Entrainment to Speech Rhythm from Perception to Production

Alistair Beith

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Neuroscience and Psychology
College of Science and Engineering
University of Glasgow

June 2024

Abstract

Human speech comprehension requires a highly coordinated engagement of linguistic processes. Speech contains complex layers of information to be decoded and often acted on just-in-time. This thesis explores the possible role of speech rhythm in connecting the speech we hear to the speech we produce. Two methodological chapters are presented. The first reviews and evaluates different approaches to creating rhythmic speech stimuli. The second highlights the challenges of collecting precise and accurate auditory response times in web-based experiments. These methods are then applied to a series of three experiments, each building on the last. In a new experimental paradigm, participants responded verbally to simple maths sums where tempo and rhythm were manipulated. Both time-domain and frequency-domain analyses found effects of stimulus tempo on the timing of responses.

Acknowledgements

Thank you to Rachel for introducing me to the topic of speech rhythm, encouraging me to consider a career in research, and for sharing an encyclopedic knowledge.

Thank you to Dale for always being there to accept any methodological, statistical, or programming challenge. I am particularly grateful to you for knowing when to encourage me to explore new ideas, and when to point me back on course. Thank you, also, for developing the "math sum" paradigm, I hope I did more than add an "s".

Loïs, I likely wouldn't have finished my undergraduate degree if I wasn't trying to impress you. Thank you for always pushing me and picking me up when necessary. I am also indebted to my family, en mijn schoonfamilie, for their love and support.

Thank you to Sarune, Jack, Alex, and Jinghui for being officemates and friends over the years.

I am grateful to Rory Ferreira, Michael Eagle, Daniel Dumile, David Cohn, and Jonathan Wolf for daily reminders of why rhythm matters, and to Kenny Segal for not quantising his drums.

Contents

List of Figures

List of Tables

Declaration

All work in this thesis was carried out by the author unless otherwise explicitly stated.

Chapter 1

General Introduction

Rhythm binds us together. Whether through song, dance, or simply co-navigating the physical world, human behaviours are connected by pulses and patterns. This thesis is concerned with how the rhythm of heard speech influences the timing of spoken responses. Specifically, it asks if speaker and listener as joint-actors are *entrained* to a shared temporal frame of reference. Firstly, this would require that speech has, or can have, rhythm. Secondly, that listeners are able to entrain to this rhythm. Thirdly, that the perceived rhythm affects production. Finally, that if present, the relationship between produced and perceived rhythm can be detected by an experimental paradigm. The following literature review addresses these questions in turn.

1.1 Literature review

1.1.1 Is speech rhythmic?

In music, rhythm can be quantified as patterns of intervals constrained to sub-divisions of units defined by a tempo and time signature. Although this definition imposes temporal characteristics of western musical notation on music that may not necessarily be present (Seeger, 1958), there is a strong cross-cultural tendency towards producing musical rhythms with integer-ratio interval relationships (Jacoby & McDermott, 2017). In contrast, attempts to define rhythm in speech have had a long and controversial history.

Speech rhythm is often used to refer to perceived language classes. These include *syllable-timed* and *stress-timed* classifications (Abercrombie, 1967), where the interval between syllables or stressed syllables in a language would be expected to be regular. However, this regularity, termed isochrony, is rarely found in the acoustic speech signal (Arvaniti, 2009; Cummins, 2012a).

Despite the lack of temporal regularity in speech, rhythm remains an important topic in speech research. Articulatory theories have proposed oscillatory mechanisms of speech production (Tilsen, 2019; Byrd & Krivokapić, 2021). Alternatively, where it is stated what is meant by *speech rhythm*, more nuanced definitions such as *perceived rhythm* can be studied without relying on stricter temporal definitions (Goswami & Leong, 2013; Turk & Shattuck-Hufnagel,

2013).

An alternative approach to understanding speech rhythm is to consider how we might construct rhythmic speech. While attempting to develop speech stimuli with regular presentation of words, Morton et al. (1976) observed that words with acoustically regular onsets did not sound perceptually regular. They coined the term p-centre (perceptual centre) to refer to the perceptual onset of a word. The p-centre is not determined by any one phonological feature, but is closely related to the rise in acoustic energy at the onset of a stressed vowel (Šturm & Volín, 2016).

In this thesis, the term rhythm is used in a general temporal sense and in reference to the acoustic speech signal. Specifically, it is used in a measurable sense where, in the time domain, speech in which p-centres are regularly spaced is described as more rhythmic than speech with irregularly spaced p-centres. In the frequency domain, the complementary definition is that more rhythmic speech has greater concentration of spectral power than less rhythmic speech (Tilsen & Johnson, 2008).

1.1.2 Can listeners entrain to speech rhythm?

While rhythm in speech continues to evade detection by speech researchers, typical listeners appear to benefit from intuitive temporal expectations in speech. It has been proposed that speech perception is facilitated by *neural entrainment*, where neural oscillations track *quasi-isochronous* modulations in the amplitude envelope of speech (Peelle & Davis, 2012; Gross et al., 2013; Giraud & Poeppel, 2012; Ghitza, 2013). The term entrainment, here, refers to the coupling of periodic neural oscillations with the quasi-periodic modulations of speech.

Neural entrainment hypotheses have been received with caution by many speech researchers. In particular, the *quasi* qualifier attached to descriptors of periodicity or isochrony in the acoustic speech signal does not satisfactorily address the complexity of natural speech (Cummins, 2012b; Turk & Shattuck-Hufnagel, 2013). This is addressed in neural entrainment theories by oscillators that are able to tolerate there lack of periodicity in amplitude modulations of the speech signal. For example, the *theta-syllable* theory (Ghitza, 2013) proposes that phase-locked oscillators facilitate parsing of the acoustic speech signal into syllabic chunks. This hypothesis proposes that the oscillator is free to gradually vary in frequency to accommodate for the natural variability of natural speech. In a similar theory, Giraud and Poeppel (2012) propose a *phase-reset* mechanism where salient edges in the amplitude envelope allow for corrections in the phase of oscillators.

Cortical tracking of amplitude modulations in speech has been widely demonstrated in neuroimaging studies (Rimmele et al., 2018). However, the functional role of this tracking remains an open question with interpretations including both passive processing and active synthesis of phonological features (Ding & Simon, 2014). Furthermore, debates remain over whether "oscillations" are discrete responses to regular stimulus onsets (Novembre & Iannetti, 2018) or actual oscillatory signals (Zhou et al., 2016).

When a periodic stimulus is presented, the neural response can be measured in the frequency domain, where a peak in spectral power at the stimulus frequency would be taken as evidence of entrainment. A limitation of this approach is that this method would not distinguish between periodic entrainment and time-domain tracking of a periodic signal that appears periodic when analysed in the frequency domain. In one experiment of a larger study, Jin et al. (2018) presented words with jittered onsets and recorded electroencephalogram (EEG) responses. They then transformed the response signal to correspond to isochronously presented stimuli (i.e., performing the same temporal distortion that would be required to make the speech isochronous). After performing this transformation they found spectral peaks corresponding to the periodicity that would have been present if the stimuli were not jittered. This finding could be taken as evidence of a robust phase reset mechanism able to adapt to temporal variations, or a time-domain explanation such as evoked response potentials (ERPs) corresponding to syllable onsets (see Alexandrou et al., 2020, for discussion). In either case, this would suggest that periodicity in the dynamics of the speech signal is not a requirement of neural entrainment.

If the oscillatory nature of neural entrainment is central to its function, it would be expected that it is easier to entrain to a periodic signal than an aperiodic signal. This would lead to the prediction that retiming speech to be more periodic would benefit speech processing as oscillators would not be required to adjust to stay in phase with the signal. This prediction was tested by Aubanel et al. (2016) in a behavioural study comparing intelligibility of isochronous, anisochronous, and unaltered speech. Isochronous speech was produced by retiming sections of the signal to achieve equal spacing of p-centre estimates. The anisochronous condition applied the same retimings as the isochronous condition, but in reverse order to match the total amount of temporal distortions without increasing periodicity. They found that intelligibility was increased in the isochronous condition compared to anisochronous, but intelligibility was lower in both conditions relative to the unaltered speech. This finding illustrates a trade-off between increasing the regularity of speech and preserving the existing prosodic structure.

Brain stimulation provides an alternative approach, focusing on the causal direction of perception benefits. This is important in establishing entrainment as an active process where the temporal dynamics of the signal affect perceptual benefits. Transcranial current stimulation (TCS) has been shown to increase the intelligibility of speech in noise when modulated to match the amplitude envelope of the speech stimulus (Riecke et al., 2017) and when periodic stimulation is applied in phase with isochronous speech (Zoefel et al., 2018).

1.1.3 Does entrainment to speech rhythm affect production?

Communication research has shown that speech rate can affect both the rate of spoken responses (Jungers & Hupp, 2009; Wynn & Borrie, 2020) and the time it takes to produce a spoken response (Corps et al., 2020). A possible interpretation of these findings is that entrainment effects in perception are carried over to production. Alternatively, these findings could result from perceiving rate as a pragmatic or social cue. For example, speech prosody can convey a wide range of intentions (Hellbernd & Sammler, 2016; D. Wilson & Wharton, 2006). Therefore,

responding in-kind to fast speech could be understood within the frameworks of Relevance Theory (Sperber & Wilson, 1987) or Communication Accommodation Theory (Gallois et al., 2005) as a cooperative behaviour. Here, synchrony between speakers could have a social function such as expressing social bonding or agreement.

Speech as joint action would go further than describing cooperative communication. Instead, integrating perception and production in a theory of joint action proposes that functional overlap between perception and production processes can simplify communication (Pickering & Garrod, 2013). For the listener this requires prediction of both the content and the timing of speech for effective communication (Garrod & Pickering, 2015). Here, neural entrainment would be a possible mechanism supporting the prediction of the timing of speech (Arnal & Giraud, 2012).

