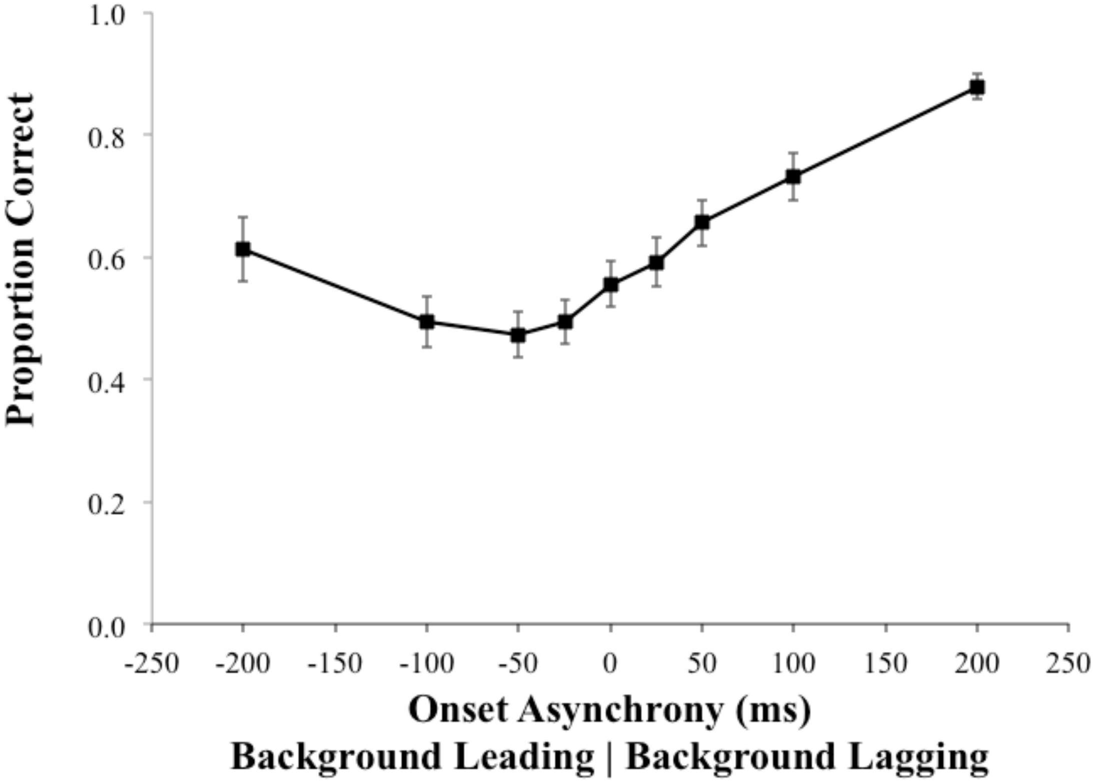

Figure 1. Results for Experiment 1: Proportion correct target Color and Number recognition for each value of onset asynchrony (0, ±25ms, ±50ms, ±100ms, ±200ms). Negative values indicate that the onset of the background Color appeared early relative to the onset of the target Color (background leading), while positive values indicate that the onset of the background Color appeared late relative to the target Color (background lagging). Error bars represent standard error.

