Journal of Memory and Language 147 (2026) 104730

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Evidence of cross-domain phase entrainment effects between nonspeech tones and speech sounds

Tzu-Han Zoe Cheng *, Sarah C. Creel

Department of Cognitive Science, UC San Diego, La Jolla, USA

ARTICLE INFO

Keywords:
Entrainment
Durational contrast
Domain-general effect
Speech categorization
Time perception

ABSTRACT

Temporal prediction and duration estimation are critical to speech perception. Studies have reported that the *rate* of nonspeech precursors affects speech categorization, potentially through an underlying non-speech-specific mechanism known as entrainment, supporting a domain-general account of language processing. However, few studies have examined the effects of precursor *phase* on speech perception, despite the centrality of phase to entrainment models and prevalent findings of phase effects in music and nonspeech tones. Experiments 1–2 looked for phase entrainment between nonspeech precursors (rhythmic series of sine tones) and speech sound targets, either *lap/lab* (Experiment 1) or *at/add* (Experiment 2), plus short/long categorization of target *tones* with speech-matched envelopes and durations. Experiment 3 mirrored Experiment 2, but tested phase entrainment from *speech* precursors (rhythmic series of speech pips) to speech targets *at/add* and the matching tones. Experiment 1 found phase entrainment for tones but not speech targets. Experiment 2 better equated the perceptual centers of words and tones, and a cross-domain phase entrainment effect appeared for words. Experiment 3 showed equivalent entrainment effects on speech and tone targets. These results suggest that phase entrainment appears in both tones and speech, and that one domain can entrain another in both directions. Still, across the three experiments, the tone-to-tone entrainment appeared more reliably and showed larger effect sizes than speech-to-speech, tone-to-speech, or speech-to-tone entrainment, which raises questions as to the strength of the influence of phase entrainment in processing of speech and other spectrotemporally complex sounds.

Introduction

Entrainment as a supramodal process of human time perception

Human listeners use event duration to accurately identify complex sounds such as speech and music. One hypothesized, supramodal neural mechanism underlying duration estimation and other temporal processing is entrainment (originally proposed by Jones, 1976), in which brain oscillations synchronize with external rhythms to scaffold the processing of temporally predictable sensory cues such as musical beats or word stress (Goswami, 2012, 2019; Haegens & Zion Golumbic, 2018; Harding et al., 2025; Lakatos et al., 2019; Poeppel & Assaneo, 2020; Rimmele et al., 2018). Researchers have claimed that language processing is shaped by entrainment (Dilley & McAuley, 2008; Doelling et al., 2014; Giraud & Poeppel, 2012; Heffner et al., 2017; Steffman, 2021; Zion Golumbic et al., 2013). Many findings further highlight a strong link between musical and language processing (Fiveash et al., 2021; Fujii & Wan, 2014; Goswami, 2012; Ladányi et al., 2020), hinting at a shared, domain-general mechanism for temporal processing (Bauer et al., 2020). Despite this intense interest, little research addresses how entrainment is transferred among different acoustic domains. Here, we test phase entrainment effects both within domains (nonspeech tone precursors, tone targets; speech precursors, speech targets) and across domains (tone precursors, speech targets; speech precursors, tone targets).

Duration perception of tone targets is influenced by phase

Existing research suggests that the *phase* of a target stimulus relative to previous events affects its processing. Numerous researchers have found better sensory detection of in-phase events when brain oscillations entrain to match stimulus rhythm (de Graaf et al., 2013; Henry et al., 2014, 2016; Hickok et al., 2015; Kizuk & Mathewson, 2017; Mathewson et al., 2010, 2012; Spaak et al., 2014; see Haegens and Zion Golumbic, 2018 Section 2 for a thorough review). Behaviorally, when a sequence of tones evenly spaced in time precedes a target tone, in-phase

* Corresponding author.
E-mail addresses: tzcheng@ucsd.edu (T.-H.Z. Cheng), screel@ucsd.edu (S.C. Creel).

https://doi.org/10.1016/j.jml.2025.104730
Received 20 August 2024; Received in revised form 17 December 2025; Accepted 18 December 2025
Available online 24 December 2025

tone targets are processed more accurately than out-of-phase (slightly early or late) targets (Barnes & Jones, 2000; Hickok et al., 2015; Jones et al., 2002; McAuley & Jones, 2003; McAuley & Kidd, 1998; see Saberi & Hickok, 2023 for a comprehensive discussion). Additionally, multiple researchers (McAuley & Jones, 2003; McAuley & Kidd, 1998) including us (Cheng & Creel, 2020) have shown that *perception of the target's duration* is distorted systematically, with early targets reported as shorter than on-time or late targets of equal duration, and late targets perceived as slightly longer (Fig. 1). This *entrainment distortion* provides a useful behavioral probe of presence or absence of entrainment.

Evidence of cross-domain entrainment from context tones to speech sound categorizations

An exciting emerging trend is the extension of the entrainment findings from nonspeech sounds (sine tones, light flashes) to speech sounds (ten Oever & Martin, 2021; ten Oever & Sack, 2015; ten Oever et al., 2024), in which duration information contributes to listeners' recognition of sounds/words. For example, coda consonant voicing in English is differentiated by (among numerous other features) the duration of the preceding vowel (Raphael, 1972), such as *lap* (shorter vowel) vs. *lab* (longer vowel; see Section 2.1.2 in Methods for more details).

Perception of speech stimuli is affected by the temporal properties of both speech and nonspeech precursors (Diehl et al., 1980; Port, 1979; Summerfield, 1981). A few recent studies have investigated entrainment-like effects on duration perception in spoken language by manipulating rate and rhythmic properties of precursor speech or nonspeech. Precursor speech *rate* (Heffner et al., 2017) and precursor speech *rhythmic patterns* (Steffman, 2021) affect perception of coda voicing of the target word (see related work by Kidd, 1989; and work on segmentation by Dilley and McAuley, 2008). Especially relevant to the current work, a few studies have reported context rate effects of *non*speech precursors on speech categorization: faster tones bias listeners to perceive target speech sounds as longer, while slower tones bias listeners to perceive target speech sounds as shorter, thus changing apparent word identity (Dutch /as/ vs. the longer-duration /a:s/ in Bosker, 2017; English [b] vs. the longer-duration [w] in Wade & Holt, 2005, but also see Pitt et al., 2016, who found that only precursors perceived as speech can generate context rate effects on speech perception). These studies are predominantly consistent with a general auditory timing mechanism underlying speech perception, counter to modular accounts of language processing (Liberman et al. 1961; Mattingly et al. 1971).

Yet only one study so far (Bosker and Kösem, 2017) has examined *phase* effects from nonspeech tone precursors on speech perception, and it found no phase influences. This is crucial for claims that speech undergoes entrainment, in that phase is the element in the originally proposed entrainment model that distinguishes interval-based timing and entrainment-based timing (see Jones, 1976; McAuley & Jones, 2003; Repp, 2002; Cheng & Creel, 2020 for detailed description) and appears critical to findings of attentional benefit for in-phase stimuli (de Graaf et al., 2013; Henry et al., 2014, 2016; Hickok et al., 2015; Kizuk & Mathewson, 2017; Mathewson et al., 2010, 2012; Spaak et al., 2014; Saberi & Hickok, 2023). In brief, it is possible that an interval model tracking the prevalence of particular durations could account for precursor *rate* effects. However, because interval models do not track phase, they cannot account for *phase* effects. If entrainment is occurring, duration distortion effects based on the phase relationship between the context and the target should also occur.

To date, it is unclear to what extent precursor entrainment, especially phase, spreads from one acoustic domain in the precursor (e.g., tones) to another domain for the target (e.g., speech), and it's also unclear whether speech is subject to phase entrainment at all, even from other speech. This leaves uncertain whether entrainment is truly domain-general, cross-cutting speech and nonspeech, or domain-specific, perhaps limited to highly simplified perceptual materials or musical contexts. If entrainment is occurring during speech perception, then duration distortion effects based on the phase relationship between the context and the target should also occur. This raises the question of whether speech is truly impervious to phase entrainment, which would

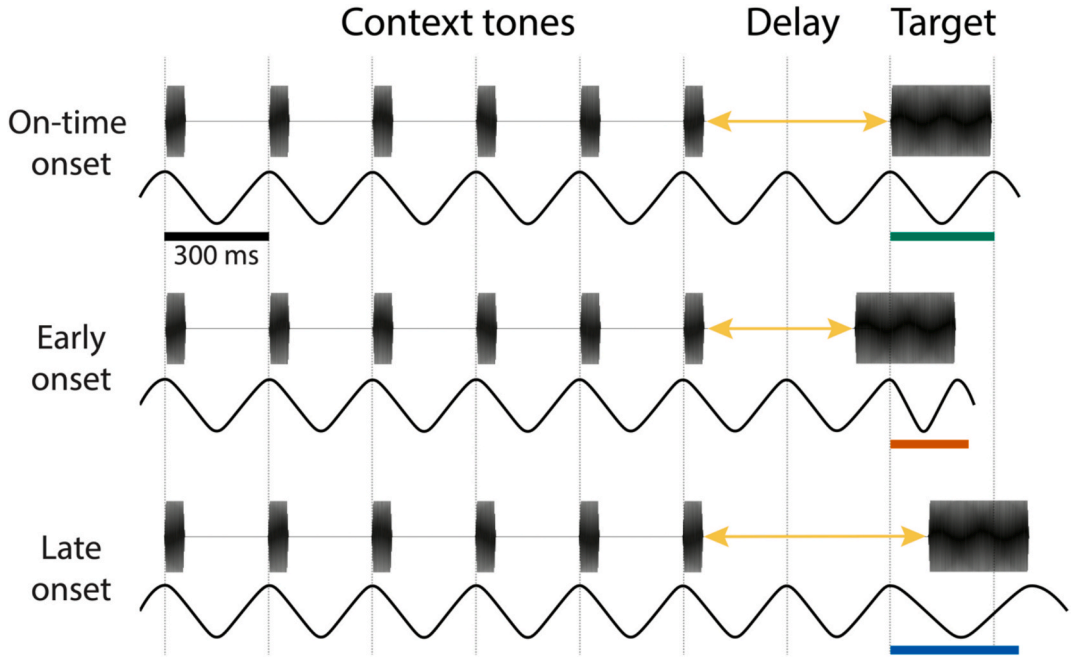

Fig. 1. Schematic illustration of entrainment distortion when the target interval is presented on-time, earlier, or later than expected. On the top row, the six short waveforms are the context sine tones (IOI = 300 ms), and the longer waveform is the target sine tone that subjects judge as being longer vs. shorter. After the delay (yellow double arrow), the target tone is presented on-time (top), early (middle), or late (bottom). The waves represent the listener's internal oscillatory state, or expected timing, entrained by the context tones. The thick horizontal color lines indicate the subjectively perceived durations such that the internal representation of duration is shortened in the early onset (red line), lengthened in the late onset (blue line), and unbiased in the on-time onset (green line). See details in our previous work (Cheng & Creel, 2020).

suggest that entrainment is domain-specific, or, instead, whether controlling for additional factors might reveal phase entrainment effects in speech, which would suggest that entrainment is a domain-general process.