Neural entrainment as a mechanism to support prediction would not be a passive tracking process. In a visual paradigm, Breska and Deouell (2016) presented isochronously timed stimuli followed by targets that were either off-beat or on-beat relative to the stimulus. When targets were presented on-beat in a higher proportion of trials, response times to on-beat targets were faster. However, when targets were presented off-beat in a higher proportion of trials, response times to targets presented on-beat were slower. This demonstrates entrainment as an active process supporting prediction.

Cortical tracking of lip-movements in motor areas has been observed in silent speech (Bourguignon et al., 2020) and shown to be enhanced by congruent auditory speech relative to an incongruent distractor (Park et al., 2016). Such connectivity could be accounted for by action-perception circuits being employed for speech processing tasks (Pulvermüller, 2018). In other words, speech perception may be supported by treating perceived speech similarly to speech that is intended to be produced.

Meyer et al. (2020) draw an important distinction between entrainment and synchrony. They argue that many observations of entrainment could be explained by a synchronisation of internal cortical rhythms to the task of speech perception. Their theory allows for a complementary relationship between *entrainment proper* where synchrony is driven by external signals such as the amplitude envelope of speech, and intrinsic synchronisation of internal rhythm to external speech. An advantage of this theory is that intrinsic oscillations would be robust to irregularities in the signal allowing for quasi-regular signals to be internalised as regular representations.

1.1.4 How would we detect an effect of entrainment in speech production?

Speech as joint action is explicitly demonstrated by the synchronous speech paradigm used by Cummins (2009). In this paradigm, participants read a text aloud in synchrony with either another participant or a recording of another participant. Synchrony was measured in terms of deviations between the timing of voiced portions of the speech signals estimated by performing dynamic time warping (DTW) analysis. This paradigm demonstrated the ability of participants to synchronise to both live and recorded speakers, with lower levels of asynchrony observed in

the live condition. Furthermore, higher levels of synchrony were observed when synchronising to a stimulus that was recorded during a previous live synchronous speech trial than to a stimulus recorded individually.

A similar paradigm referred to as Spontaneous Synchronisation to Speech (SSS) was demonstrated by Assaneo et al. (2019) with two crucial differences. Firstly, rather than explicitly instructing participants to produce the same text as the recorded stimulus, they were asked to whisper /ta ta ta.../ while listening to an isochronously timed stream of syllables. At the end of each block, participants were asked to indicate if a target syllable was present in the stimulus. Importantly, participants were not given explicit instruction to synchronise with the stimulus. Secondly, synchrony was measured in the frequency domain using a measure of coherence (PLV; phase-locking value) between the stimulus and response signals. Assaneo et al. observed a bimodal distribution in the PLVs where some participants produced the response in phase with the stimulus while others were no more likely to synchronise to an isochronous stimulus than a white noise control.

Although synchrony, as observed in both synchronous speech paradigms, would be an expected behavioural presentation of entrainment, this ability to align with a speaker or stimulus could also be explained by accommodation. In particular, the finding of lower asynchronies to recorded synchronous speech (Cummins, 2009) suggests that the adjustments made when speaking in synchrony with another speaker has the effect of making the speech produced more predictable. However, in the prosodically rich continuous speech produced in this paradigm, prosodic cues such as intonational contours and preboundary lengthening (Shattuck-Hufnagel & Turk, 1996) could provide an alternative explanation of speaker accommodation. Similarly, in the SSS paradigm, synchronisation could be explained by the predictability of the isochronous stimulus rather than entrainment to the phase of the stimulus. The detection of a coherence could alternatively be explained by tracking in the time domain.

Strong evidence of entrainment to the phase of a stimulus would come from a sustained influence after discontinuation of the stimulus. An example of a paradigm that could produce this type of evidence is found in paradigms using a phoneme detection task. Quené and Port (2005) presented participants with isochronous or jittered (irregular) words followed by a final target word. They responded by indicating whether a given phoneme was present in the target word using a button box. In the isochronous condition, the interval between stressed syllable onsets (p-centre estimates) was 1.1 s and in the jittered condition a random uniform offset was added (1.1 ± 0.5 s). Additionally, they manipulated the metrical structure of stimuli to create matched and mismatched conditions, but found no effect of this manipulation. Across metrical structures, response times were faster in the isochronous condition, demonstrating an effect of periodic temporal expectations from the prior stimulus. This finding is particularly interesting as it suggests that the phase of the stimulus continued to modulate attention after stimulus offset.

A study by Cason et al. (2015) employed a phoneme detection task with a musical prime. This

paradigm used four different metrical structures that either matched or mismatched the metrical structure of the subsequent speech target. However, they deliberately presented speech targets out of phase with the musical prime to avoid measuring entrainment. Additionally, participants were grouped with an *audio-motor* group being given training to identify metrical structures in speech, and a control group receiving no training. They found a main effect of condition (matched or mismatched) across groups which they reported as evidence of a metrical priming effect. However, they did not report the main effect of group despite response times appearing to be slower in the audio-motor group (Figure 3 in Cason et al., 2015). This apparent effect suggests the alternative interpretation that audio-motor training, and thus attention to metrical structure, inhibited task performance. The important distinction between this study and that of Quené and Port (2005) may have been the intrinsic predictability of isochronous phase-aligned stimuli.

In addition to phoneme detection tasks, a continued influence of stimulus tempo after stimulus offset can be observed via context effects of tempo on other perceptual decisions. Presenting speech surrounded by slower or faster speech can affect whether or not function words are perceived (Dilley & Pitt, 2010), and in languages with vowel length distinctions what word is spoken (Reinisch et al., 2011). Kösem et al. (2018) extended this paradigm by using fast and slow context speech with high periodicity followed by a neutrally timed speech target without periodicity at either of the prior tempos in either of the context tempos. Additionally, they used magnetoencephalography (MEG) to measure neural entrainment to the context speech tempo. They found increased power at context speech tempos both during the presentation of the context speech and during the subsequent presentation of the target speech. In the behavioural task, participants identified whether a word presented in the target speech was the long or short vowel candidate of a Dutch minimal pair (e.g., TAK or TAAK; meaning *branch* or *task* respectively). Participants were more likely to perceive the word as the long vowel candidate when the context speech was presented at the faster tempo. In combination, the neuroimaging and behavioural findings suggest that neural entrainment affects how speech is perceived.

A more cautious interpretation of the findings of Kösem et al. (2018) would be that while they demonstrate continued periodic entrainment after discontinuation of periodicity and a context effect of rate, the two finding could be independent. For example, the observed entrainment could be evidence of low-level speech processing, while the context effect could be evidence of a time-domain effect of context rate on perception. In contrast with the design of Quené and Port (2005), the phase of the target was not aligned with stimulus. An effect of neural entrainment on perception in this case would be more complex than modulation of attention. Kösem et al. propose that the cycle length of the entrained tempo chunks speech, resulting in expectations about what information to expect from a chunk based on the period of a cycle. This interpretation would support the theories that neural entrainment has a facilitatory role in the perception of time (Henry & Herrmann, 2014).

Phoneme detection and context effect studies can provide valuable insights into the effect of

tempo manipulations on perception, but they do not address this question of whether perception effects can be carried over into production. Conversely, the two synchronous speech paradigms discussed are informative about the transfer of perceived temporal characteristics of speech to production, but do not allow entrainment to be distinguished from synchrony. Speech rate priming paradigms address this question by demonstrating influences of the rate of speech in a stimulus on the spoken response time in single word responses or speech rate in continuous responses. In these studies participants respond faster to questions presented at faster rates (Corps et al., 2020), and produce speech at a faster rate following a fast speech prime compared to a slow speech prime (Wynn & Borrie, 2020; Jungers & Hupp, 2009).

1.2 Overview

This thesis presents three empirical chapters, each documenting an experiment that investigates the effects of tempo and rhythm manipulations on the timings of spoken responses. The experiments share a paradigm which was refined to reflect both theoretical and methodological limitations of the previous iterations. For the first two experiments, due to the COVID-19 pandemic and subsequent restrictions, data collection was carried out remotely using a web-based experiment.

Two methodological chapters are presented prior to the empirical chapters to familiarise the reader with the techniques and challenges that informed the design. In the first of these chapters, different approaches to retiming speech are evaluated in terms of the amount of temporal distortion caused by the transformation. The chapter was written for inclusion in a forthcoming book (Beith et al., in press), and therefore attempts to appeal to a wide audience. The second methodological chapter documents software written to make the experiments in this thesis possible with remote web-based data collection.

The distinctions between variations of the experimental paradigm employed are, at times, subtle. To avoid confusion they are outlined here, following the hypotheses that were tested.

1.2.1 Hypotheses

These paradigms were used to test predictions of two hypotheses. The *tempo entrainment hypothesis* makes predictions about the effect of the tempo of speech on the timing of responses, while the *rhythm entrainment hypothesis* makes predictions about the effect of the rhythm of speech on the timing of responses.

Tempo entrainment hypothesis Here, *tempo* is defined as a frequency expressed in beats per minute (BPM). This usage implies the metrical regularity typically associated with music, and equivalent to the expression of frequency as hertz (Hz) in neural oscillation studies (i.e., 60 BPM = 1 Hz). Tempo can also be expressed as the period of a cycle, where 60 BPM would correspond to a period of 1 s. This use of the term tempo is not equivalent to terms typically expressing average rates such as *words per minute* or *syllables per minute*, where periods between onsets may vary.

The general prediction of the tempo entrainment hypothesis is that participants will show a tendency to respond *on the beat* of the stimulus, where the beat is a cyclical pulse induced by word onsets. This is operationalised in the time domain as the *period* (in seconds) between stimulus onsets. In the frequency domain, the beat is operationalised as the *phase* (in radians or degrees) of a response onset. In both the stimulus and response the *onset* is the p-centre estimate.

Time-domain predictions are made for both the central tendency and the dispersion of responses. Relating to central tendency, the response time prediction is that response times will be faster when trials are presented at faster tempos. The response dispersion prediction is that response times will be less dispersed when trials are presented at faster tempos. This prediction also has the frequency-domain interpretation that responses dispersed within a shorter phase cycle (i.e., faster tempo) will have a lower dispersion when measured in the time domain.