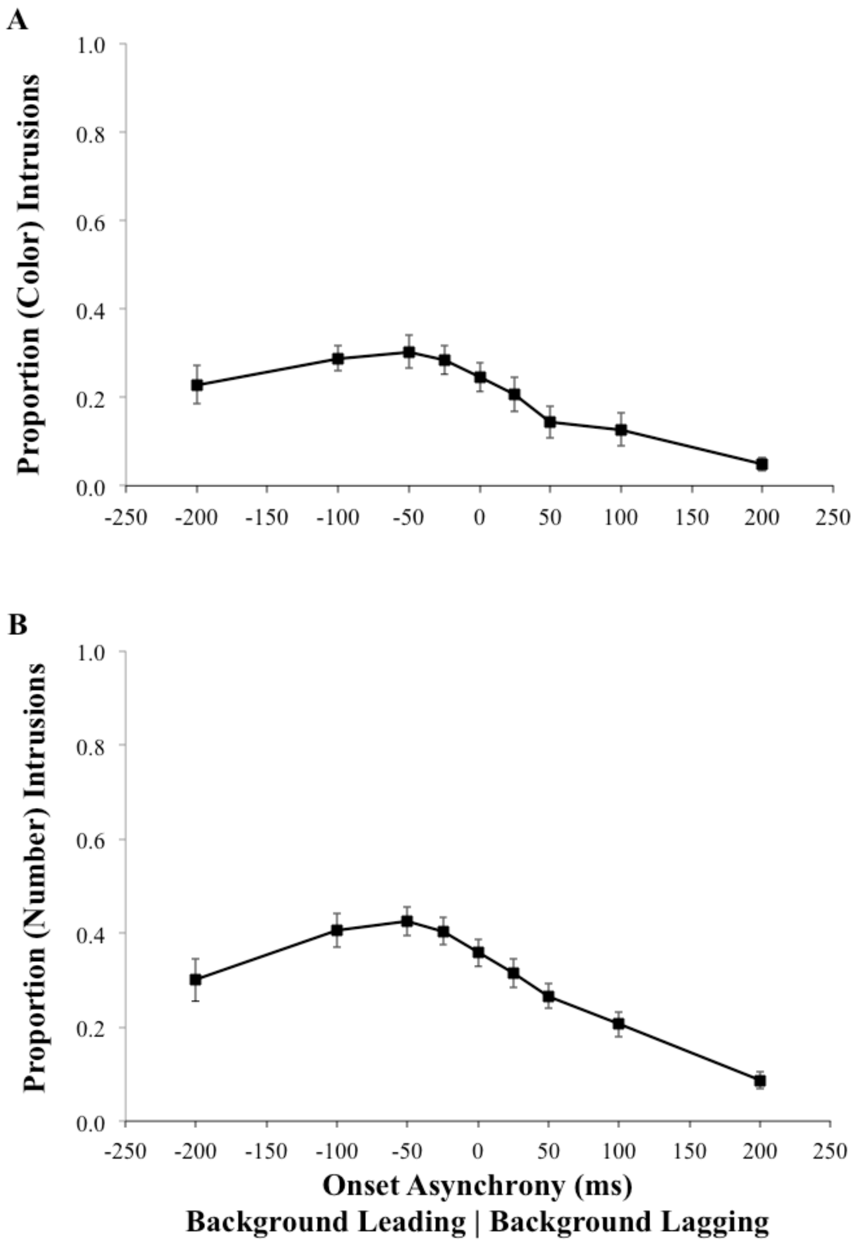

Figure 2. Results for Experiment 1: Proportion of Color (Panel A) and Number (Panel B) intrusions for each value of onset asynchrony (0, ±25ms, ±50ms, ±100ms, ±200ms). Negative values indicate that the onset of the background Color appeared early relative to the onset of the target Color (background leading), while positive values indicate that the onset of the background Color appeared late relative to the target Color (background lagging). Error bars represent standard error.

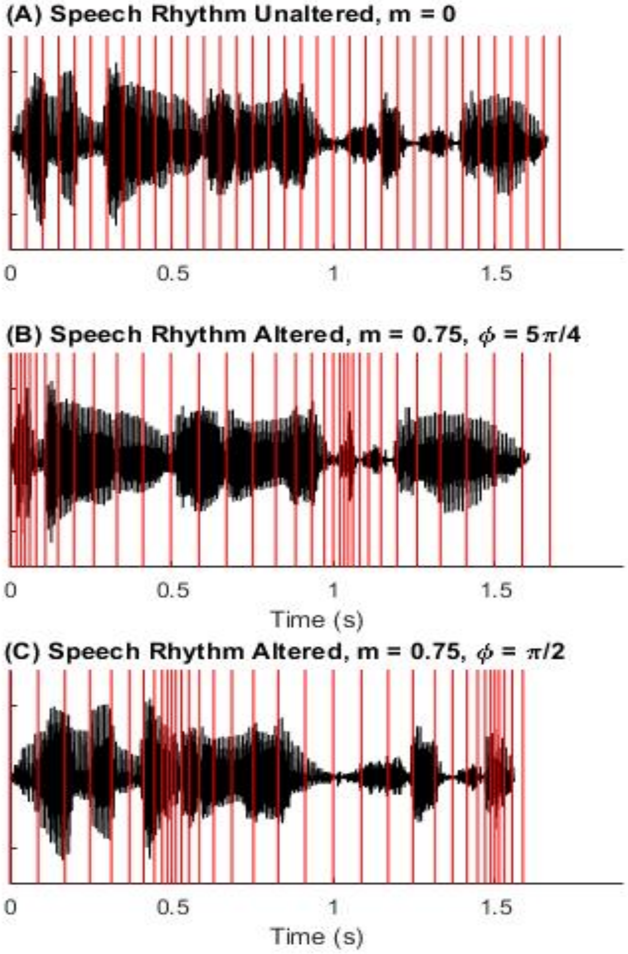

Figure 3. Examples of rhythm unaltered and altered versions of a spoken CRM sentence of the form "Ready [call sign] go to [color] [number] now.' The top panel (Panel A) shows the sample sentence where the rhythm is unaltered (*m* = 0), as represented by the bars equally spaced in time. The middle and 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 (Panel B, phi = 5π/4; Panel C, phi = π/2).