The current study

Here we asked whether there are cross-domain phase entrainment effects between nonspeech tones and speech sounds, and whether speech sounds are subject to entrainment generally. An answer to this question contributes to the long-standing debate between domain-specific and domain-general accounts of speech processing as well as a fundamental understanding of entrainment mechanisms in speech and nonspeech tone perception. We measured entrainment distortion (Fig. 1) as a function of the phase of the target onset (early, on-time, late) as in our previous study (Cheng & Creel, 2020) across different auditory targets (Speech vs. speech-shaped Tone). We had two goals. Our primary goal was to ask whether phase entrainment is domain-specific vs. domain-general. If entrainment is domain-general, entrainment distortion should occur for both speech-shaped tone and speech target regardless of the tone/speech precursors. If it is domain-specific in the sense that only nonspeech sounds are susceptible to phase entrainment, then speech targets may not entrain at all. These predictions were addressed in three experiments, with different sets of speech sound targets and corresponding tone targets (Experiment 1: *lab/lap*; Experiments 2 and 3: *add/at*). Note that Experiment 1′s outcome was foreseen in two earlier, separately-run pilot experiments (Supplementary Section 1), but here we explored a different set of speech sounds/matched tones and ran each of the current studies as a unitary experiment with random assignment to conditions to verify that previous outcome. Also note that, unlike the sine-wave context tones, all tone targets in this study are speech-shaped tones, modulated by the speech envelope to enhance their similarity to speech. For convenience, speech-shaped tone targets will henceforth be referred to simply as "tone targets."

Our secondary goal was, if no entrainment is observed for speech (as implied by our earlier experiments), to make an initial probe as to potential reasons why. One possibility is that auditory dissimilarity between context tones and target words blocks phase effects of entrainment, which we term the *auditory grouping hypothesis*. To this end, we included in Experiment 1 an additional Tones-as-Speech condition which asked subjects to classify Tone targets as speech stimuli, and across Experiments 2–3, we varied the nature of the entrainers themselves (tones in Experiment 2 vs. speech sounds in Experiment 3). On the auditory grouping hypothesis, Tones-as-Speech targets, which *do* auditorily group with context tones while subjects use their internal representations to categorize the words, should show entrainment. Further, entrainment distortion should occur for tone targets but not for speech targets after tone precursors (Experiments 1–2), and it should occur for speech targets but not for tone targets after speech precursors (Experiment 3). Another possible reason why speech might not show entrainment is that *subjectively* perceived event onset time between tone and speech, known as the perceptual center ("p-center") is misaligned. Various researchers (Cooper, Whalen, & Fowler, 1986; Marcus, 1981; Morton, Marcus, & Frankish, 1976) have found p-centers to be influenced by many factors such as the type of consonants prior to the vowel, amplitude envelope, and more. We attempted to control for this in Experiment 1 by matching amplitude envelopes between speech and tone stimuli, but to prefigure our findings somewhat, a control study revealed p-center differences between Experiment 1's tone and speech stimuli. This and a second control study also identified a set of speech stimuli with p-centers similar to tones, which we used in Experiment 2 and 3.

Experiment 1

Methods

Participants

We tested 300 participants from UCSD via the online platform FindingFive (FindingFive Team, 2019; findingfive.com), as pre-registered and approved by UCSD's Institutional Review Board. Participants electronically acknowledged informed consent to participate and received one course credit for participation. Power analysis was done on the main effect of the Onset Times from pilot experiments to determine the target sample size. We estimated an effect size of partial eta squared = .106. With this effect size, G*Power (Faul et al., 2007) indicated that a sample size of 49 or higher was needed to achieve power of.85. We rounded this up slightly to 50 to reach a multiple of 2, given two response key assignments. To ensure we collected enough high-quality data collected via online experiment, we doubled the number of target sizes to 100 for each group of subjects.

Participants were excluded for: scores below 90% for catch trials, suggesting inattentiveness (n = 44); reporting intrusive environmental noise (n = 4); reversed (i.e., positive) or zero identification slopes from individual logistic fits (n = 29); aberrant patterns in 50 % point data (n = 31), where "aberrant" was characterized as more than 1.5 times the interquartile range (IQR) below the first or above the third quartile (Rousseeuw & Croux, 1993). Importantly, all exclusions were blind to condition. Final analyzed sample sizes were: Speech condition, $n = 62$; Tone condition, $n = 72$; Tone-as-Speech condition, $n = 57$.

Stimuli

Each trial in the main study presented six context tones followed by a single target (speech or amplitude-modulated tone). Context tones were 60-ms, 440-Hz sine tones generated in MATLAB (MathWorks), with inter-onset intervals (IOIs) of 300 ms. In the Speech condition, target words were constructed from natural recordings of a monolingual English female talker sampled at 44,100 Hz in a soundproof room on Adobe Audition software. We tested English coda consonant voicing perception because it is influenced by vowel duration (Raphael, 1972), with vowels varying on a roughly similar time scale to tone durations in the prior work (Cheng & Creel, 2020). Note that many other cues to coda consonant voicing have also been identified (e.g. Hillenbrand et al., 1984; Revoile et al., 1982; Wolf, 1978), but this is immaterial to the question at hand, which is whether perceived vowel duration (and thus speech sound identification) is *affected by precursor rate*. We chose the word pair *lab* vs. *lap* in Experiment 1 because the coarticulation between the word-initial consonant and the vowel has a smooth transition, allowing us to vary syllable duration continuously to avoid acoustic unnaturalness. We lengthened *lap* in 20-ms increments, yielding eight steps ranging from 246 ms total word duration (perceived as *lap*) to 386 ms total word duration (perceived as *lab*). The word-final consonant release, another potential cue to voicing (e.g. Revoile et al., 1982; Wolf, 1978), was removed so that participants could not use it to identify words. All words were normalized to the same loudness of 70 dB SPL. The effects of vowel duration on subjects' sound identification were robust, providing evidence that lengthening the "*lap*" stimulus resulted in substantial increases in "*lab*" responses (see Fig. 3 and the statistics in Results).

For Tone and Tone-as-Speech conditions, we created speech-shaped tone targets matched in duration with the *lab* vs. *lap* continuum by convolving the sine tone with the upper half of the peak envelope of each speech stimulus step, thereby matching the speech for length and amplitude modulation. We did this because the amplitude envelope strongly influences p-center (Cooper, Whalen, & Fowler, 1986; Marcus, 1981; Morton, Marcus, & Frankish, 1976; see also Danielsen et al., 2019).

Design and procedure

As outlined in our pre-registration, the two independent variables

were Auditory Targets (between-groups; Speech, Tone, Tone-as-Speech) and Onset Times (within-participants; Early, On-time, Late). We had planned to assess two dependent variables: (1) proportion "short" responses and (2) 50% point in the logistic categorization curve. In hindsight, we determined that logistic mixed-effects regression models were more appropriate for capturing trial-by-trial two-alternative forced choice (2AFC) responses and the full shape of the response curve, rather than averaging responses across trials and along the acoustic continuum. Therefore, we used the glmer function from the lme4 package in R (Bates et al., 2015) to create logistic regression models of "short"/"*lap*" responses to the three Onset Times (with Early as the reference level) and eight Target Durations (coded as a mean-centered linear variable) for each of the three Auditory Targets (Speech, Tone, Tone-as-Speech).

The three Onset Times were coded using simple coding, which is similar to dummy coding with a reference level of Early but with predictors centered. One predictor compared early to on-time target onset, and the other compared early to late target onset. To test the significance of (interactions with) the overall Onset Time effect, we compared the full model to a "holdout" model which removed only the (interaction including the) Onset Time fixed effect predictors. The three Auditory Targets were coded by Helmert coding that contrasts Speech vs. the other two, and the other two (Tone, Tone-as-Speech) against each other. The logistic regression results are reported in the main text. The pre-registered proportion "short" responses and 50% points are visualized in the main text, while the ANOVA results are reported in the Supplementary Materials. Technically, it was not necessary to exclude aberrant 50% point data for the logistic regression model, as this measure is only relevant to the ANOVA analysis. However, we chose to maintain a consistent final sample across both the pre-registered ANOVA and logistic regression for comparability. Notably, the logistic regression results remained qualitatively similar regardless of whether these aberrant data points were included or excluded. The experiment randomly assigned each participant to one of the three conditions (total time < 60 min).

Pretests. Conditions differed slightly in their instructions and pre-tests. In the Speech condition, participants completed 120 identification trials with six coda-voicing word pairs (*bag/back*, *lag/lack*, *fad/fat*, *bead/beat*, *cab/cap*, *lab/lap*) without feedback. Next, they completed 240 *lab/lap* identification trials using the full 8-step continuum without feedback. In the Tone condition, participants heard 4 examples (2 each) of the shortest and longest isolated tones and then performed a *short/long* identification task (240 trials total) for the shortest and longest tones without feedback. They then heard 4 examples of the full test sequence (on-time condition only, 2 each for the shortest and longest target tones) and practiced those examples *with* feedback for 24–36 trials depending on their performance. In the Tone-as-Speech condition, stimuli and procedure matched the Tone condition, except that we told participants that the sine tones were made from l*ab* or *lap*, and they should respond to which word they thought the tone was.

Main task. In each trial of the main task (288 trials), six context sounds with an IOI of 300 ms were followed by a target sound (Fig. 2). The speech or tone target occurred early (510 ms after the last context tone onset), on-time (600 ms, or 2 "beats," after) or late (690 ms after). Interspersed catch trials (n = 26) asked participants to report what animal sound occurred. Timing choices followed our original studies (Cheng & Creel, 2020), which were based on a classic duration discrimination task (McAuley & Jones, 2003). McAuley and Jones (2003) presented "early" and "late" items 180 ms before/after the beat at a beat rate of 600 ms, so the deviations were ± 30% of the beat duration. We retained these values (30% of the beat duration early or

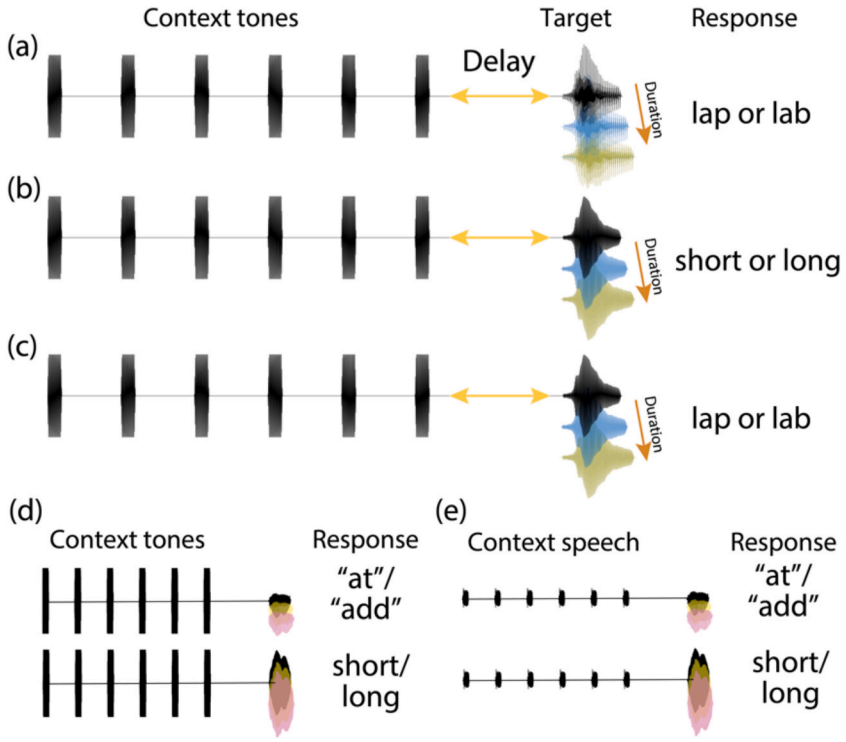

Fig. 2. Schematic illustration of the experimental tasks, including context, target and response. (a–c) Experiment 1 paradigms. In the word discrimination task, context tones were followed by (a) a speech target from the "*lap*"/"*lab*" continuum or (c) a tone-as-speech target from a short/long tone continuum. In the duration discrimination task, context tones were followed by (b) a tone target from a duration continuum. Tone and tone-as-speech stimuli were shaped with speech-matched envelopes and durations (246–386 ms in eight steps); only the first three durations are shown here (black, blue, green). The orange arrow indicates increasing target duration. The yellow double-headed arrow denotes the delay duration corresponding to early, on-time, and late target onsets. (d) Experiment 2 paradigm with context tones followed by either a speech target from the "*at*"/"*add*" continuum or a matched tone target. (e) Experiment 3 paradigm with context speech pips followed by the targets used in Experiment 2.