Improved response time measurement in the second and third experiment allows for a phase prediction to be tested. This is that responses will be more likely to be in phase with the trial tempo than a competitor tempo. Phase alignment was operationalised as *phase amplitude*, and tested with a method that borrows aspects of signal detection and coherence analysis methods (see Chapter 5).

Rhythm entrainment hypothesis Each of the predictions of the tempo entrainment hypothesis assume that participants entrain to the regularity or perceived regularity of the stimulus. This leads to the expectation that effects will be stronger when the *rhythm* of the speech is more regular. Here, rhythm refers to *isochronous* and *anisochronous* manipulations. Importantly, this hypothesis does not test the validity of isochrony as a property of speech rhythm, but rather as a parameter of the temporal aspects of speech rhythm that can be manipulated with regard to phase.

The predictions of the rhythm entrainment hypothesis mirror those of the tempo entrainment hypothesis. Firstly, there is the time-domain response dispersion prediction that dispersion of response times will be lower when the stimulus rhythm is more regular. Secondly, there is the beat concentration prediction, that response will be more likely to be in phase when the stimulus is more regular.

1.2.2 Paradigm

All three experiments employ what is referred to as the *maths sum evaluation* paradigm. The variations used in each experiment are illustrated in Figure 1.1. In this paradigm, participants hear an addition or subtraction maths sum (e.g., FOUR PLUS FIVE IS NINE, THREE MINUS TWO IS ONE) and evaluate its correctness with the word RIGHT or WRONG. The choice of words used in the stimulus production and the response are crucial to the paradigm. Producing the = operator as the monosyllabic IS rather than disyllabic EQUALS allows for a naturally quasi-isochronous rhythm while maintaining a natural production. Consideration was given to producing the − operator as LESS for the same reason but this was not felt to be a sufficiently

natural choice in Standard Scottish English. Sums containing the number 7 were not included in stimulus sets to ensure that all included numbers were monosyllabic. The choice of RIGHT and WRONG for evaluation responses minimises differences between phonological onsets (c.f., CORRECT and INCORRECT). This is important because sum correctness is an independent variable, and therefore phonological onset differences would introduce a potential confound. Furthermore, the choice of monosyllabic response words allowed for greater consistency between stimulus and response.

In all three experiments the tempo of the stimulus was manipulated. In Experiment 1 tempo was randomly manipulated as a pseudo-continuous variable within blocks, whereas in Experiment 2 and Experiment 3 tempo was manipulated between blocks. Rhythm was manipulated within blocks in Experiment 1, and between blocks in Experiment 2. There was no rhythm manipulation in Experiment 3. The motivations and implications of these choices are discussed in more detail within experiment chapters.

In Experiment 1 the visual presentation of the sum was static text displayed for the full trial. In Experiment 2 this was changed to a rapid serial visual presentation (RSVP; Potter, 1984) with visual stimuli being presented in synchrony with the onsets of their auditory counterparts. In Experiment 3 visual stimuli reverted to the static presentation of Experiment 1, but with a dynamic element in the form of a *karaoke dot* added to blocks of trials at the end of the experiment. The karaoke dot provided an explicit cue to the rhythmic nature of the stimulus. Additionally, in these explicit blocks, participants were instructed to respond on-beat. In contrast, in implicit blocks the karaoke dot was absent and participants received no specific instructions about response timing. The use of RSVP in Experiment 2 was not intended or expected to induce entrainment visually, however, reversion to static presentation in Experiment 3 makes a clearer distinction between the explicit and implicit conditions.

Experiment 3 also introduced the *maths sum completion* paradigm shown in the right panels of Figure 1.1. In this variation, participants were instructed to complete the sum in which the number was removed from the auditory stimulus and replaced with a question mark in the visual stimulus. The terms *evaluation* and *completion* are chosen with the intention of being neutral interpretation. They may alternatively be considered as *across-turn* and *within-turn* in reference to speaker turns, where the completion of a sum would be within a single turn and the evaluation of a sum would be across turns.

In Experiment 1 audio stimuli were presented as individual audio files with timing controlled by the experiment software, whereas in Experiment 2 and Experiment 3 full sums were created in advance. This reflects subtle developments in stimulus design techniques. However, across all three experiments auditory stimuli would be best described as *serialisation* in the same sense as visual stimuli were serialised in the RSVP employed in Experiment 2. This term is also used in Chapter 2 to describe this presentation technique in the context of the retiming of speech. In Experiment 3, the timing of individual tokens was normalised by aligning each token with the centroid of amplitude envelopes within word and position. This type of retiming is described

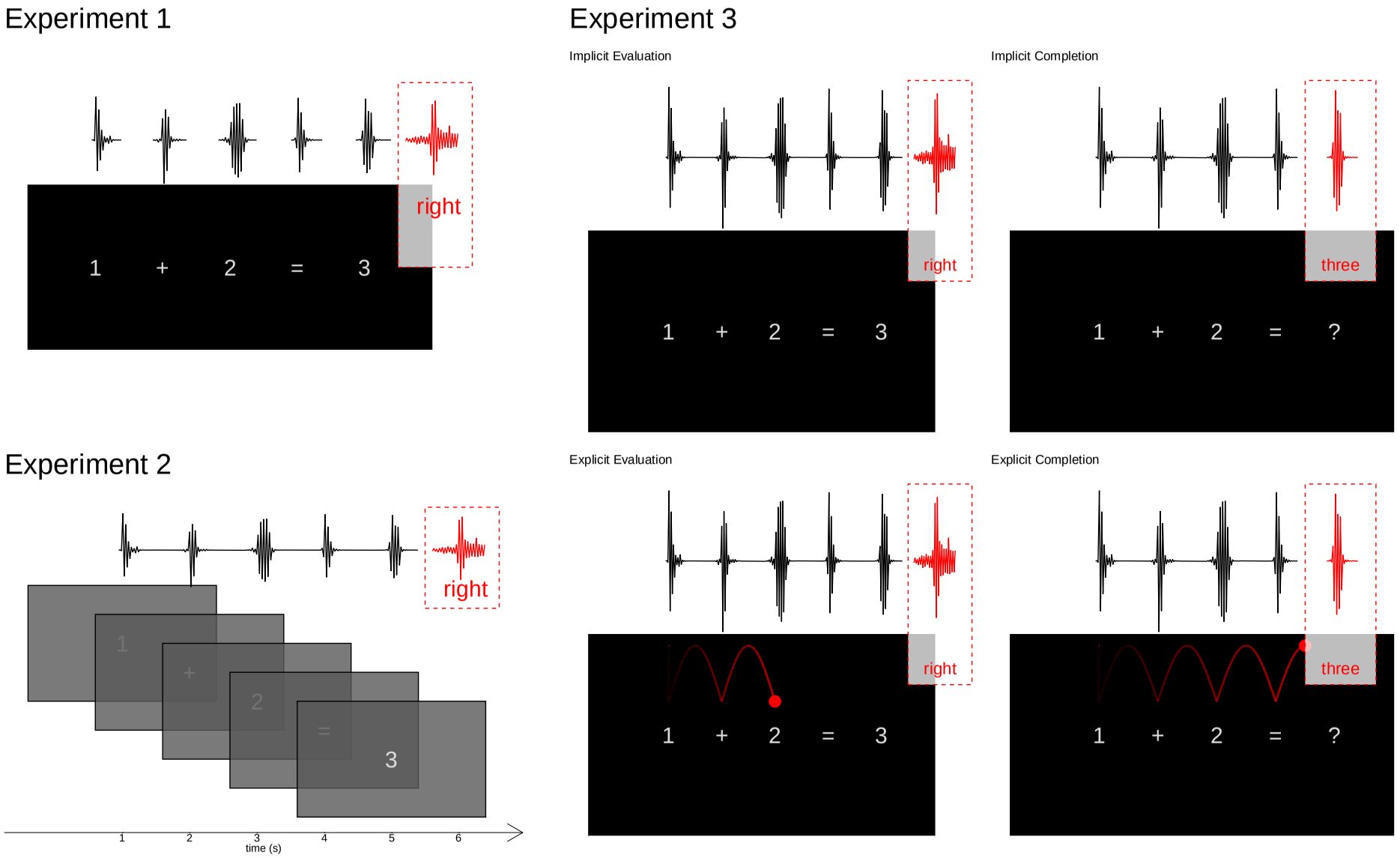

Figure 1.1: Schematics of the variations of the paradigm used in the three experiments reported here.

as *continuous* in Chapter 2, but as it was performed at the token level the full sums remain a serialised presentation.

All three experiments were created using variations of the *audio-audio* jsPsych (de Leeuw, 2015) plugin which is introduced in Chapter 3. This plugin was developed to ensure that the timing of auditory presentation, dynamic visual elements, and the recording of auditory responses were all scheduled using the same temporal software context. However, Experiment 1 did not benefit from the *latency-test* plugin which was developed to correct hardware related measurement errors that could not be directly controlled by software. While this is not strictly an element of the experimental paradigm, this development allows a phase interpretation of the timing of the response.

Chapter 2

Preserving Prosody in Temporal Distortions of Speech

To be published as:

> Beith, A., Barr, D. J., & Smith, R. (in press). Preserving prosody in temporal distortions of speech. In L. Meyer & A. Strauss (Eds.), Rhythms of Speech and Language: Culture, Cognition, and the Brain. Cambridge University Press.

2.1 Introduction

Rhythm in speech is the temporal structure created by patterns of variation in duration, fundamental frequency (f0), and intensity. These patterns can result from articulatory processes (Tilsen, 2019), characteristics of the spoken language (Cutler et al., 1997), and pressures arising from comprehension processes (Henke et al., in press; Grillo et al., in press). The rhythmic patterns in naturally-produced speech vary both within and between speakers.

In instances where we wish to control or manipulate speech prosody, this natural variation presents a challenge. For example, if contrasting the effect of different intonational contours on speech perception, it would be necessary to ensure that non-temporal parameters (f0, intensity, phonation type) were not correlated with temporal patterns that could offer an alternative interpretation of effects. Conversely, if contrasting different temporal patterns, it would be necessary to ensure that the independent variable is not confounded by non-temporal parameters.