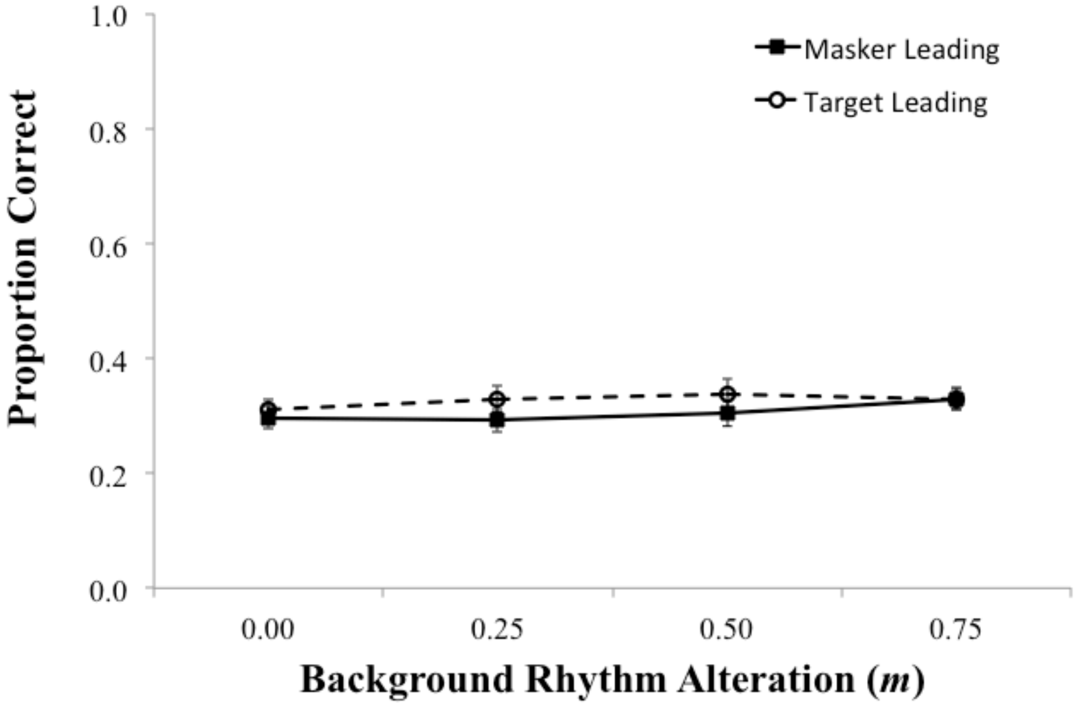

Figure 4. Results for Experiment 2: Proportion correct target Color and Number recognition for each level of background rhythm alteration (*m* = 0.0, 0.50, 0.25, 0.75). Black squares with a solid line represent performance when the background Color/Number was leading (OA = -50ms) and open circles with a dashed line represent performance when the target Color/Number was leading (OA = +50ms). Error bars represent standard error.