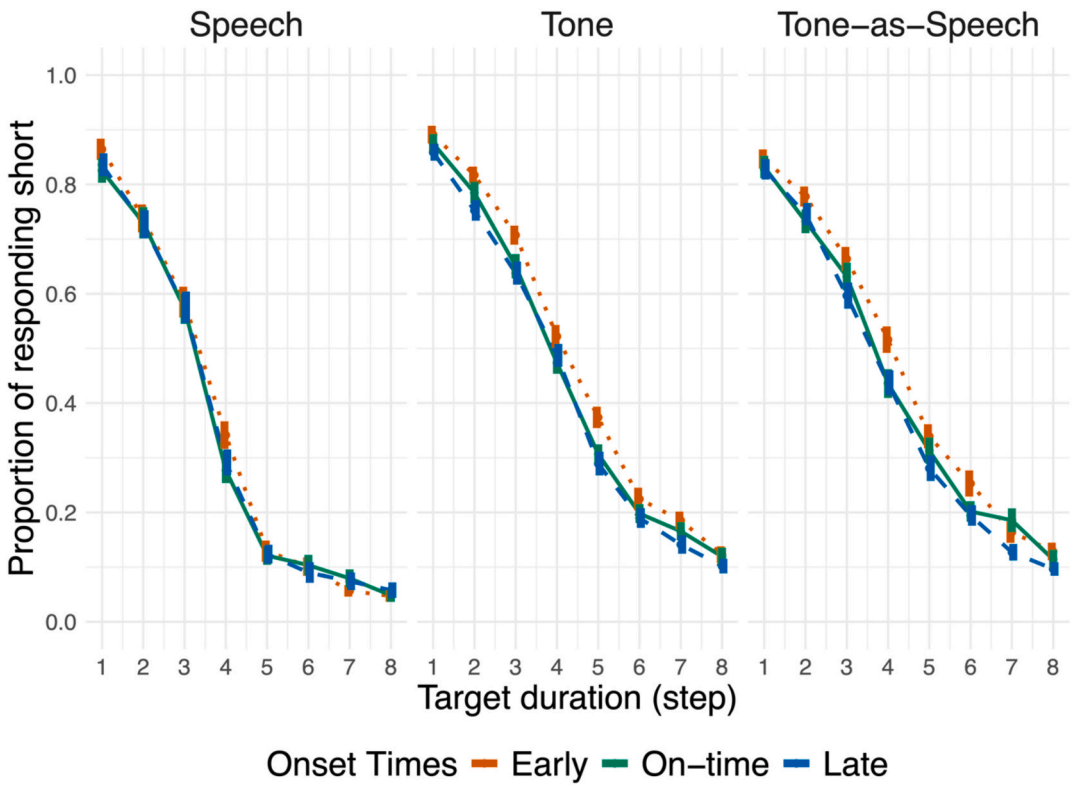

Fig. 3. Response curves averaged across subjects for each Auditory Target, with standard errors, in Experiment 1. The x-axis represents step duration from shorter to longer sounds. The y-axis represents the proportion of responding short (or "*lap*") averaged across all subjects for targets presented early (red dotted lines), on-time (green solid lines) and late (blue dashed lines). Note that participants were sensitive to changes in target stimulus duration (x-axis). Each vertical bar represents standard errors across subjects.

late, i.e. 90 ms before/after the beat at a beat rate of 300 ms) in our current study because we consider they are a reasonably large deviation from "on time."

Questionnaires and debriefing. Finally, each participant completed demographic, language, and music background questionnaires and debriefing questions.

Results

The results were shown in Figs. 3 and 4. We had planned to use ANOVA to detect the phase entrainment effect in our pre-registration (osf.io/r37ax). In hindsight, and in light of suggestions from the review process, we determined that logistic mixed-effects regression models were more appropriate for capturing the trial-by-trial, full range of 2AFC responses along the acoustic continuum. The corresponding

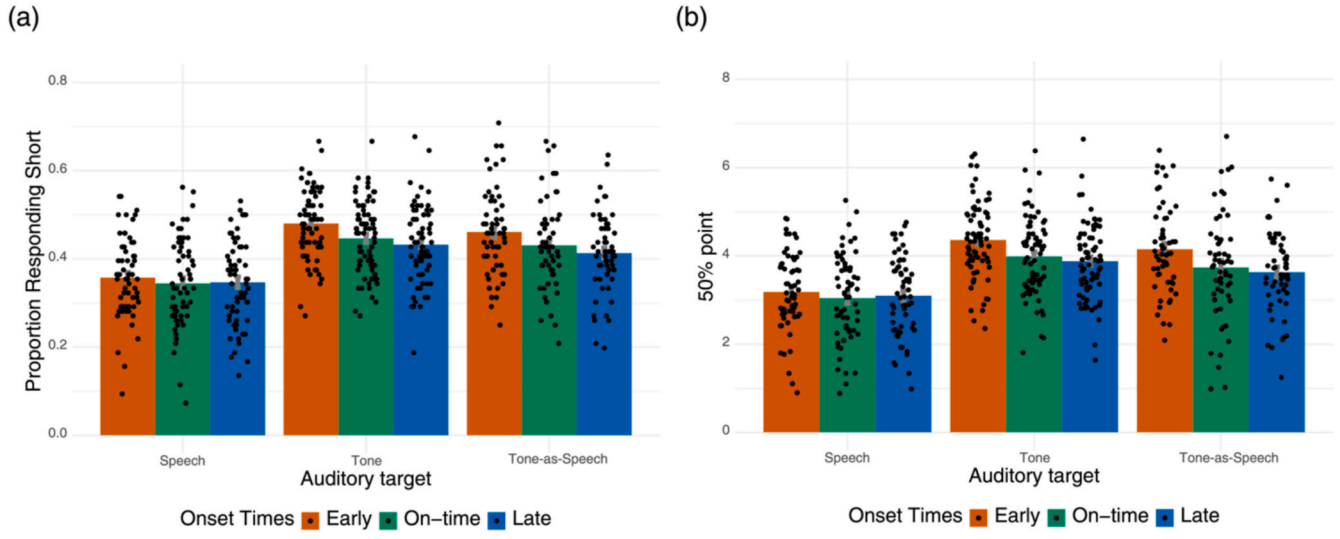

Fig. 4. Bar plots of proportion short and 50% point for each onset time among Auditory Target in Experiment 1. The x-axis represents different auditory targets. The y-axis represents (a) the proportion of responding short and (b) 50% points for targets presented early (red bars), on-time (green bars) and late (blue bars). Bar plots show the group means and standard errors. Each black dot represents one individual.

ANOVA results, as outlined in the pre-registration, are available in the Supplementary Materials.

Cross-domain phase entrainment

If entrainment occurs across domains, listeners should respond "short" (for tones) or "*lap*" (for speech; *lap* is acoustically shorter than *lab*) more often in the Early than the other two onset times (see Fig. 1 and our previous work; Cheng and Creel, 2020, for rationale). Note that more negative estimates indicate stronger entrainment distortion effects, corresponding to lower probability of "short" responses for e.g. Late compared to Early onset times. If entrainment is fully domain-general, then we should observe an effect of Onset Times across all conditions, with no Auditory Targets x Onset Times interaction. If entrainment does *not* occur across domains, we should observe a significant interaction between Auditory Targets and Onset Times, with entrainment in the Tone condition but not the Speech condition. To assess significance, we compared the full model to holdout models which removed only a single fixed effect at a time. We also compared the model of each Auditory Target to a corresponding holdout model that removed Onset Times to test for entrainment effects. Stepwise elimination with the buildmer package was used to determine the random effects structure of the largest regression model that could still converge.

In the full Auditory Target x Onset x Target Duration model, we first verified that there is a significant effect of Target Duration in the full model ($\chi^2 = 351.22, p < .000$), confirming that there is a strong decrease in "short"/"*lap*" responses with increasing target durations in our stimuli. There was a marginally significant 2-way interaction between the three Auditory Targets and the three Onset Times ($\chi^2 = 8.73, p = .068$). Further, for the contrast comparing Speech to the other two targets (Tone and Tone-as-Speech targets), there was a significant 2-way interaction between the Auditory Targets (Speech vs. others) and the Onset Times ($\chi^2 = 8.49, p = .014$), suggesting Speech targets showed different Onset effects than the other two auditory targets. The 3-way interaction among the Auditory Targets, Onset Times and Target Durations was not significant ($\chi^2 = 6.77, p = .562$).

To understand the nature of the interactions, we compared pairs of Auditory Target conditions, in each case testing whether Auditory Target interacted with Onset Times. There was a significant 2-way interaction between Speech vs. Tone targets and Onset Times ($\chi^2 = 9.71, p = .008$), so we further tested Speech and Tone conditions individually. The Tone target model showed a significant effect of Onset Times ($\chi^2 = 26.01, p < .000$), with lower probability of responding short in the Late compared to Early onset times ($B = -0.32$, $SE = 0.06$, $p < .000$), as well as the On-time compared to Early onset times ($B = -0.20$, $SE = 0.05$, $p = .0002$). The model conducted on Speech did not show an effect of Onset Times ($\chi^2 = 2.85, p = .240$). Neither Onset Time contrast showed a significant difference. These results suggest a stronger entrainment effect (i.e. significant effect of Onset Times) in the Tone condition than in the Speech condition.