Faced with this type of challenge there are three options: To record speech produced by a trained phonetician, to synthesise speech, alter the timing of a speech recording (retiming), or some combination of these (e.g., retiming and resynthesis of speech produced by a trained phonetician in R. Smith and Rathcke, 2017). Speech produced by a phonetician or by speech synthesis are both valuable tools for research. However, even trained phoneticians may not be able to control timing very precisely, whereas controlling timing in synthetic speech requires *a priori* decisions to be made in the parametrisation of the tool. In contrast, retiming speech can produce multiple signals with different temporal characteristics from a single source.

Although it would be preferable to avoid signal degradation or disruption of local prosody, there are ways of mitigating the impact. One approach is to counterbalance these effects across experimental conditions. For example, when producing isochronous stimulus items, Aubanel et al. (2016) created a matched anisochronous condition with the same amount of absolute temporal distortion. While this allows for comparisons between different retimings, it does not allow for naturalistic stimulus design. A further risk is that the experimental manipulation is more apparent to the participant.

Where it is a priority that stimuli sound natural it is important to consider subtle differences in how a retiming might affect an utterance. Two important factors to this are the amount of disruption to the signal and the location of the disruption. The least disruptive retiming would be no retiming at all, with the most disruptive retimings approaching the hypothetical physical limits of the transformation where parts of the signal are sped or slowed by infinite factors. Therefore, a possible approach is to perform the minimally disruptive retiming that achieves the desired temporal structure. However, extreme disruptions may have subtle effects in certain positions. For example, a part of a signal sped by an infinite factor would be equivalent to the removal of that part of this signal. In most cases this would be an unacceptable disruption to the signal. However, there would be special cases with ecologically valid interpretations such as retiming a recording of the word LIBRARY produced as /laɪbrəri/ in a manner that results in the elision of the medial syllable to produce /laɪbri/, amounting to complete deletion of one or more segments, as observed for many phenomena cross-linguistically (Bürki et al., 2011; Johnson, 2004; Dilley & Pitt, 2010; Turnbull, 2018). Similarly, infinite slowing of a section of the signal would be an extreme disruption, but at a word or syntactic boundary it would be equivalent to a speaker taking a pause (Zellner, 1994; Kentner et al., 2023). These retimings may be subtle in the sense that they would not necessarily sound unusual, but if not applied with careful consideration of the prosodic context could result in unacceptable disruptions to the signal.

The aim of this chapter is to provide an overview of the possibilities available to researchers producing retimed stimuli. Three methods of producing retimed stimuli are presented and their utility is demonstrated by creating utterances to have an isochronous rhythm. It is expected that all three methods will enhance the periodicity of the utterances to match the frequency implied by the isochronous intervals. Following this, we present an example of how the last of these three methods may be generalised to produce a wider variety of stimulus types.

2.2 Isochronous speech

2.2.1 Materials

A corpus composed of a male speaker of Scottish English producing simple mathematical sums was recorded and segmented into words for the purpose of demonstrating the three retiming methods. The corpus was composed of all 66 possible correct sums containing only the numbers from ONE to TEN, and the operators indicated by the words PLUS, MINUS and IS, e.g., FOUR

PLUS FIVE IS NINE, THREE MINUS TWO IS ONE. All sums containing SEVEN were excluded leaving MINUS (produced as /ˈmaɪnəs/) as the only polysyllabic word. Sums were spoken with a quasi-isochronous rhythm at an approximate rate of one word per second with short silences between words.

2.2.2 Retiming methods

Three approaches to retiming speech are considered here. In each case the interval between perceptual centres (p-centres; Morton et al., 1976; Rathcke, in press) is held constant at 1 second. Here p-centres are defined as the maximum rate of change of the amplitude envelope for a syllable; however, these approaches could be applied to alternative definitions. The first method achieves isochronous retiming by increasing the rate of the entire utterance and extending or inserting silences. The second method alters the rate of the signal between p-centres and the third method continuously varies the rate of the speech.

In order to compare these methods, we applied them to create stimuli with the shared specification that the p-centres of each word should be equally spaced to achieve an isochronous utterance. For this reason, the term *anchor point* is used here, somewhat interchangeably with the term *p-centre*. However, these methods are not limited to producing isochronous stimuli. Any feature that can be annotated as a fixed point could be used if there is a motivation to control the timing of the associated event.

Isochrony has advantages for the purposes of evaluation. The effectiveness of the retiming technique can be measured by the extent to which the signal is made more isochronous. Isochrony is defined in the time domain as a signal with equal time intervals between events and can be measured using autocorrelation[1]. Furthermore, isochrony would also be expected to be observable in the frequency domain as the rate of intervals per unit time. This can be measured using discrete Fourier transformation (DFT). Typically, the DFT representation will show a high level of spectral power at the frequency corresponding to the interval between events. There are however potential theoretical implications for the choice of measure. A DFT decomposes a signal into sinusoidal waves whether or not the signal is composed of sinusoidal waves. As a result, it may not be best suited to detect recurrences of irregular waveforms (Zhou et al., 2016). For example, the amplitude envelope of an utterance with a regular 1 s interval between syllables would typically have high spectral power at a frequency of 1 Hz. However, an irregular signal could have a high concentration of spectral power at 1 Hz because of a strong 1 Hz component to the signal such as a single cycle of a 1 Hz sinusoidal wave repeating at irregular intervals. In contrast, evaluating isochrony in the time domain using autocorrelation will capture recurrences even if no sinusoidal pattern is apparent in the amplitude envelope.

To provide a representation of the rhythmic structure of the input signals, the amplitude envelope was extracted using the `extract_env` function from the `retimer` package (Beith, 2023). This function is an implementation of the vocalic envelope extraction method proposed by

[1]The correlation between a signal itself at a lag. See Venables and Ripley (2000) for implementation used in this chapter.

Tilsen and Johnson (2008), with the additional option to control the low-pass filter and output sampling frequency. These modifications make it possible to extract a smoothed envelope by setting the sampling frequency higher than the low-pass filter frequency. Onsets were then detected by finding peaks in the rate of change in amplitude. An amplitude envelope of an example utterance is shown in the upper panel of Figure 2.1, annotated with the p-centres that are used as anchor points.

All retimings were performed in R with the `wsola` function from the `retimer` package. This function is a translation of the Wave-similarity Overlap-and-Add algorithm (WSOLA) implemented in the TSM toolbox (Driedger & Müller, 2014) for Matlab. A translation for Python users is also available in the PyTSMod library (PyTSMod, 2022). Like OLA, WSOLA involves overlapping windowed slices of the signal to compress or extend the duration of a part of the signal. The advantage of WSOLA, is that it allows the positioning of the overlaps to be adjusted to minimise phase discontinuity. The techniques discussed here could also be applied to any retiming algorithm implemented in either TSM toolbox or PyTSMod. Additionally, an OLA transformation can also be produced by using the WSOLA algorithm with the tolerance parameter set to 0 (Driedger & Müller, 2014). Many speech researchers will also be familiar with time-domain pitch-synchronous overlap-and-add (TD-PSOLA; Valbret et al., 1992) due to its implementation in Praat (Boersma & Weenink, 2023). This method performs a source-filter decomposition which makes it additionally useful for pitch transformations, but still shares the limitations of OLA in terms of phase discontinuities.

Method 1: Serialisation Most simply, what is referred to here as serialisation is an utterance that is created by presenting individual words in a sequence. This is the same stimulus presentation method that led to the observation that an utterance created with regularly spaced word onsets does not result in a perceptually regular utterance (Morton et al., 1976). It is also an auditory analogue to the common rapid serial visual presentation paradigm (RSVP; Potter, 1984).

Despite the intuitive appeal of this method, the actual process of producing stimulus items can be challenging. Altering the temporal position of a word in a continuous utterance without silences at word boundaries will result in an overlap with one of the neighbouring words. Therefore, to prevent overlaps, either the rate of words must be increased to shorten their durations, or the length of the utterance must be increased to provide additional space.

The implementation presented here is constrained by the previously mentioned requirement that the p-centres of each word are equally spaced by a fixed interval of one second. The function used here is provided in the `retimer` package (Beith, 2023) as `get_serial_anchors()`. It takes two sets of anchor points in seconds – corresponding to the original p-centres and desired p-centres – as arguments, along with the onsets and offsets of the words. This assumes that there is silence between words in the original utterance or that the researcher has inserted silences prior to the transformation. The function calculates the minimum reduction in speech rate that will prevent overlaps and returns a set of anchor points corresponding to word onsets

and offsets that will result in the desired p-centre alignment. If a sample rate is provided these anchors are transformed to index time in samples rather than seconds. A retiming factor can also be provided if – as is the case here – all items are to be retimed by a common factor. The resulting anchor points, along with the input signal and sample rate, can then be passed to the `wsola` function to return the retimed signal.

An alternative approach would be to use the anchor points returned by `get_serial_anchors()` or an equivalent calculation to individually present words or concatenate signals with silence. Presenting individual words may be more appropriate when generating stimuli in real-time during an experiment. However, concatenation may also require the insertion of spectrally matched noise rather than silences.

Method 2 (point to point) A logical progression from method 1 would be to relax the constraint that the rate increase applied to each word is uniform across words. However, the benefits of this approach would be limited as the necessary rate increase to prevent overlaps would be constrained by the position of neighbouring words. Therefore, it would still be necessary to insert silence in most cases. A solution to this is to annotate the utterance from p-centre to p-centre and apply the retiming to these intervals.

One of the advantages of this method over method 1 is that there is no need to rely on heuristics to calculate anchors without overlaps. The researcher is required only to provide a list of anchors referring to the p-centres in the input signal and a list of anchors referring to the desired timing of the p-centres in the output signal.

As this method does not insert any silence, or change overall utterance duration, there is no change to the average rate of the utterance. Instead, sections that are sped are compensated for by sections that are slowed. Additionally, the durations of pauses between words are preserved relative to the surrounding context. The cost of this is that tempo can change within a word and phone as p-centres are not typically located at word or even phone boundaries. This discontinuity in the speech rate will be most pronounced when the speech rate in the source recording alternates between fast and slow as this would result in one part of the word being sped and another part slowed.