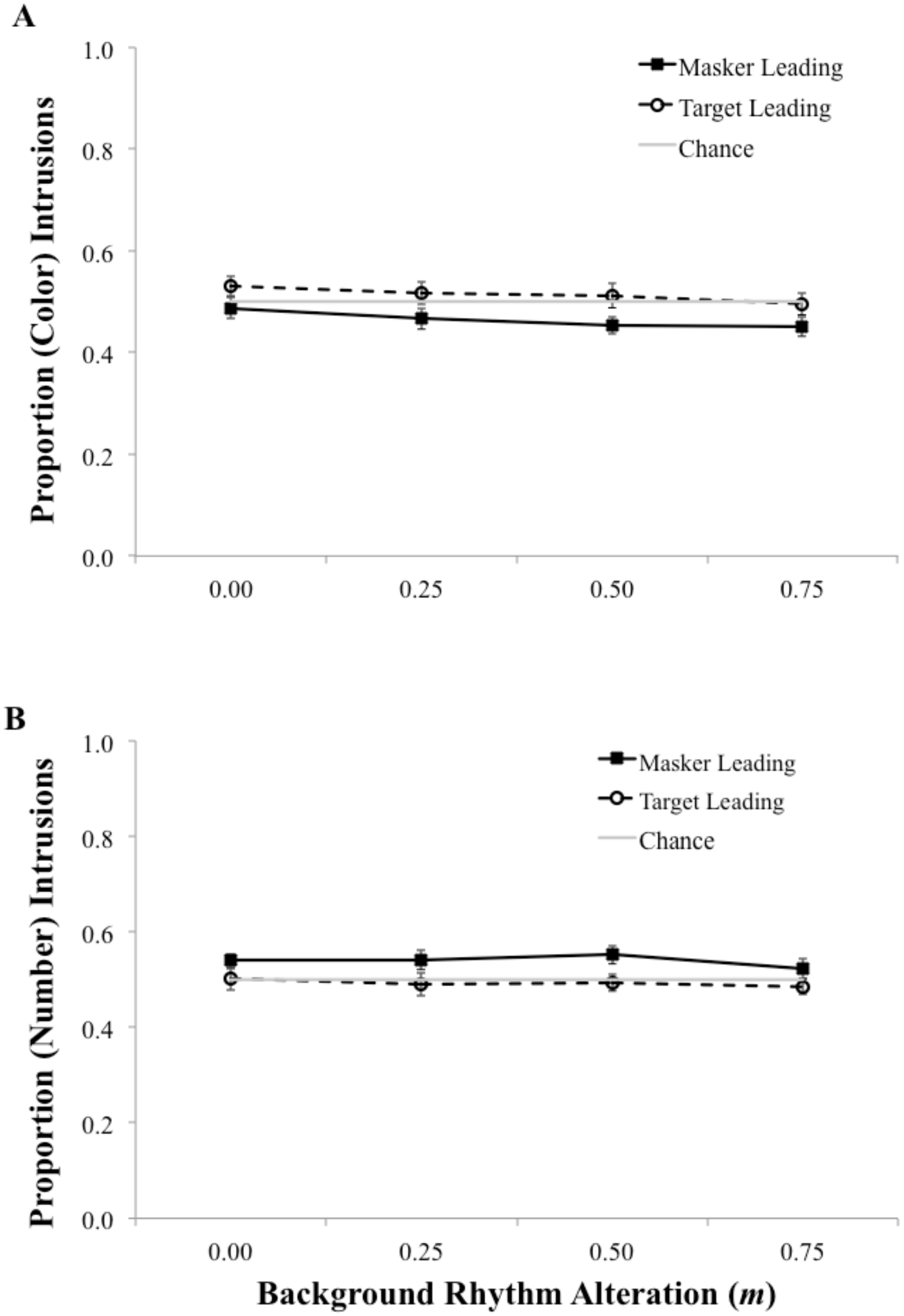

Figure 5. Results for Experiment 2: Proportion Color (panel A) and Number (Panel B) intrusions for each level of background rhythm alteration (*m* = 0.0, 0.50, 0.25, 0.75). The solid grey lines represent the chance of choosing the background Color or Number at random when both background and target Colors and Both numbers are heard. Black squares with a solid line represent performance when the background Color/Number was leading (OA = -50ms) and open circles with a dashed line represent performance when the target Color/Number was leading (OA = +50ms). Error bars represent standard error.

REFERENCES

REFERENCES

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.

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.

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*(2), 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.

Booth, A. J., & Elliott, M. T. (2015). Early, but not late visual distractors affect movement synchronization to a temporal-spatial visual cue. *Frontiers in psychology*, *6*, 866.

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

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. Journal of Phonetics, 10(1), 23-36.

Darwin, C. J., & Ciocca, V. (1992). Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component. *The Journal of the Acoustical Society of America*, *91*(6), 3381-3390.

Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *The Quarterly Journal of Experimental Psychology Section A*, *33*(2), 185-207.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. Journal of Phonetics, 11, 51-62.

Desjardins, J. L., & Doherty, K. A. (2013). Age-related changes in listening effort for various types of masker noises. *Ear and hearing*, *34*(3), 261-272.

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(29), 11854-11859.

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

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

Dowling, W. J. (1973). The perception of interleaved melodies. *Cognitive psychology*, *5*(3), 322-337.

Frings, C., & Moeller, B. (2012). The horserace between distractors and targets: Retrieval-based probe responding depends on distractor–target asynchrony. *Journal of Cognitive Psychology*, *24*(5), 582-590.

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.

Goswami, U. (2019). Speech rhythm and language acquisition: an amplitude modulation phase hierarchy perspective. Annals of the New York Academy of Sciences.

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.

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*(3), 741-754.

Johnson, T. A., Cooper, S., Stamper, G. C., & Chertoff, M. (2017). Noise exposure questionnaire: A tool for quantifying annual noise exposure. *Journal of the American Academy of Audiology*, *28*(1), 14-35.

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.

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

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.

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

McAuley, J. D., Shen, Y., Smith, T., & Kidd, G. R. (2021). Effects of speech-rhythm disruption on selective listening with a single background talker. *Attention, Perception & Psychophysics,* 1-12

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.

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.

Noble, W., Jensen, N. S., Naylor, G., Bhullar, N., & Akeroyd, M. A. (2013). A short form of the Speech, Spatial and Qualities of Hearing scale suitable for clinical use: The SSQ12. *International journal of audiology*, *52*(6), 409-412.

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

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.

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.

Turgeon, M., Bregman, A. S., & Roberts, B. (2005). Rhythmic masking release: effects of asynchrony, temporal overlap, harmonic relations, and source separation on cross-spectral grouping. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(5), 939.

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.