Auditory grouping: The Tone-as-Speech condition

We continued by comparing the speech-as-tone targets to the other two target conditions. We did not find robust cross-domain phase entrainment from tones to speech sounds, consistent with our preliminary work (Supplementary Section 1). One possible reason for the lack of tone-to-speech entrainment could be the auditory dissimilarity between tone precursors and speech targets (auditory grouping hypothesis). If speech categories *do* undergo entrainment as long as auditory properties match between precursors and targets, then Tone-as-Speech should look *similar to the Tone condition* (indicated by *nonsignificant* interaction effect) and should show significant effects of Onset Times. We first compared the Tone-as-Speech and the Tone conditions. There was no significant interaction between the Tone/Tone-as-Speech targets and Onset Times ($\chi^2 = 0.19, p = .911$). This is consistent with the nonsignificant contrast x Onset Time interaction in the full model Helmert contrast between these two conditions. We then compared the Tone-as-Speech target to the Speech target. There was a significant interaction between the Speech/Tone-as-Speech target and Onset Times ($\chi^2 = 8.12, p = .017$). These findings suggest that the responses to the Tone-as-Speech target were more similar to Tone target responses, and dissimilar to Speech target responses.

The model conducted on the Tone-as-Speech condition alone showed a significant effect of the Onset Times ($\chi^2 = 22.20, p < .000$), with lower probability of responding short in the Late compared to the Early onset times ($B = -0.31$, $SE = 0.06$, $p < .000$), and the On-time compared to the Early onset times ($B = -0.18$, $SE = 0.06$, $p = .002$). These results imply that the Tone-as-Speech condition patterns with the Tone condition, with both showing phase entrainment effects.

This might be taken as evidence that speech, as represented by tone targets heard as speech, can undergo phase entrainment, but how well did listeners process these tone targets as speech? To check how successful listeners were at processing amplitude-modulated tones as speech in the Tone-as-Speech condition, debriefing questions asked whether they perceived/felt the tones as being *lab* and *lap*. Only 14 participants said yes, an unexpectedly small proportion. Still, as planned, we tested their responses alone, and found a significant effect of Onset Times ($\chi^2 = 7.068, p = .029$), with significantly lower probability of responding short in the Late compared to the Early onset times ($B = -0.36$, $SE = 0.12$, $p = .004$), despite the small *n*. There was a marginally significant difference between the Early vs. On-time onset times ($B = -0.17$, $SE = 0.10$, $p = .079$). Because this small sample size was not foreseen, we explored the "speech hearers" effect by collecting additional participants who reported successfully hearing speech sounds as requested in the experiment instruction (total $n = 26$). These data continued to support the phase entrainment effect (see Supplementary 2.5 for details).

Discussion

We found significant phase effects, shown as entrainment distortion, from entrainer tones to target tones, yet we did not observe significant phase effects in target words *lab/lap*. Results were consistent in logistic regression and supported by ANOVA both in proportion short and 50% point (Supplementary Section 2). This finding is, on its surface, consistent with a domain-specific view of language processing and inconsistent with the general auditory processing account. The lack of tone-to-speech phase entrainment replicated our preliminary studies and Bosker and Kösem's (2017) null result for phase effects on speech stimuli. Note that Bosker and Kösem (2017) did not have a tone-target control condition that matched the word targets as we did, but we still did not observe phase effects on speech sound perception.

The lack of phase entrainment for speech stimuli is particularly interesting in the context of other findings that tone precursors *do* affect spectral (Holt, 2005, 2006; Lotto et al., 2003) and durational (Bosker, 2017; Wade & Holt, 2005) aspects of speech perception (summarized in Table 1), as well as neural evidence on cross-domain prosodic priming from music to speech (Sun et al., 2024). The lack of tone-to-speech phase effects is even more surprising given that studies have demonstrated cross-modal or supra-modal entrainment effects, for instance, between auditory and visual modalities (ten Oever & Sack, 2015; Zoefel & VanRullen, 2017; and a review by Bauer et al., 2020).

Table 1
Effects of nonspeech precursors on perception of speech sounds, with example studies.

Nonspeech precursor property	Affects speech targets
Spectrum	Yes (Holt 2005, 2006; Lotto et al., 2003)
Rate	Yes (Bosker, 2017; Bosker and Kösem, 2017; Wade & Holt, 2005)
Phase	No (Bosker and Kösem, 2017; our Experiment 1)

Within Experiment 1, we probed one possible reason for the lack of cross-domain phase entrainment in our Tone-as-Speech condition, that is, the auditory dissimilarity between the precursors and the target. On this auditory grouping account, phase effects from entrainment may occur in a wide range of modalities, including speech, but are limited by degree of acoustic match. Therefore, auditory dissimilarity may hinder the phase effect of the precursors on the targets because sequential temporal relationships (of which phase is one type) are harder to apprehend between dissimilar sounds (e.g., Bregman, 1990; Miller & Heise, 1950). The outcome of our Tone-as-Speech condition provided an initial attempt to test this hypothesis. We found entrainment effects for speech categorization even if listeners were asked to categorize *tone* targets as speech categories (the word *lap* or *lab*). Using tones as both precursors and targets yields higher acoustic similarity between the two, and may permit entrainment effects while categorizing stimuli as speech. However, given that there is only a small number of subjects who can actually hear the tones as words, additional research needs to be done to verify that auditory similarity modulates entrainment effects, a point we return to in Experiment 3.

A second possible reason for lack of speech entrainment relates to the issue of p-center of Tone and Word onsets–that is, the perceived onset time, which can differ from acoustic onset time (e.g. Morton et al., 1976). Even though we controlled tone targets to match the durations and amplitude envelopes of the word targets, it is still possible that the *subjectively* perceived onsets of tones and words are not identical. Specifically, the entrainment distortion could have been reduced in the Speech condition if the perceived onset times of words were later than those of the tones. Namely, speech stimuli *lap/lab* in the early, on-time, late conditions may have been perceived more as on-time, late, late, effectively removing the "early" effect and thus weakening entrainment distortion. This is addressed in our Experiment 2 by selecting speech stimuli which have p-centers more aligned with the tones. A control study (N = 12) was conducted within our lab to estimate the p-center of both *lab/lap* and a new vowel-initial word pair *add/at*, as well as the *lab/lap*-based tones. Only the endpoint stimuli from each condition were tested. In each trial of a beat alignment perception test, listeners heard trains of 10 single woodblock hits and 10 repetitions of a word or tone target (mixed to a single channel), with matched IOI (i.e. rate) but varying phase relationships, with the woodblock hits at 0–120 ms phase lag in 30-ms increments. Subjects were asked to judge whether the timing of the words/tones were aligned with the metronomes or not. Results suggested that *lab/lap* indeed had a later and possibly wider p-center than the tones (see Danielsen et al., 2019 on wider p-centers), with higher "beat aligned" judgments at greater degrees of target earliness (see the solid light gray vs. dotted dark gray curves in Fig. 5). This was somewhat surprising to us, as we had controlled speech and tones to have the same amplitude envelopes, which have been reported as critical to the p-center in speech (e.g. Šturm & Volín, 2016; though see Cooper et al., 1986). However, *add/at* (dashed black curve in Fig. 5) had a p-center similar to the tones, both of which were perceived around the sound onset. In Experiment 2, we investigated the cross-domain phase entrainment using an *add/at* speech continuum and new *add/at*-matched tones, so that the p-centers of the Speech and the Tone conditions were similarly early.

Experiment 2

Methods

Participants

We tested 200 participants in the manner described for Experiment 1. The sample size and exclusion criteria were identical to Experiment 1, resulting in final sample sizes of n = 71 for the Speech condition and n = 70 for the Tone condition. Participants were excluded for: scores below 90% for catch trials, suggesting inattentiveness (n = 40); reporting intrusive environmental noise (n = 5); reversed or zero identification slopes (n = 4); aberrant response patterns in 50% point data (n = 10).

Stimuli

The word pair *add* vs. *at* was chosen for Experiment 2 to better match the p-center between speech sounds and tones. The p-center of *add* and *at* was found to be around the sound onset, and thus better aligned to the p-center of tones based on our control study. Recordings were made by the same speaker as Experiment 1, using Praat 6.1.16 (Boersma & Weenink, 2020) in a quiet room. The stimuli were constructed in a similar manner (gradually lengthened "at") as in Experiment 1, with the only difference being how the envelope was extracted to construct the Tone condition. Instead of using the upper half of the envelope, we used the average envelope across the upper and lower half to generate tones matching the speech sounds. Please note that there is a strong correlation between the upper envelope and the lower half of the envelope (r = -0.91), as well as with the averaged envelope. Therefore, the choice between using the upper or lower half is likely inconsequential. However, for the sake of maximizing similarity to the speech signal, we opted to utilize the averaged envelope.

Design and procedure

The design and procedure closely mirrored Experiment 1 with the following adjustments. First, we exclusively tested the Speech condition and Tone condition. Second, we aligned the trial number and procedure across both conditions to mitigate any procedural disparities that might influence the results. Participants completed 240 pretest trials of identifying *at/add* (Speech condition) or *short/long* (Tone condition) in the full 8-step continua, followed by 288 main trials. After the main task, they undertook 120 trials discerning six word pairs (Speech condition) or six corresponding tone pairs (Tone condition). The planned analyses remained consistent with those of the first experiment.

Results

Following the same procedure as Experiment 1, we used logistic mixed-effects models to test whether there is cross-domain phase entrainment from context tones to words and tones. Results are presented in Figs. 6 and 7.

We first confirmed that there was a strong decrease in "short"/"*at*" responses with increasing target durations in our stimuli in the full model, demonstrated by the effect of Target Duration ($\chi^2 = 258.05$, $p <$ .000). The 2-way interaction between the Auditory Targets and the Onset Times was significant ($\chi^2 = 7.96$, $p = .02$). The two-way interaction suggests a stronger entrainment effect for tones than for speech targets. To understand the two-way interaction, we then examined Speech and Tone conditions individually. Consistent with Experiment 1, the model conducted on the Tone condition showed a significant effect of Onset Times ($\chi^2 = 43.97$, $p <$ .000), with a significantly lower probability of responding short or "*at*" for the Late compared to the Early onset times (B = −0.29, SE = 0.04, $p <$ .000) and the On-time compared to the Early onset times (B = −0.18, SE = 0.04, $p <$ .000). Unlike Experiment 1, the model conducted on Speech also showed an effect of Onset Times ($\chi^2 = 7.61$, $p = .022$) with a significantly lower probability of responding short or "at" at the Late compared to the Early onset times (B = −0.12, SE = 0.04, $p = .006$). No significant difference was found between the Early and On-time onset times (B = −0.07, SE = 0.04, p = .116).

Note that in the full model, there was also a 3-way interaction between Auditory Targets, Onset Times, and Target Durations ($\chi^2 = 13.22$, $p = .001$). This three-way interaction (see also Experiment 3) implies differences in curve shape, most likely the more asymmetric effects of Onset Time for speech targets (mostly in the shorter durations) but symmetric effects for tone targets. We briefly return to this in Experiment 3.