Method 3: Continuous In the first method there was a discontinuity in time as silences interrupted the speech. In the second method there was a discontinuity in rate as fast intervals followed slow and slow followed fast. Method three smooths the transitions at rate changes by continuously altering the speech rate.

A convenient way of achieving this is to interpolate a spline passing through the same anchor points as would be used in the prior method. A spline provides a smooth, continuous function between anchor points so that abrupt discontinuities are avoided. In some cases, fitting a spline could result in a set of anchors that would imply that parts of the signal were to be reversed. This can be avoided by using the spline algorithm proposed by Hyman (1983) which fits a cubic spline to a series of points while preserving the monotonicity of the input. This algorithm is

available as the `spline` function in R and through the SciPy (Virtanen et al., 2020) library for Python.

2.2.3 Analysis

Each of these three retiming methods will result in the desired p-centre onset time. However, they do so with different degrees of temporal distortion and with differences in the distribution of these distortions over the duration of the utterance.

Retimings can be visualised by plotting the *warp path* of the original and retimed speech. The warp path is given by the intersection of matched points projected from adjacent axes as shown in the lower panel of Figure 2.1.

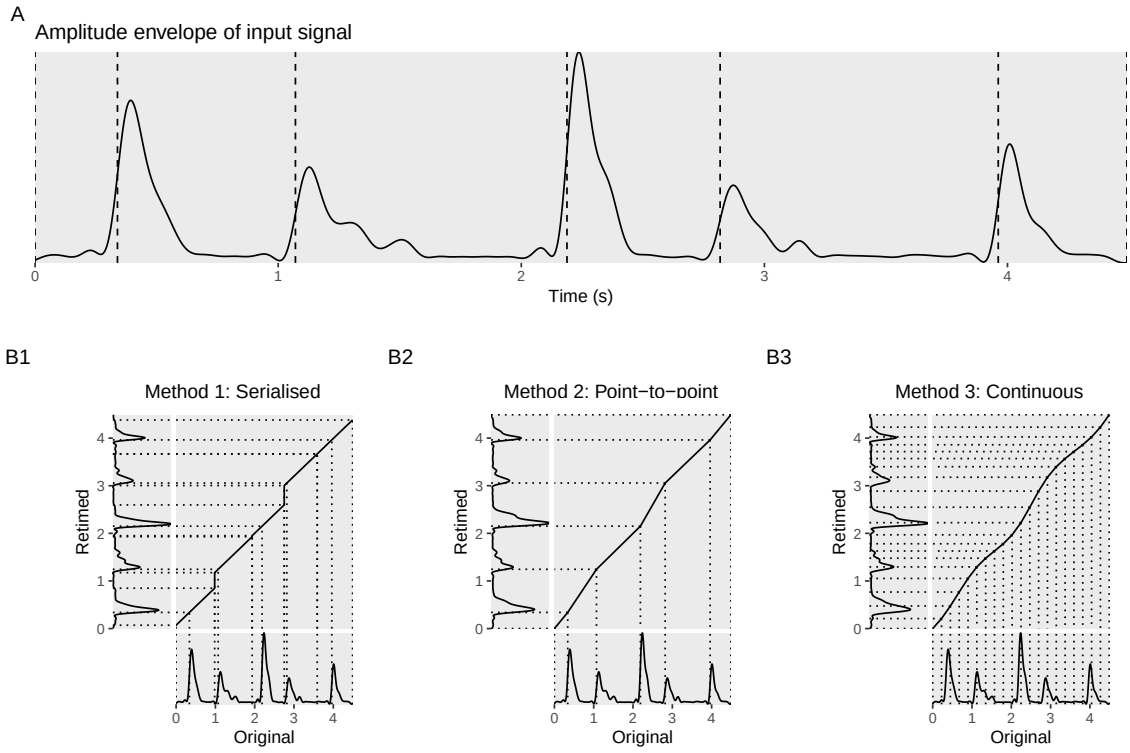

Figure 2.1: Upper panel (A) shows amplitude envelope of input signal annotated with p-centre estimates. Lower panel (B) shows the mapping of each of the three retimings (left axes) to the input signal (bottom axes)

The first method results in a warp path made up of two types of lines: The diagonal segments which are all projected at a common angle corresponding to the retiming factor, and the vertical segments corresponding to insertions of silence. This is indicative of a discontinuity in the speech itself. In contrast, the warp path resulting from method 2 is exclusively made up of diagonal segments with different slopes. While continuity of speech content is preserved, there are discontinuities in the speech rate at each of the anchor points annotated with dotted lines in Figure 2.1. Method 3 is made up of continuously changing slopes, while achieving a similar overall path to method 2.

Three retimed utterances, one using each of the three methods, were created for each of the 66

		95% HDCI	
Method	Median	Lower	Upper
1: Serialisation	0.057	0.009	0.207
2: Point-to-point	0.007	0.000	0.085
3: Continuous	0.008	0.000	0.176

Table 2.1: Warp path cost. Values show lengths of warp path over-and-above the shortest possible path and scaled to show the hypothetical limit as 1. Lower and upper bounds of 95% highest density continuous interval (HDCI) shown.

sums. They were evaluated in terms of the disruption to the signal and their effectiveness in producing isochronous stimuli. The full set of stimulus examples and a vignette of one example is available in the supplementary repository (Beith et al., 2024).

Temporal distortion The distinctions between different realisations of isochronous retimings are not merely conceptual. The warp path represents disruptions to the original signal, with the least disruptive path being the shortest path – a diagonal line of fixed slope – representing an unaltered signal. The least disruptive retiming would be the shortest path that passes through each of the anchor points. By definition, this is method 2 as each segment of the path is the shortest path between sequential pairs of anchor points.

Temporal distortion was measured by calculating the additional length of the warp path. In order to normalise the path length measure, and not to favour shorter items or penalise longer items, anchor points were first scaled between 0 and 1. This makes $\sqrt{2}$ the shortest possible warp path length between the input and output signal. There is also a theoretical upper limit of 2 which would correspond to a path made entirely of vertical and horizontal steps. As these limits are known, the path length is expressed here as a cost over-and-above the shortest possible path, and additionally scaled so that a warp path of length $\sqrt{2}$ would have a cost of 0 and a warp path of length 2 would have a cost of 1.

The results of the warp path cost analysis are shown in Table 2.1. They show a distinction between the similarly short paths of methods 2 and 3 and the longer path of method 1. Here there appears to be a small advantage to the point-to-point retiming compared to the continuous method when looking at the median. A larger difference between these methods is seen at the upper bound of the 95% highest density continuous interval (HDCI).

Increased isochrony In order to validate the retiming methods presented here, the increase in isochrony was measured in both the time domain and frequency domain. This was achieved using the same item set as for the previous warp path length measure. An additional control condition was included as a reference. For the control condition, the average interval between syllables was calculated and the utterance was sped or slowed to have the same average interval duration as the isochronous stimuli. This condition would have a normalised path length of $\sqrt{2}$, or path length cost of 0 for all items. It was expected that all three retiming methods would

have increased spectral power at the retimed frequency of 1 Hz, and increased autocorrelation at the retimed period of 1 s. Additionally, retimed signals would be expected to increase autocorrelation at multiples of the retimed period.

The FFT method used here is adapted from Tilsen and Johnson (2008). A vocalic amplitude envelope was extracted using the function `extract_env` in the `retimer` package. While Tilsen and Johnson suggest low-passing the signal at 80 Hz and then also downsampling the signal to 80 Hz, here the signal was low-passed at 32 Hz and downsampled to 1024 Hz. The lower low-pass frequency smooths out more high-frequency information and the higher sample rate provides a higher resolution in the FFT. Setting these values to powers of 2 ensured that the FFT of the signal would have frequency bins at whole numbers, and crucially the 1 Hz bin of interest. The same amplitude envelope was used for the autocorrelation function (ACF) analysis. For both analyses the spectra were averaged across all items.

Results of these analyses are shown in Figure 2.2. The gain in spectral power shown in the upper panel at 1 Hz appears modest for all three conditions with a small advantage for the latter two, less destructive, methods. The frequency spectrum also shows peaks for all three retimings at harmonic frequencies of 2, 3, and 4 Hz. The difference between control and retimings is more apparent in the autocorrelation analysis shown in the lower panel. Here there may be a slight advantage to the first method[2].

Discussion Depending on the hypothesis being investigated, the definition of isochrony may vary. For example, if the researcher hypothesises that the oscillatory structure of the amplitude envelope will elicit neural entrainment, it would be preferable to demonstrate a cyclical periodic structure in the stimulus. Alternatively, if it was hypothesised that greater regularity of edges in the amplitude envelope would increase the salience of the rhythmic structure, recurrence might be more important than cyclical periodicity.

Method 1 appears to be less appropriate for increasing periodicity. Due to the insertion of silence, any naturally occurring cyclical pattern could be disrupted by the creation of discontinuities. Furthermore, as the speech rate is increased without reducing the duration of the utterance, any existing periodicity in the envelope would no longer be coherent with the imposed frequency. This may explain the more modest power increase shown in the close up view of the 1 Hz bin in the upper right of Figure 2.2. For all retimings, power is also increased at the harmonic bins (2, 3, and 4 Hz) as would be expected from a periodic signal, with no evidence of recurrences at time-domain equivalences (0.5, 0.25, or 0.125 s) in the ACF spectrum. This again highlights that power at a frequency in an FFT does not necessarily correspond to recurrence at that frequency. In contrast, there appears to be an advantage to the first method over the others in the ACF measure. This may be because all words were increased by the same rate and therefore, regularities in the original production of the words were retained.

[2]No statistical tests are reported for either FFT or ACF analyses as no specific effects were hypothesised prior to analysis. However, non-overlapping error bars for the 1 Hz and 1 s peaks would indicate significance where no correction for multiple comparisons is made.

The advantage of methods 2 and 3 in the FFT analysis may not be enough to justify their selection over method 1. Similarly, their disadvantage compared to method 1 in the ACF analysis may not be enough to justify selection of method 1. The clearer difference is apparent in the warp-path cost measure. Both of these measures have lower median values than the lower bound of the 95% HDCI for method 1. This is unsurprising, particularly in the case of method 2, as it defines the shortest path passing through the required anchors. The additional cost of the third method appears to be minimal except with a higher upper bound. This would suggest that more careful checking of outliers may be required.