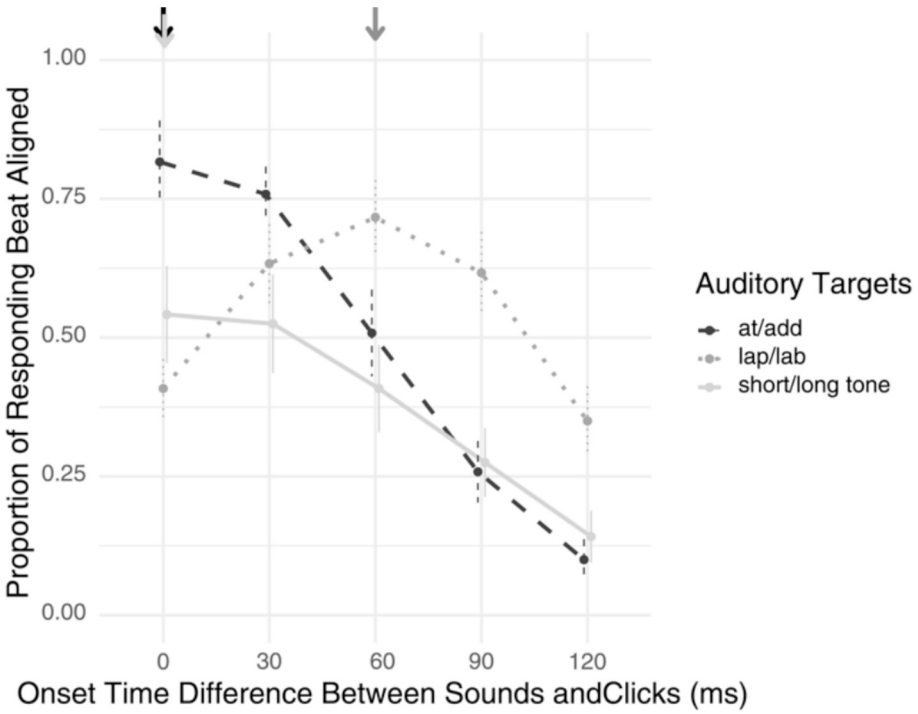

Fig. 5. Control study response curves averaged across 12 participants. The x-axis represents the perceived start of sound (in ms) of words/tones according to the metronome clicks. The y-axis represents the proportion of responding "aligned" between beat and words/tones averaged across all subjects for short and long tones (in solid light gray), word "*lap*" and "*lab*" (in dotted dark gray), and word "*at*" and "*add*" (in dashed black). The arrows on top indicate the perception onsets of different sounds.

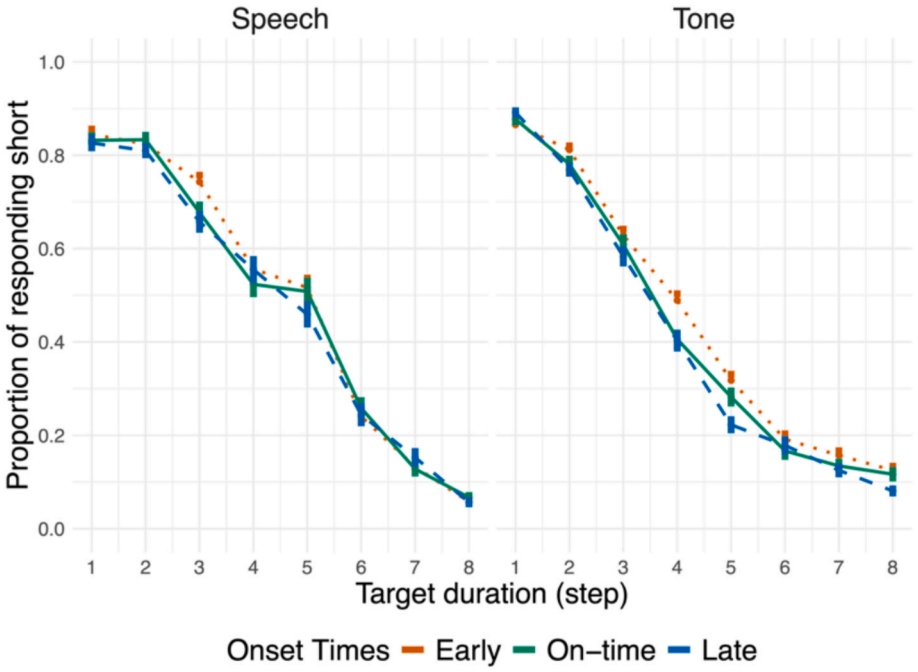

Fig. 6. Response curves averaged across subjects for each Auditory Target in Experiment 2. The x-axis represents step duration from shorter to longer sounds. The y-axis represents the proportion of responding short (or "*at*") averaged across all subjects for targets presented early (red dotted lines), on-time (green solid lines) and late (blue dashed lines). Each vertical bar represents standard errors across subjects.

Discussion

After aligning the p-center of words and tones, we observed more evidence of cross-domain phase entrainment from precursor tones to target words *add/at* compared to Experiment 1. This cross-domain entrainment effect on speech perception achieved statistical significance in the logistic regression and ANOVA proportion short responses (Supplementary Section 3). At first glance, the significant Onset Times in the Speech condition may suggest a different result than Experiment 1, hinting that entrainment is domain-general and affects both target

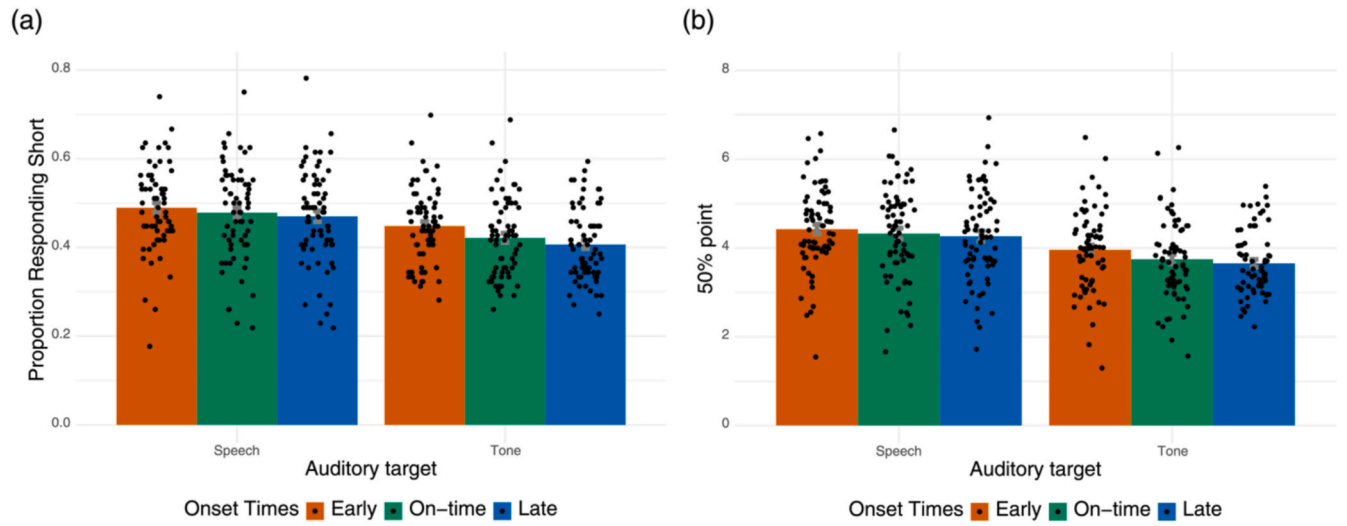

Fig. 7. Bar plots of proportion short and 50% point for each Onset Time among Auditory Targets in Experiment 2. The x-axis represents different auditory targets. The y-axis represents (a) the proportion of responding short and (b) 50% points for targets presented early (red bars), on-time (green bars) and late (blue bars). Bar plots show the group means and standard errors. Each black dot represents one individual.

words and target tones as long as the p-center is equated. However, the magnitude of the effect was significantly greater for tones than for speech, suggesting that, even equating p-center, speech may be less susceptible to entrainment. Further, the response curves of the Speech targets and the Tone targets showed distinct shapes (Fig. 6), hinting that despite speech sounds and tones both being subject to tone entrainment, they may have been processed differently. For example, listeners may compute durations for tone targets only with respect to recent memory traces (in the task itself), while they compute durations for speech targets with respect to long-term speech category knowledge. On the other hand, perhaps speech target phase effects are hindered by the auditory dissimilarity to the tone entrainers we used here, as on our auditory grouping hypothesis.

To fully compare cross-domain phase entrainment between nonspeech and speech sounds and to probe our auditory grouping hypothesis, we conducted a final experiment using the same word and tone targets, but replaced the tone precursors with speech precursors. This final experiment mirrored Experiment 2 and tested entrainment effect from speech sounds to target words/tones. According to the auditory grouping hypothesis, entrainment should be stronger when the target is acoustically similar to the precursor. Thus, we expected a stronger entrainment effect for speech targets (acoustically similar to the speech precursors) and a weaker effect for tone targets (acoustically dissimilar to the speech precursors) in Experiment 3 compared to Experiment 2.

Experiment 3

Methods

Participants

We tested 168 participants in a manner identical to Experiment 2. The sample size and exclusion criteria were identical to Experiment 1 and 2, resulting in final sample sizes of n = 50 for the Speech condition and n = 57 for the Tone condition. Participants were excluded for: scores below 90% on catch trials, suggesting inattentiveness (n = 31); reporting intrusive environmental noise (n = 5); reversed or zero identification slopes (n = 8); aberrant response patterns in 50 % point data (n = 19).

Stimuli

The same word pair *add* vs. *at* used in Experiment 2 was also used in Experiment 3. Importantly, we used *speech precursors* instead of *tone precursors* to test the phase entrainment effects from speech sounds to speech/nonspeech sounds, mirroring Experiment 2′s tone precursors. The stimuli were otherwise constructed in the same manner as in Experiment 2. The speech precursors were 60-ms "speech pips", edited from the first 60 ms of the word "*at*", recorded by the same speaker as Experiment 1 in a quiet room. In a small control study (N = 11) using the same procedure described in Section 2.3, the p-center of *add* and *at* and the speech precursors was found to be around the sound onset, similar to those of the tone targets and precursors (Fig. 8).

Design and procedure

The design and procedure were identical to Experiment 2.

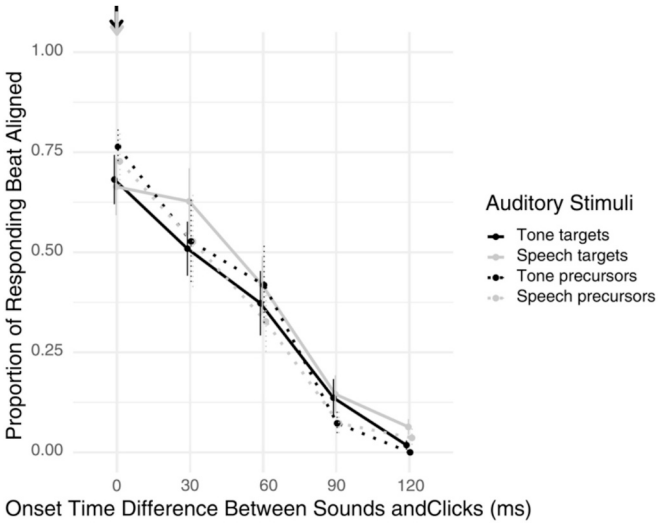

Fig. 8. Control study II response curves averaged across 11 participants. The x-axis represents the perceived start of sound (in ms) of words/tones according to the metronome clicks. The y-axis represents the proportion of responding "aligned" between beat and the tone targets (solid black) and the speech targets (solid gray) used in both Experiment 2 and 3, as well as the tone precursors in Experiment 2 (dotted black) and speech precursors in Experiment 3 (dotted gray). The arrows on top indicate the perception onsets of different sounds.

Results

Results are summarized in Figs. 9 and 10. Following the same procedure as in Experiments 1 and 2, we first confirmed that there was a strong decrease in "short"/"*at*" responses with increasing target durations (χ^2 = 186.05, p < .000). The 2-way interaction between the Auditory Targets and the Onset Times was not significant (χ^2 = 0.42, p = .809). Consistent with Experiment 1 and 2, the model conducted on the Tone condition showed a significant effect of Onset Times (χ2 = 15.98, p = .000), with a significantly lower probability of responding short or "*at*" for the Late compared to the Early onset times (B = −0.22, SE = 0.05, p < .000) and the On-time compared to the Early onset times (B = −0.10, SE = 0.05, p = .047). The model conducted on the Speech condition also showed an effect of Onset Times (χ2 = 12.44, p = .002) with a significant difference between the Early and Late onset times (B = −0.20, SE = 0.06, p = .000), and a marginal difference between the Early and On-time onset times (B = −0.08, SE = 0.05, p = 0.09). In the full model, the 3-way interaction among the Auditory Targets, the Onset Times, and the Target Durations just reached significance (χ^2 = 6.12, p = .047). As in Experiment 2, this seems to result from the more asymmetric effects of onset time on speech vs. more symmetric effects on tone, hinting that despite the indistinguishable overall effects of onset time, speech categorization and tone duration categorization are computed somewhat differently. Our favored explanation is that speech categorization relies on long-term representations whereas tone categorization relies on recently learned (short-term) representations, but other explanations are possible.

Discussion

This final experiment demonstrated significant phase entrainment effects from speech precursors to both target words and tones. The logistic regression and ANOVA results (Supplementary Section 4) suggest that speech precursors can exert cross-domain phase entrainment effects on tone targets, mirroring the findings from Experiment 2, which demonstrated a cross-domain effect from tone precursors to speech targets. Together, these two experiments provide evidence for cross-domain phase entrainment under specific conditions, namely when the p-centers of the speech stimuli are controlled (Fig. 8). In addition, we replicated the phase entrainment effect on speech sounds (Experiment 2). Further, moving from tone precursors (Experiment 2) to speech precursors that were acoustically more similar to the speech targets (Experiment 3) led to a larger effect size for speech entrainment, which is consistent with the auditory grouping hypothesis. See General Discussion effect size comparisons for further elucidations. As observed in Experiment 2, there were some subtle differences between the Tone and the Speech curves (Fig. 9). Across the two experiments, the effects of Onset Times on the Speech conditions happened more on Target Durations 2, 3 and 4 (i.e. shorter end of the "*at/add*" continuum, more "*at*" responses), while it was more evenly spread out cross the full continuum for the Tone condition. These nuances surfaced in our logistic mixed-effects regression as significant 3-way interactions in both Experiment 2 and 3. It is interesting that this finding persists across two studies and two different precursors; future studies should replicate this effect with different stimuli. Nonetheless, we find clear evidence in both studies that precursor phase relationships affect duration perception.

General discussion

We asked whether there is cross-domain phase entrainment between tones and speech, with tone-to-tone and speech-to-speech entrainment conditions included for comparison. We consistently observed phase entrainment effects with amplitude-modulated tones. This in itself was not a foregone conclusion, as previous research often used tones with relatively sudden onsets, whereas we used tones with more gradual onsets that closely matched our speech stimuli. The picture for speech entrainment was more complex. We did not observe significant phase entrainment effects from precursor tones to target words *lab/lap* (Experiment 1). However, we did observe phase entrainment effects from precursor tones to target words *add/at* (Experiment 2) in which we aligned the p-centers of the speech targets and tone targets. We also observed phase entrainment effects from speech precursors to speech targets (Experiment 3).

To succinctly summarize effect magnitude across all conditions in all

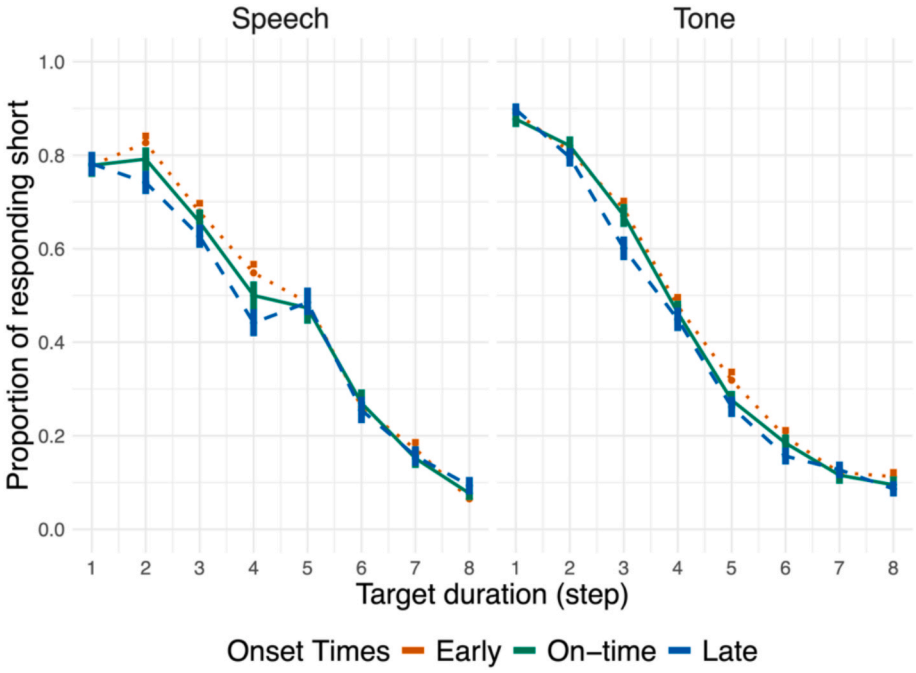

Fig. 9. Response curves averaged across subjects for each Auditory Target in Experiment 3. The x-axis represents step duration from shorter to longer sounds. The y-axis represents the proportion of responding short (or "*at*") averaged across all subjects for targets presented early (red dotted lines), on-time (green solid lines) and late (blue dashed lines). Each vertical bar represents standard errors across subjects.

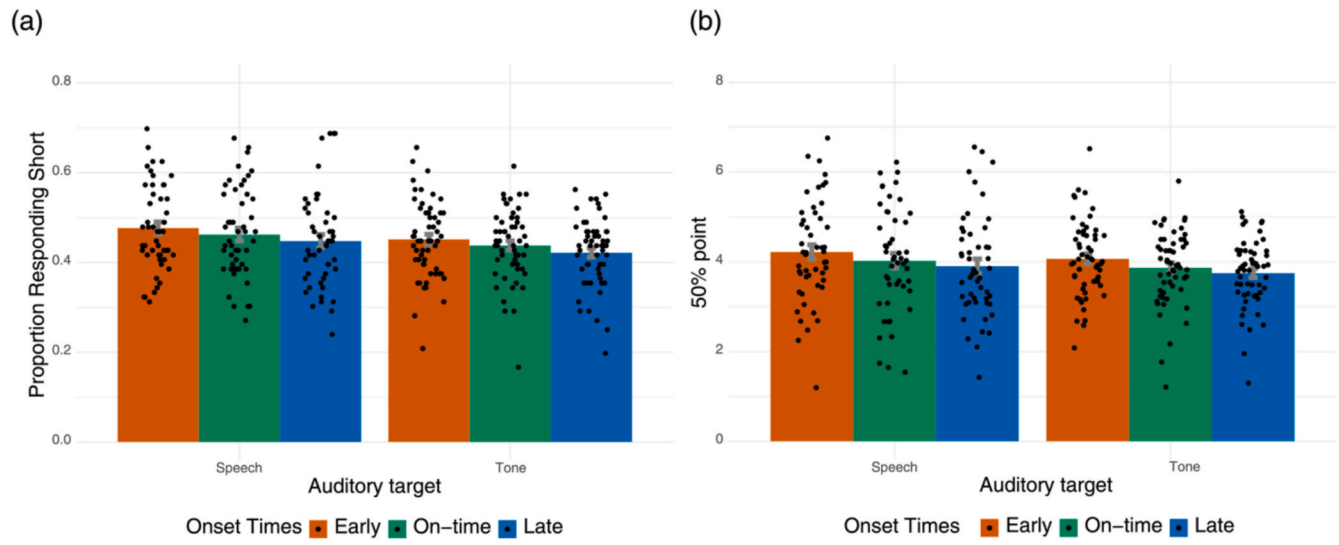

Fig. 10. Bar plots of proportion short and 50% point for each Onset Times among Auditory Targets in Experiment 3. The x-axis represents different auditory targets. The y-axis represents (a) the proportion of responding short and (b) 50% points for targets presented early (red bars), on-time (green bars) and late (blue bars). Bar plots show the group means and standard errors. Each black dot represents one individual.

experiments, we used the conceptually largest possible Onset Time effect, the beta value accounting for the difference between Early and Late onset times for each condition.[1] Across the first two experiments, the tone-to-tone entrainment (B ranging from −0.29 to −0.32) showed a larger effect size than tone-to-speech entrainment (−0.08 to −0.12). Moreover, these tone-to-tone effects also surpassed the speech-to-speech effect in Experiment 3 (−0.20), which was similar to the cross-domain and acoustically dissimilar speech-to-tone effect (0.22). It is notable that of the cross-modal conditions, the speech-to-tone effect was numerically larger than the tone-to-speech effects in Experiments 1–2. It is also notable that the largest magnitude effect on speech targets was the speech-to-speech condition in Experiment 3, though these are descriptive numerical comparisons and not statistical tests. These results are summarized in Table 2 and Supplementary Section 5.

At this point we feel safe in concluding that phase entrainment can occur for speech. Still, overall effects of phase appear weaker for speech than for nonspeech tones. Our perspective is that a combination of factors conspires to increase entrainment effects in tones, or inversely, to decrease entrainment effects in speech. A potential factor already discussed is auditory similarity between precursors and targets. We view the differences in outcome between Experiment 2 (tone precursors) and Experiment 3 (speech precursors) as support for auditory similarity effects. Future studies should also explore how the degree of auditory similarity affects entrainment. For example, based on the auditory grouping hypothesis, two *dis*similar types of nonspeech should not show entrainment, such as noise burst precursors and tone targets (or vice versa), or instrumental sounds with distinguishable timbres and or pitch heights. We discuss other potential differences between tones and speech below.

One factor that might affect phase entrainment is the top-down knowledge of whether one is listening to tones versus speech sounds, which might shift the listener's perception of the target. This view is aligned with previous findings (e.g. Pitt et al., 2016 Experiment 3),

Table 2
Cross-domain entrainment effect, indicated by a significant effect of Onset Times and the difference between Early and Late onset times across the three experiments. The statistically significant effects ($p < .05$) are highlighted in bold.