The results presented here may not necessarily generalise to other stimulus sets. Consider that neither point-to-point nor continuous retiming methods would alter an already isochronous stimulus item, while the serialisation method would insert silences. By extension, the measures reported here capture not only the performance of the retiming method, but also the irregularity of speech in the source recording.

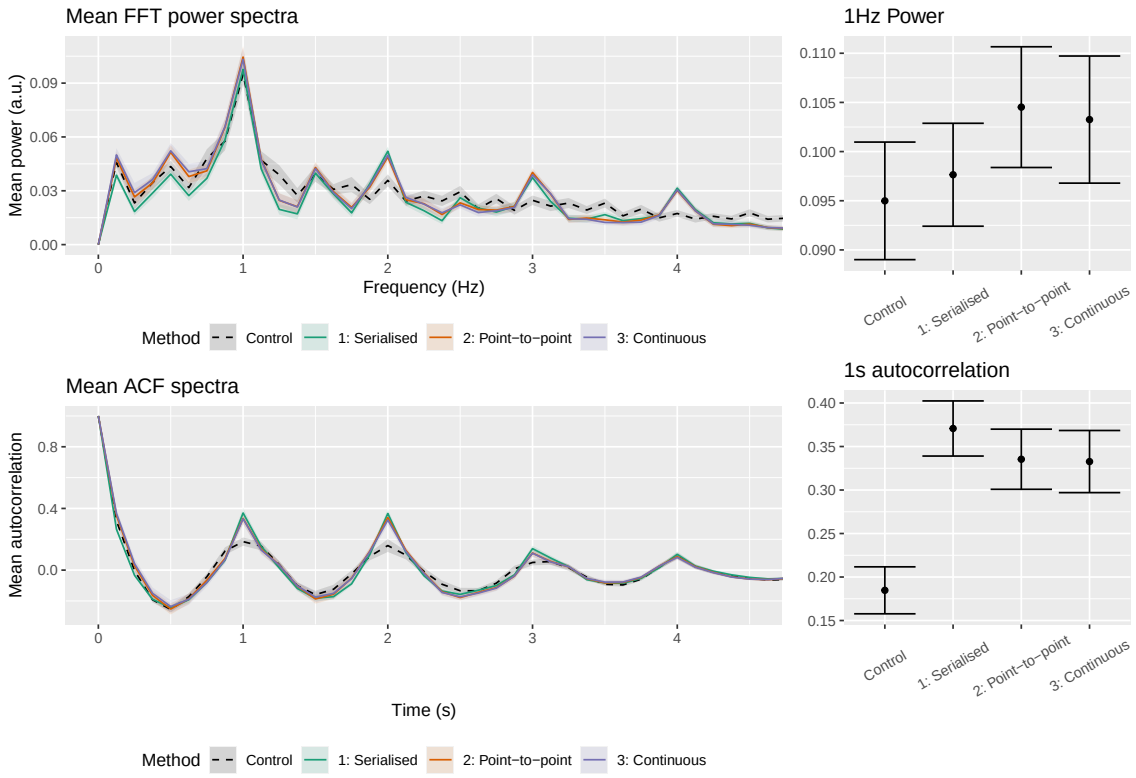

Figure 2.2: Left: FFT and ACF (AutoCorrelation Function) spectra of original amplitude envelope and amplitude envelopes resulting from each of the three retimings. Ribbons indicate +/- 1.96 SE. Right: View of only the 1 Hz and 1 s peak heights with error bars showing same confidence interval as ribbons.

2.3 Rhythmic chimeras

The flexibility of continuous retiming demonstrated in method 3 makes it useful for many applications beyond creating isochronous speech. In natural speech the temporal structure varies in subtle ways that would be challenging to define *a priori*. Two utterances could consist

of the same sequence of words and phrase structure but the rhythmic structure may vary between speakers or settings.

This section shows how continuous retiming can be used to create a *rhythmic chimera*. The analogy of a chimera is borrowed here from the term *auditory chimera* used by Z. M. Smith et al. (2002) to refer to combining the fine detail of one auditory signal with the amplitude envelope of another to create a hybrid of the two. Here, the rhythmic chimera does not replace the amplitude envelope, but instead uses it as a reference to warp the rhythmic structure of one utterance to that of another.

2.3.1 Materials

Example utterances were taken from the CHAINS corpus (Cummins et al., 2006). This corpus was created to support speaker identification research and consists of 36 speakers producing a range of sentences and longer texts under different reading conditions. Here the *solo* sentence tasks were used where the speakers read the sentences at their own pace. Recordings of a male and female speaker, each producing the same sentence, were sampled from these sentences to demonstrate the creation of rhythmic chimeras.

2.3.2 Method

There are three steps to the creation of a rhythmic chimera. These are feature extraction, alignment, and retiming. A vignette of the full stimulus creation process using R is provided in the supplementary repository (Beith et al., 2024).

Feature extraction The first step is to produce time series of each of the signals. The simplest option would be to use the amplitude envelope as with the previous examples. However, an improved alignment can be achieved by using a multivariate signal.

Mel Frequency Cepstral Coefficients (MFCC) have been used to perform alignments of utterances in previous studies (e.g., Cummins, 2009) and are used extensively in automatic speech recognition (ASR) tools (e.g., McAuliffe et al., 2017; Young et al., 2015). The advantage of MFCCs is that they reduce the feature representation from a full spectrogram to a smaller number (12 in the examples mentioned) of dimensions. More recent ASR tools such as OpenAI's Whisper (Aldarmaki et al., 2022) use full mel-spectrograms. Following this approach, mel-spectrograms with 80 bins and 10 ms time steps (i.e., 100 Hz sampling frequency) were used for feature extraction. As the mel-spectrogram is an intermediate step of MFCC extraction, the `melfcc` function from the `tuneR` package (Ligges et al., 2023) was adapted for this purpose.

Alignment Following feature extraction, alignment was performed by Dynamic Time Warping (DTW) using the `dtw` package (Giorgino, 2009). The warp paths shown in Figure 2.1 are similar to those that would be obtained from a DTW analysis. Each point in one time-series refers to a matched point in the other time-series. For DTW, this is done by finding the optimal path through the distance matrix of two signals that meets the defined parameters. When using a

multivariate signal such as MFCC or a mel-spectrogram, the distance method must be specified as "Manhattan".

As demonstrated by the retiming methods in this chapter, there are many warp paths that can produce similar results. In DTW, paths can be subjected to local constraints to specify particular characteristics. As the intention here is to use the alignment to perform a retiming of the signal, the alignment should be usable as anchor points, preferably with minimal additional processing. Therefore, the alignment path between the two signals must be monotonically increasing.

This constraint can be achieved by specifying a step pattern. An extensive review of step pattern options is included in Rabiner and Juang (1993) and further documentation of options is available in the companion paper for the `dtw` package (Giorgino, 2009). Although all step patterns are monotonically non-decreasing, many patterns, as with the serialisation approach (method 1), allow sequential points in one signal to be mapped to the same point in the other signal. Effectively, this would require the insertion of silence or omission of part of the signal to implement. The Rabiner-Juang Type III step pattern (section 4.7 in Rabiner and Juang, 1993) was used here to produce a monotonically increasing alignment. With this pattern, each step taken in one time series must correspond to one or two steps in the other time series.

Retiming The alignment maps points in the mel-spectrogram of one signal to the corresponding point in the mel-spectrogram of the other signal. Therefore, the only step required to produce anchor points for a retiming is to re-scale the alignment to the sampling frequency of the signal. Simply reversing the anchors allows for the retiming to be performed in the opposite direction.

This retiming would be useful if the intention was to retime one speaker to have the rhythmic structure of another. A further possibility is to produce ambiguous rhythmic chimeras. This can be done by averaging the anchor points to create a series of anchors representing the mid point between the time a sound occurs in one speaker's utterance and the time it occurs in the other speaker's utterance. For example, if speaker A produces a sound at the time 500 ms and speaker B produces that same sound at 1000 ms, both recordings could be altered so that both speakers produce the same sound at 750 ms. The same principle could be extended to produce a continuum of rhythmic structures by weighting the averages.

The vignette provides examples of unaltered recordings of the two speakers and all combinations of full rhythmic chimeras and half-warped rhythmic chimeras. A final example places the two half-warped rhythmic chimeras in left and right channels of a stereo audio file to allow for aural inspection of the alignments.

2.3.3 Results

Rhythmic chimeras are achieved by using DTW both as a means of analysis and of producing the stimulus. Therefore, insofar as DTW is an appropriate means of analysis, the retiming

is perfect. Furthermore, many criticisms of the retiming could be addressed by altering the parameters (most likely the step pattern) of the DTW analysis.

The primary limitation is that more extreme differences between speakers' rhythms will result in more disruption in the retimed signal. Where suitable to the experimental paradigm, the half-warped retiming presented here would be an appropriate method of minimising disruption. A further option would be to add additional constraints to the DTW step pattern. In the examples, type "a" slope weighting was used, incurring no additional cost for steps with steeper or shallower slopes. However, weighted slopes could be used to prefer unwarped steps at the local level over steeper or shallower slopes that provide a better fit. Global constraint can also be applied via a windowing function to prevent large deviations from the original signal.

2.4 Discussion

This chapter set out to demonstrate three methods available to researchers for producing stimulus items where temporal structure needs to be controlled or manipulated. To the extent that they were able to produce isochronous speech, all three methods performed well. By design all three methods ensured that the timing of the p-centre estimate occurred at the intended time. This was demonstrated by increased autocorrelation at the intended period in the time-domain (ACF) analysis. When measured in the frequency domain (FFT), the increase in periodicity was less apparent, but with a potential benefit of point-to-point and continuous retiming over control and serialisation.

A limitation of the isochronous stimulus examples is that the speech used was initially produced with quasi-isochronous rhythm and with silences between words. It should not be assumed that these results would generalise to connected speech where a weaker relationship between periodicity in the amplitude envelope and syllable rate would be expected (Zhang et al., 2023). In these cases segmentation will be more challenging and the effects of retiming methods should be assessed for suitability. Furthermore, as found by Aubanel et al. (2016), imposing strict regularity on speech with complex temporal structure can reduce intelligibility.

There are certain constraints that cannot be changed when altering the temporal structure of speech. Most crucially, information in the speech signal increases monotonically over time and any transformation would be expected to be monotonically non-decreasing at a minimum. Or more simply, it is unlikely that a researcher would want any part of the speech to be played backwards. The serialisation method for producing isochronous speech exemplifies one extreme of this limit, where silences are inserted reducing the rate at which the signal progresses to zero. In contrast the point-to-point method minimises deviation from the rate of the source signal to produce the most efficient path that passes through the desired points. Finally, the continuous retiming method attempts to smooth transitions at rate changes.

For the purposes of this chapter it is assumed that minimising disruption of temporal structure will in-turn minimise disruption of existing prosodic structure. This is reflected in the choice of quantitative measures used to operationalise prosodic disruption. However, future research

could benefit from expanding this scope to include qualitative and perceptual measures of prosodic structure.

A potential disadvantage of expressing these methods as warp paths is that it obfuscates some of the more intuitive differences. An equivalence of method 1 could be achieved by simply presenting individual words at predefined times. When constructing a stimulus item in this way there may be no expectation that the item will sound like a natural utterance or even that it should have the grammatical structure of a sentence or phrase. There are cases where stimuli of this type will have value such as the design adopted by Quené and Port (2005) where effect of different timings of word presentation were compared. However, caution should be exercised if there is any intention to interpret findings as being generalisable to spontaneous connected speech, (see Alexandrou et al., 2020 for discussion).

Choosing to focus on the commonalities of each of the three retiming methods highlights the complementary relationship between retiming and dynamic time warping. While retiming is used for performing manipulations, and DTW is used for analysis, both techniques can be expressed as a warp path. An accurate DTW analysis of a source signal and its retiming will result in a warp path resembling the anchor points used to perform the retiming. The steps allowed within these paths allow the researcher to meet the specific needs of their study.

In the case of an *a priori* temporal structure, such as the imposition of isochrony, the warp path maps a limited set of predefined anchors. Nonetheless, each of the three methods demonstrated had different local continuity constraints. In method 1 the angle could be either a diagonal of a fixed slope or vertical. In method 2 the angle of the slope could change, but only at fixed points. In method 3 the angle of the slope could change but the local change in angle was constrained by the effect it would have on the surrounding points. Recognising this relationship between retiming and DTW provides the stimulus designer with valuable insights from the DTW and speech recognition literature.

While isochrony provides a useful example, a researcher may not always be able to explicitly define the desired anchors in this way. Rhythmic chimeras provide an example of a more flexible approach to retiming a signal. In the example provided a rhythmic chimera can be used to present one speaker's utterance with another speaker's temporal structure. This does not require the researcher to define the temporal structure *a priori* in terms of onsets or boundaries, yet it allows for a meaningful retiming.

Thus the approach is valuable for exploring the contribution of temporal organisation to recognition of individual voices (Kello, 2003) and accents (Mareüil & Vieru-Dimulescu, 2006; Kolly et al., 2017; R. Smith & Rathcke, 2017). Furthermore, the half-warped rhythmic chimera would have applications in experimental control of items. Multiple speakers with different voice qualities or realisations of segments could be presented as stimulus items with the same temporal structure. Overall, DTW-based methods offer a flexible and practical way to apply the temporal organisation of one utterance to the spectral structure of another, such as in situations where researchers want to explore how temporal cues yield different structural interpretations of the

same phonemic content (e.g., different locations of word, morpheme, or prosodic boundaries; White et al. 2015; R. Smith and Hawkins 2012) or where the contribution of speech rate to perception of segmental, lexical, or larger units is to be explored (e.g., Dilley and Pitt, 2010; Reinisch et al., 2011). The methods do so without requiring *a priori* linguistic assumptions, and allowing for straightforward generation of intermediate values, and their advantages in terms of convenience and naturalness can be explored in future research.

2.5 Conclusion

Retiming speech is a convenient and powerful way of investigating effects of rhythm. However, these manipulations degrade and distort the source signal with the method chosen to perform the retiming impacting on the naturalness of the resulting stimuli. Dynamic Time Warping provides an analytical framework to describe the properties of different approaches to retiming. Recognition of this complementary relationship gives the experimenter greater ability to parametrise stimulus design to meet specific needs of the experiment. In doing so, this opens up new avenues for experimental stimulus creation that could investigate questions about the contribution of temporal structure to speech perception.

Chapter 3

Precision and Accuracy in Web-based Auditory Experiments

3.1 Introduction

The first two of the three experiments presented in this thesis were conducted during the COVID-19 pandemic. As a consequence of restrictions on in-person experiments, the first two studies were remote and web-based[1]. Furthermore, due to the temporal nature of the research questions addressed by this thesis, precise and accurate timing of responses relative to stimulus presentation was necessary.

Precision and accuracy are closely related concepts. Here, precision refers to the aspect of measurement affected by random error. In the context of psychological research this is closely related to reliability, where a precise measurement would be one that could be reliably obtained by repeating the same measure under identical conditions. However, in this case the source of random error is not the variability of human responses, but the variability of sources of sources of hardware and software error. Precision is also closely related to statistical power as, a measure must be sufficiently precise to detect an effect. Where a large effect is expected, a more imprecise measure may be acceptable, whereas finer precision would be necessary to detect a smaller effect. In contrast, accuracy refers here to the aspect of measurement affected by systematic error. This is similar to the concept of bias in psychological research, where a measurement would be accurate if it is able to measure a ground truth without bias. The importance of accuracy will depend on the experimental paradigm. For example, differences between response times (RTs) at different levels of a within-subject measure could be detected where the systematic error is captured by modelling random by-subject intercepts. Crucially, a measure can be accurate if the source of systematic error can be quantified and corrected for.

The current chapter provides a brief review of the current state-of-the art of web-based experiment software, before proposing and evaluating two software plugins that address current

[1]The term *web-based* is used here rather than *online* due to the potential for confusion between studies investigating online (real-time) speech processing and online (web-based) speech processing studies.

limitations. The first plugin, referred to as audio-audio, addresses unknown sources of error in auditory responses by aligning the context of stimuli presentation and response recording within a shared software framework. The second plugin, referred to as latency-test, uses the participant's microphone and headphones to measure systematic errors in measurement trials that can be used to adjust response times in experiment trials. Responses from Chapter 5 are used to evaluate the variability of latency-test measurements in a remote web-based study and responses from Chapter 6 are used to evaluate the accuracy of latency-test measurements using a version of the web-based software in a lab setting.

3.1.1 Experiment builders

A wide range of tools for building web-based experiments are available. Bridges et al. (2020) compared 5 experiment building tools (Gorilla, jsPsych, Lab.js, PsychoPy, Testable.org) across 3 operating systems (MacOS, Ubuntu, Windows) and 4 web browsers (Chrome, Edge, Firefox, Safari) on two computers (i.e., one Mac running MacOS and one PC running both Ubuntu and Windows). Across measures of errors in key response time, visual duration, and audiovisual (AV) asynchrony, most combinations of computer, operating system, and web browser achieved a mean precision within 10 ms. The authors found that their own PsychoPy tool had the best precision results across operating systems and browsers with a mean precision of 4.84 ms in the worst performing operating system and browser combination (MacOS and Chrome). Bridges et al. did not explicitly measure round-trip latency (RTL) in their study, however the AV asynchrony measure would be affected by audio output latency. Where a positive AV asynchrony would indicate audio lagging presentation of visual stimulus and a negative AV asynchrony would indicate visual presentation lagging auditory presentation, they found lags ranging between -44.36 ms (Ubuntu/Firefox/Testable) and 198.05 ms (Ubuntu/Chrome/Lab.js), with the majority of combinations resulting in positive lags. A negative AV asynchrony could be explained by delays in processing presentation instructions and the relatively low temporal resolution of visual presentation (i.e., an interval of 16.7 ms between frames for a monitor with a refresh rate of 60 Hz). A positive AV asynchrony could also be partially explained by delays in processing instructions, however output latency caused by buffering of audio samples is likely to be a large contributor. Despite the high magnitude of lags and variability between operating systems, browsers and experiment builders; the variability of these lags within combinations were still generally below 10 ms. This would suggest that if the lag could be measured during an experiment, AV asynchronies could be corrected. Although this particular challenge is beyond the scope of the current chapter, the principles of the solution for measurement of RTL presented in Section 3.2.2 could be applied to the measurement of AV asynchrony.

While Bridges et al. tested a range of combinations of operating system, web browser and experiment builders, performance on the computers used for testing may not generalise to the wide range of laptop and desktop computers used by participants in a typical study. A similar study carried out by the creators of the Gorilla experiment builder (Anwyl-Irvine et al., 2021) measured the precision of a range of web-based experiment builders on Windows and MacOS.

However, they used computers with lower specifications than Bridges et al. to capture a more conservative performance measure. They found that precision ranged between 4.69 ms (Lab.js) and 17.40 ms (jsPsych) in visual duration errors. Although they did not measure auditory timing errors, they did find longer delays in reaction times for keyboard responses than Bridges et al. (2020). Despite both Bridges et al. (2020) and Anwyl-Irvine et al. (2021) demonstrating promising levels of precision across a wide range of software combinations, the combined sample of four computers is a concerning limitation.

A more pragmatic approach to assessing the viability of web-based experiment software is to attempt to replicate findings from lab-based studies. Recent studies have successfully replicated auditory response experiments using remote web-based data collection. In a picture naming task, Fairs and Strijkers (2021) found similar effect sizes for both lab-based and remote web-based data collection using a platform named FindingFive. However, they note that mean response times were approximately 100 ms slower in the remote experiment. Using both a platform named SoSciSurvey and jsPsych (de Leeuw, 2015), Vogt et al. (2021) replicated significant effects in a picture-word interference experiment using both experiment builders. They also noted that response times were slower in the remote studies than in the previous lab-based study. They did not state the size of this difference but a visual inspection of the modes of response distributions (fig. 4 in Vogt et al., 2021) suggests a similar difference to the 100 ms observed by Fairs and Strijkers (2021). A third study (Stark et al., 2022), using a different interference task also successfully replicated a lab-based study using SoSciSurvey in a web-based experiment.