		Speech	Tone	Tone-as-Speech
Experiment 1: tone precursors + tone/words (*lab/lap*)	Effect of Onset Times	$\chi^2 = 2.85$, $p = .240$	**$\chi^2 = 26.01$, $p < .000$**	**$\chi^2 = 22.20$, $p < .000$**
	Early vs. Late	B = −0.08, SE = 0.06, $p = 0.221$	**B = −0.32, SE = 0.06, $p < .000$**	**B = −0.31, SE = 0.06, $p < .000$**
Experiment 2: tone precursors + tone/words (*add/at*)	Effect of Onset Times	**$\chi^2 = 7.61$, $p = .022$**	**$\chi^2 = 43.97$, $p < .000$**	.
	Early vs. Late	**B = −0.12, SE = 0.04, $p = .006$**	**B = −0.29, SE = 0.04, $p < .000$**	.
Experiment 3: speech precursors + tone/words (*add/at*)	Effect of Onset Times	**$\chi2 = 12.44$, $p = .002$**	**$\chi2 = 15.98$, $p = .000$**	.
	Early vs. Late	**B = −0.20, SE = 0.06, $p = .000$**	**B = −0.22, SE = 0.05, $p < .000$**	.

which showed that the ambiguous sinewave speech precursors only had a context rate effect on speech perception when the sinewave speech was heard as speech. We attempted to manipulate top-down perception with our Tone-as-Speech condition in Experiment 1 and found that even when instructed to hear tones as speech, listeners showed entrainment. However, as we discussed earlier, this could also be explained by the auditory grouping hypothesis, suggesting that phase entrainment occurs when the precursors and targets have similar acoustic properties, regardless of whether participants were primed to listen to words or tones. Also problematic for a "top-down" explanation, our Tone-as-Speech condition left open the degree to which listeners actually processed tone targets in "speech mode," thus leaving unclear whether top-down perception can affect entrainment. Future research might investigate more effective top-down manipulations where listeners are cued to interpret the same ambiguous stimulus as either speech or nonspeech. Another factor that might affect cross-domain entrainment is the degree of alignment of subjectively perceived event onset time between tone and speech. Experiments 2 and 3 both showed that after controlling for

[1] Throughout the study, as in our previous work (Cheng & Creel, 2020), the early condition shows the strongest effects vs. the other two, such that the smaller-magnitude on-time vs. late difference on its own does not always reach the threshold for significance (See Supplementary Materials 2.3, 3.3, 4.3 for a summary of pre-registered ANOVAs and t-tests). This asymmetry, however, is not limited to speech stimuli and does not bear on the general question of sensitivity to entrainment effects.

p-center, speech targets are detectably susceptible to entrainment from both tones (Experiment 2) and speech (Experiment 3). Thus, speech appears at least mildly susceptible to phase entrainment under certain contrived conditions. While this limited susceptibility might seem to suggest some degree of modularity of speech and nonspeech sounds, it is quite possible that nonspeech sounds too are also only susceptible under limited conditions, and phase entrainment research simply has not explored sufficiently naturalistic nonspeech materials yet, for example expressively-timed music.

Theoretical implications

In the broader context, a novel contribution of our study is the focus on the phase aspect of entrainment, which is the key distinction between two classic models of how humans perceive time duration and onsets: entrainment models and interval models (McAuley & Jones, 2003). Our findings of the weaker effect of tone-to-speech entrainment and the apparent impact of auditory similarity appear to specifically apply to the *phase,* rather than the *rate,* of the precursors (Bosker, 2017; Wade & Holt, 2005). Taking our findings and previous related work as a whole (see Table 1), the important question is why speech sound perception appears to be influenced more by the rate and spectral features of the tone precursor than by the precursor's phase.

One possibility is that temporal cues, including duration estimation and onset prediction, are inherently represented and processed differently in speech and nonspeech domains. This maps onto a much-researched distinction between entrainment timing (or beat-based timing) and interval timing (or memory-based timing), where interval timing refers to recognition of exact duration without respect to phase relationship or context/precursors (McAuley & Jones, 2003). In invoking *a priori* processing differences between speech and nonspeech, this explanation falls closest to a modular account of timing perception. Assuming there is an actual distinction between these two timing systems (Teki et al., 2011; Breska & Ivry, 2018, but see Teki et al., 2012; Rimmele et al., 2018), it may be that musical/nonspeech sounds are subject to both interval and entrainment timing, while duration perception in speech is mostly subject to interval timing that does not care much about the phase relationship. Admittedly, the isochronous precursors used in our design represent a special case of rhythm and temporal prediction, lying at the intersection of top-down and bottom-up processes of entrainment. These stimuli are useful for providing simplicity and precise experimental control over the phase relationship (the Onset Times in this study) between speech and nonspeech sounds. In contrast to isochronous beats, running speech and live music are typically quasi-rhythmic, yet they contain predictable cues that allow the brain to entrain and make temporal predictions. Some existing research suggests that phase-*like* effects may occur in speech segmentation (Dilley & McAuley, 2008). These rhythmic effects in speech may result from setting up expectations of repeating cycles of stress alternations. It is noteworthy that Dilley and McAuley's segmentation-inducing speech precursors are not isochronous, unlike our entraining tones and syllables. Future research may investigate the domain specificity or generality of entrainment using more natural speech materials or musical materials from styles with less isochrony, with an eye toward real-world relevance. A major open theoretical question is whether events that occur earlier or later than expected relative to a predictive but nonisochronous preceding sequence experience similar duration distortions. If so, phase effects could potentially apply in a much wider range of circumstances.

A related but theoretically distinct account is that speech and nonspeech sound perception may rely on different timing systems due to the outcome of a *learning mechanism*. Critical durational cues in speech sounds, such as vowel durations and voice onset times, are repeatedly heard and learned. This learning process may yield speech representations that contain durational information, and these activated, interval-type memories of speech duration may participate in the recognition process for speech in a way that does not occur for nonspeech stimuli. In addition, speech sound perception may rely more on rate information than phase information because phase is a very noisy cue in natural speech, and perceivers therefore learn to downweight phase information in speech perception. Thus, even though we used isochronous precursors, listeners who are categorizing speech may have already downweighted phase cues that phase cues have little impact on their perceptions of words. If this is the case, then manipulations of auditory similarity should not increase entrainment in speech, although training that causes listeners to upweight phase cues might do so. For example, listeners could receive extensive exposure to speakers who produce more regular IOIs, taking p-center into account of course, to determine whether this shifts speech entrainment sensitivity. Further, some languages contain duration contrasts (e.g. Japanese, Finnish, Arabic, Thai, Dinka) such that, for example, in Japanese, /kado/, /kaado/, and /kadoo/ are all different words (Hirata, 2004; see Remijsen & Gilley, 2008, on 3-way duration contrasts). It is also possible that such listeners may be generally more sensitive to duration, which could in principle heighten phase entrainment.

A final consideration is, if speech is susceptible to entrainment distortion as we observed in Experiments 2 and 3, is this likely to affect perception in real-world contexts? There is ample evidence for rate-based entrainment effects in speech perception (Baese-Berk et al., 2014; Bosker, 2017; Bosker & Kösem, 2017; Heffner et al., 2017; Wade & Holt, 2005; Steffman, 2021), but not for phase-based entrainment effects. Consequently, it is unclear how small-magnitude phase perturbations in speech functionally impact listeners' speech processing. While we observed robust phase effects in tone targets, the relatively weaker effects for speech raise further questions about to what extent and under what natural circumstances (and in what listeners) perception of nonspeech is affected by phase perturbations. The current study implies that there may be stronger nonspeech effects overall, but this may be the result of the simplified tone stimuli with simpler acoustic features. At the risk of repeating another stimulus that appears periodically, further research is needed.

Future directions

The various interpretations we have posited here (top-down knowledge, auditory grouping, learned phase downweighting) lead to testable hypotheses that should be examined in future studies. For example, studies could explore different techniques of generating targets that are ambiguous between speech and nonspeech, allowing for manipulation of top-down perception (such as degraded speech or sinewave speech as in Pitt et al., 2016). Additionally, in future testing, stimuli that optimize entrainment effects should be considered. The current study included a 600 ms gap (i.e. 2 cycles) after the precursors based on a classic duration judgment task used in our original studies (Cheng & Creel, 2020), which in turn were based on work by McAuley & Jones (2003). In hindsight, making the target sound continuous with the precursor would likely boost entrainment effects, allowing for more sensitive tests. Another valuable avenue would be to use neuroimaging methods to obtain objective measurements (not self-report) of perceptual distortion in tones versus in speech sounds, and further investigate the underlying phase resetting mechanisms (Rimmele et al., 2018). For example, M/EEG could test *when* entrainment occurs, whether during an earlier process such as perceptual encoding or a later process like phonological categorization (see Toscano et al., 2010 on voice onset time). In addition, fMRI could be used to test whether neural regions associated with the two posited timing systems (Breska & Ivry, 2018; Buhusi & Meck, 2005; Grondin, 2010; Teki et al., 2011), namely the cerebellar circuit (sensitive to interval timing) or basal ganglia circuit (sensitive to entrainment timing), were differentially activated during duration judgment in speech vs. nonspeech.

Conclusion

We tested whether phase entrainment effects occur for speech, both with cross-domain entrainers (tones) and within-domain entrainers (speech). We did not observe significant phase entrainment effects from precursor tones to target words *lab/lap*, but we did observe some phase entrainment effects in target words *add/at* with both speech and tone precursors. Still, across the three experiments, tone-to-tone entrainment showed larger effect sizes than within-domain speech-to-speech entrainment, or cross-domain tone-to-speech or speech-to-tone entrainment. These findings suggest that speech sounds are subject to cross-modal phase entrainment effects, though reduced in magnitude compared to the tone entrainment effect on tones. Our work contributes to the debate between domain-specific and domain-general accounts of speech processing by making an initial exploration of factors that may influence speech entrainment based on the relative *phase* between precursor sounds and target words. Taking into account the related literature on precursor effects, results suggest that speech can be affected by phase in addition to spectral and rate cues, but these effects tend to be smaller than for nonspeech tones, raising questions about the strength of the influence of phase entrainment in speech processing.

CRediT authorship contribution statement

Tzu-Han Zoe Cheng: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Sarah C. Creel:** Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2025.104730

Data availability

This project was pre-registered and the stimuli, data, analysis codes are posted on Open Science Framework (osf.io/r37ax).

References

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. (2014). Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science, 25*(8), 1546–1553.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bauer, A. K. R., Debener, S., & Nobre, A. C. (2020). Synchronisation of neural oscillations and cross-modal influences. *Trends in Cognitive Sciences, 24*(6), 481–495.

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

Boersma, P., & Weenink (2020). Praat: doing phonetics by computer [Computer program]. *http://www.praat.org/*.

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

Bosker, H. R., & Kösem, A. (2017). An entrained rhythm's frequency, not phase, influences temporal sampling of speech. In *Interspeech 2017* (pp. 2416-2420).

Bregman, A. B. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge: MIT Press.

Breska, A., & Ivry, R. B. (2018). Double dissociation of single-interval and rhythmic temporal prediction in cerebellar degeneration and Parkinson's disease. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 12283–12288. https://doi.org/10.1073/pnas.1810596115

Buhusi, C. V., & Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews. Neuroscience, 6*(10), 755–765.

Cheng, T. H. Z., & Creel, S. C. (2020). The interplay of interval models and entrainment models in duration perception. *Journal of Experimental Psychology. Human Perception and Performance, 46*(10), 1088.

Cooper, A. M., Whalen, D. H., & Fowler, C. A. (1986). P-centers are unaffected by phonetic categorization. *Perception & Psychophysics, 39*(3), 187–196.

Danielsen, A., Nymoen, K., Anderson, E., Câmara, G. S., Langerød, M. T., Thompson, M. R., & London, J. (2019). Where is the beat in that note? Effects of attack, duration, and frequency on the perceived timing of musical and quasi-musical sounds. *Journal of Experimental Psychology. Human Perception and Performance, 45*(3), 402–418. https://doi.org/10.1037/xhp0000611

de Graaf, T. A., Gross, J., Paterson, G., Rusch, T., Sack, A. T., & Thut, G. (2013). Alpha-Band Rhythms in Visual Task Performance: Phase-Locking by Rhythmic Sensory Stimulation. *PLoS One1, 8*(3), Article e60035. https://doi.org/10.1371/journal.pone.0060035

Diehl, R. L., Souther, A. F., & Convis, C. L. (1980). Conditions on rate normalization in speech perception. *Perception & Psychophysics, 27*, 435–443.

Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language, 59*(3), 294–311. https://doi.org/10.1016/j.jml.2008.06.006

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

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

FindingFive Team (2019). FindingFive: A web platform for creating, running, and managing your studies in one place. FindingFive Corporation (nonprofit), NJ, USA. https://www.findingfive.com.

Fiveash, A., Bedoin, N., Gordon, R. L., & Tillmann, B. (2021). Processing rhythm in speech and music: Shared mechanisms and implications for developmental speech and language disorders. *Neuropsychology, 35*(8), 771.

Fujii, S., & Wan, C. Y. (2014). The role of rhythm in speech and language rehabilitation: The SEP hypothesis. *Frontiers in Human Neuroscience, 8*, 777.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience, 15*(4), 511–517.

Goswami, U. (2012). Entraining the brain: applications to language research and links to musical entrainment.

Goswami, U. (2019). Speech rhythm and language acquisition: An amplitude modulation phase hierarchy perspective. *Annals of the New York Academy of Sciences, 1453*(1), 67–78.

Grondin, S. (2010). Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics, 72*(3), 561–582.

Haegens, S., & Zion Golumbic, E. (2018). Rhythmic facilitation of sensory processing: A critical review. *Neuroscience and Biobehavioral Reviews*, *86*(December 2017), 150–165. https://doi.org/10.1016/j.neubiorev.2017.12.002.

Harding, E. E., Kim, J. C., Demos, A. P., Roman, I. R., Tichko, P., Palmer, C., & Large, E. W. (2025). Musical neurodynamics. *Nature Reviews. Neuroscience*, 1–15.

Heffner, C. C., Newman, R. S., & Idsardi, W. J. (2017). Support for context effects on segmentation and segments depends on the context. *Attention, Perception, & Psychophysics, 79*, 964–988.

Henry, M. J., Herrmann, B., & Obleser, J. (2014). Entrained neural oscillations in multiple frequency bands comodulate behavior. *Proceedings of the National Academy of Sciences of the United States of America, 111*(41), 14935–14940. https://doi.org/10.1073/pnas.1408741111

Henry, M. J., Herrmann, B., & Obleser, J. (2016). Neural microstates govern perception of auditory input without rhythmic structure. *The Journal of Neuroscience, 36*(3), 860–871. https://doi.org/10.1523/JNEUROSCI.2191-15.2016

Hickok, G., Farahbod, H., & Saberi, K. (2015). The Rhythm of perception: Entrainment to Acoustic Rhythms Induces subsequent Perceptual Oscillation. *Psychological Science, 26*(7), 1006–1013. https://doi.org/10.1177/0956797615576533

Hillenbrand, J., Ingrisano, D. R., Smith, B. L., & Flege, J. E. (1984). Perception of the voiced–voiceless contrast in syllable-final stops. *The Journal of the Acoustical Society of America, 76*(1), 18–26.

Hirata, Y. (2004). Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics, 32*(4), 565–589.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science, 16*(4), Article 305–312.https://doi.org/10.1111/j.0956-7976.2005.01532.x.

Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *The Journal of the Acoustical Society of America, 120*(5), 2801–2817. https://doi.org/10.1121/1.2354071

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

Jones, M. R., Moynihan, H., MacKenzie, N., & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science, 13*(4), 313–319.

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

Kizuk, S. A. D., & Mathewson, K. E. (2017). Power and phase of alpha oscillations reveal an interaction between spatial and temporal visual attention. *Journal of Cognitive Neuroscience, 29*(3), 480–494. https://doi.org/10.1162/jocn_a_01058

Ladányi, E., Persici, V., Fiveash, A., Tillmann, B., & Gordon, R. L. (2020). Is atypical rhythm a risk factor for developmental speech and language disorders? *Wiley Interdisciplinary Reviews: Cognitive Science, 11*(5), e1528.

Lakatos, P., Gross, J., & Thut, G. (2019). A new unifying account of the roles of neuronal entrainment. *Current Biology, 29*(18), R890–R905.

Liberman, A. M., Harris, K. S., Kinney, J. A., & Lane, H. (1961). The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology, 61*(5), 379–388. https://doi.org/10.1037/h0049038

Lotto, A. J., Sullivan, S. C., & Holt, L. L. (2003). Central locus for nonspeech context effects on phonetic identification (L). *The Journal of the Acoustical Society of America, 113*(1), 53–56. https://doi.org/10.1121/1.1527959

Marcus, S. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception and Psychophysics, 30*, 247–256.

Mathewson, K. E., Fabiani, M., Gratton, G., Beck, D. M., & Lleras, A. (2010). Rescuing stimuli from invisibility: Inducing a momentary release from visual masking with pre-target entrainment. *Cognition, 115*(1), 186–191. https://doi.org/10.1016/j.cognition.2009.11.010

Mathewson, K. E., Prudhomme, C., Fabiani, M., Beck, D. M., Lleras, A., & Gratton, G. (2012). Making waves in the stream of consciousness: Entraining oscillations in EEG alpha and fluctuations in visual awareness with rhythmic visual stimulation. *Journal of Cognitive Neuroscience, 24*(12), 2321–2333. https://doi.org/10.1162/jocn_a_00288

Mattingly, I. G., Liberman, A. M., Syrdal, A. K., & Halwes, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology, 2*(2), 131–157. https://doi.org/10.1016/0010-0285(71)90006-5

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

McAuley, J. D., & Kidd, G. R. (1998). Effect of Deviations from Temporal expectations on Tempo Discrimination of Isochronous Tone Sequences. *Journal of Experimental Psychology. Human Perception and Performance, 24*(6), 1786–1800. https://doi.org/10.1037/0096-1523.24.6.1786

Miller, G. A., & Heise, G. A. (1950). The trill threshold. *The Journal of the Acoustical Society of America, 22*(5), 637–638.

Morton, J., Marcus, S., & Frankish, C. (1976). Perception centers (P-centers). *Psychological Review, 83*, 405–408.

Pitt, M. A., Szostak, C., & Dilley, L. C. (2016). Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate. *Attention, Perception, & Psychophysics, 78*, 334–345.

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews. Neuroscience, 21*(6), 322–334.

Port, R. F. (1979). The influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics, 7*(1), 45–56.

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America, 51*(4, Pt. 2), 1296–1303. doi:10.1121/1.1912974.

Remijsen, B., & Gilley, L. (2008). Why are three-level vowel length systems rare? Insights from Dinka (Luanyjang dialect). *Journal of Phonetics, 36*(2), 318–344.

Repp, B. H. (2002). Phase correction in sensorimotor synchronization: Nonlinearities in voluntary and involuntary responses to perturbations. *Human Movement Science, 21*, 1–37. https://doi.org/10.1016/S0167- 9457(02)00076-3

Revoile, S., Pickett, J. M., Holden, L. D., & Talkin, D. (1982). Acoustic cues to final stop voicing for impaired-and normal-hearing listeners. *The Journal of the Acoustical Society of America, 72*(4), 1145–1154.

Rimmele, J. M., Morillon, B., Poeppel, D., & Arnal, L. H. (2018). Proactive sensing of periodic and aperiodic auditory patterns. *Trends in Cognitive Sciences, 22*(10), 870–882.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association, 88*(424), 1273–1283.

Spaak, E., de Lange, F. P., & Jensen, O. (2014). Local entrainment of alpha oscillations by visual stimuli causes cyclic modulation of perception. *The Journal of Neuroscience, 34*(10), 3536–3544. https://doi.org/10.1523/JNEUROSCI.4385-13.2014

Steffman, J. (2021). Rhythmic and speech rate effects in the perception of durational cues. *Attention, Perception, & Psychophysics, 83*(8), 3162–3182.

Šturm, P., & Volín, J. (2016). P-centres in natural disyllabic Czech words in a large-scale speech-metronome synchronization experiment. *Journal of Phonetics, 55*, 38–52. https://doi.org/10.1016/j.wocn.2015.11.003

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology. Human Perception and Performance, 7*(5), 1074.

Sun, M., Xing, W., Yu, W., Slevc, L. R., & Li, W. (2024). ERP evidence for cross-domain prosodic priming from music to speech. *Brain and Language, 254*, 105439.

Teki, S., Grube, M., Kumar, S., & Griffiths, T. D. (2011). Distinct neural substrates of duration-based and beat-based auditory timing. *The Journal of Neuroscience, 31*, 3805–3812. https://doi.org/10.1523/JNEURO- SCI.5561-10.2011

Teki, S., Grube, & M., Griffiths, T. D. (2012). A unified model of time perception accounts for duration-based and beat-based timing mechanisms.

Ten Oever, S., & Martin, A. E. (2021). An oscillating computational model can track pseudo-rhythmic speech by using linguistic predictions. *Elife, 10*, Article e68066.

Ten Oever, S., & Sack, A. T. (2015). Oscillatory phase shapes syllable perception. *Proceedings of the National Academy of Sciences, 112*(52), 15833–15837.

Ten Oever, S., Titone, L., Te Rietmolen, N., & Martin, A. E. (2024). Phase-dependent word perception emerges from region-specific sensitivity to the statistics of language. *Proceedings of the National Academy of Sciences, 121*(23), Article e2320489121.

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science, 21*(10), 1532–1540.

Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception and Psychophysics, 67*(6), 939–950. https://doi.org/10.3758/BF03193621

Wolf, C. G. (1978). Voicing cues in English final stops. *Journal of Phonetics, 6*(4), 299–309.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron, 77*(5), 980–991.

Zoefel, B., & VanRullen, R. (2017). Oscillatory mechanisms of stimulus processing and selection in the visual and auditory systems: State-of-the-art, speculations and suggestions. *Frontiers in Neuroscience, 11*, 1–13. article 296.