3.1.2 Limitations of current software

Despite apparent differences in the web-based experiment builders discussed they are all implemented using JavaScript and leverage web-browser application programming interfaces (APIs; Table 3.1) to capture audio responses in experiments. For example, in the SoSciSurvey experiments reported by Vogt et al. (2021) and Stark et al. (2022), audio was captured using the RecordRTC extension for WebRTC; and the audio-audio-response plugin developed by the jsPsych team (Gilbert, 2020) uses the MediaRecorder technology. Individually the WebRTC and MediaStream Recording APIs perform well for their intended purposes. However, they are not optimised for precise scheduling such as synchronising audio capture with audio presentation, forcing developers to rely on imprecise timers that have to compete with other processes in the DOM API.

The WebAudio API does allow precise scheduling of auditory processes, but does not implement audio recording directly. Vogt et al. (2021) addressed this issue in the jsPsych version of their experiment by using a ScriptProcessor to capture raw audio data. This method benefits from WebAudio's precise scheduling, but ScriptProcessors were deprecated in 2014 in favour of AudioWorkletProcessors [2]. Although ScriptProcessors use the WebAudio AudioContext for

[2]ScriptProcessors continued to be used after their deprecation as AudioWorkletProcessors were not implemented in a web browser until 2018 and were not supported by all major web browsers until 2021. Detailed

API	Technology	Uses
WebAudio	AudioDestinationNode	Audio playback
	ScriptProcessor	Processing (e.g., recording) audio. Now deprecated in favour of AudioWorkletProcessors.
	AudioWorkletProcessor	Processing (e.g., recording) audio
	AudioContext	Precise scheduling of audio processes
Media Capture and Streams	MediaStream	Capturing media (e.g., microphone or webcam)
MediaStream Recording	MediaRecorder	Recording media captured by MediaStream
WebRTC	WebRTC	Real-time audio and video applications (e.g., video conferencing)
	RecordRTC	Recording media
Performance	Performance	High-resolution timing for performance measurement
DOM (Document Object Model)	setTimeout/setInterval	Low precision scheduling

Table 3.1: Web API technologies. This table provides a list of technologies implemented by APIs and their uses for presenting, recordings, and timing audio stimuli. These APIs do not need to be installed by users or provided by the web server as they are specifications implemented in all major web browsers. All APIs refered to in this chapter are included in this table. Documentation can be found at https://developer.mozilla.org/en-US/docs/Web/API

scheduling, processing of audio is executed in the main processing thread. As a result of this a ScriptProcessors must compete with other processes making it vulnerable to distortion if available resources do not allow processing to keep up with the sampling rate of the recorder. Their successor, the AudioWorkletProcessor, addresses this by processing audio in a thread dedicated to WebAudio. This means that playing and recording audio can be scheduled with sample level precision. Although not implemented in the experiment builders discussed here at the time of writing, AudioWorkletProcessors can also be used for recording (Salomatin, 2018).

The problem that AudioWorkletProcessors solve is distinct from other factors affecting timing. For example, audio files must be read into a buffer before they can be played making it essential that audio files are preloaded. However, even then there are layers of software and hardware processing that delay (output latency) the signal actually reaching a loudspeaker of headphones. Similarly, the signal from the microphone only reaches the AudioContext after first passing through layers of hardware and software processed, including the MediaStream Recording API that finally directs the microphone signal to an AudioWorkletProcessor with the associate cumulative delay (input latency). The sum of these sources of input and output latency are refereed to as round-trip latency (RTL).

In the previously mentioned web-based replications of auditory experiments (Fairs & Strijkers, 2021; Stark et al., 2022; Vogt et al., 2021) researchers noted that, while they replicated effects (i.e., differences between conditions), response times were longer than lab-based studies. As noted by Bridges et al. (2020), relative precision is sufficient to detect an effect in many common experimental paradigms. However, this would not be the case in studies where the actual response time needs to be known.

In a lab-based study the systematic error attributed to RTL can be precisely measured. A standard procedure for this is to create a *loopback* where the audio output is connected directly to the audio input. In a study highlighting the importance of audio latencies in real-time auditory feedback experiments, Kim et al. (2020) found hardware RTLs ranging between 5.3 ms and 30.6 ms and combined hardware and software RTLs ranging between 19.3 ms and 39.0 ms. In doing so they also demonstrated the ability to achieve sub-millisecond precision (i.e., the tolerances of RTL measures) using a loopback. However, this procedure requires physical access to the computer making it impractical, if not impossible, to implement in remote experiments.

Although a physical loopback is not feasible, when a stimulus is presented over loudspeakers (as opposed to headphones), a similar measure can be achieved by using the participant's microphone to record the audio stimulus. Implementation of this solution is challenging due to a combination of background noise and noise cancellation features. Furthermore, participants' hardware, software, and physical environments can vary greatly resulting in different technical challenges. One solution to this is to design auditory stimuli with robust acoustic markers that can be detected in noise. Anglada-Tort et al. (2021) built the REPP experiment platform using this approach. In their study, participants were asked to reproduce a musical rhythm by

information can be found in the documentation cited in Table 3.1.

tapping it out on the body of their laptop. Additionally, the study dynamically adjusted stimuli based on participants' responses in prior trials. In order to achieve the accurate measurement of responses necessary for this paradigm, a burst of low frequency markers was played over the device speakers prior to each trial. After the trial the onsets of the markers were detected in the recorded signal and used as a reference for measuring the timing of the response events. The signal was filtered to remove the low frequency markers, allowing the response to be isolated.

While this technique is effective, it is optimised for a very specific paradigm. 1) The experiment is deployed on a server with the necessary signal processing software available in order to perform calculations during the experiment. 2) The acoustic markers may distract from stimuli. 3) The acoustic properties of a participant tapping on the body of their laptop will be much more predictable than other responses such as human speech. 4) The requirement to present the stimulus over loudspeaker precludes to possibility of asking participants to wear headphones, or of participants choosing to wear headphones. 5) The authors noted that noise cancellation technology can hinder detection of markers.

A more general solution is needed in order to allow a wider range of researchers to benefit from accurate response timing in web-based studies. Here, it is proposed that such a solution would ensure precise scheduling in experiment trials, and measure systematic errors in separate trials. To ensure validity of error measurements, the two trials types would share the same underlying design. Additionally, three further requirements are considered. Firstly, that both trials types should be compatible with headphones to allow greater control of audio presentation. Secondly, that the solutions are implemented in *client-side* code, meaning that processing is carried out on the participant's device without requiring specialist server software or sending unnecessary data. Thirdly, the measurement solution should be robust enough to work without filtering or otherwise preprocessing responses. The following section addresses these requirements by introducing two new types of experiment trial.

3.2 Plugins

Two plugins were developed for use with the jsPsych experiment builder. An important advantage of jsPsych is that it is open-source and has a modular design. This allows experimenters to combine useful existing plugins with custom plugins designed to meet the needs of a specific experiment. The first plugin is referred to here as *audio-audio* is an improved version of the previously mentioned audio-audio-response trial type (Gilbert, 2020) and is intended to be used to present auditory stimuli and collect auditory responses. The second plugin, referred to as *latency-test* performs real-time measurement of RTL in a dedicated trial to be presented between blocks of audio-audio trials.

3.2.1 audio-audio plugin

The audio-audio plugin addresses the issue of mismatched processing contexts in other approaches by implementing the recording of participant responses in an AudioWorkletProcessor.

This allows the audio-audio trial to benefit from the precise scheduling and dedicated processing thread of the AudioContext of the WebAudio API. The audio-audio plugin, like existing jsPsych plugins, use the WebAudio for scheduling playback of audio stimuli. Scheduling in the AudioContext has sample level precision on the condition that any requirements of the scheduled task (e.g., preloading an audio file) are complete before the scheduled time. If an instruction is scheduled at a time in the past, the instruction with be executed immediately. For example, a common practice in situations where timing is not crucial is to schedule playback at $t = 0$.

The recorder worklet was adapted from an existing example (Salomatin, 2018) to include visual presentation triggers. This is achieved via a messaging protocol that allows communication between the dedicated WebAudio processing thread and the main processing thread. By sending instructions in this direction the AudioContext clock can be used to synchronise timing between auditory and visual presentation. The type of visual element control varied throughout implementations. In Experiment 1 a simple volume indicator was displayed to show participants that the experiment was recording. In Experiment 2 the timing of a RSVP (rapid serial visual presentation; Potter, 1984) paradigm was controlled to match the timing of auditory presentation. In Experiment 3 the recorder worklet controlled the spatial position of a visual element corresponding to the progression of the trial.

In the versions used in Experiments 1 and Experiment 2, in addition to the response, the stimulus was directly captured by the recorder worklet to confirm that no delay was added during scheduling or recording. As Experiment 3 was lab-based, the stimulus captured along with the microphone signal via a physical loopback. A schematic diagram of this final version of the plugin is shown in Figure 3.1. This diagram shows where latencies occur and demonstrates that, with the exception of the latency associated with the propogation of sound in air, the only additional latency is the participant's response time. This visualisation also helps to illustrate the potentially unintuitive fact that it is not necessary to measure input and output latencies as it is the RTL that determines the error. The output latency would correspond to the delay that the participant experiences and would be relevant in experiments with precise audio-visual manipulations.

3.2.2 latency-test plugin

The latency-test plugin is a version of the audio-audio plugin configured to measure RTL. This is achieved by using the participant's headphones and microphone as a form of loopback. The stimulus is sent both directly to a worklet and via the microphone-headphone loopback to the same worklet. The RTL is then measured by calculating the lag between two signals. In Experiment 2, where the latency-test plugin was first implemented, the worklet was a recorder worklet and calculations were carried out using recordings. In Experiment 3 the same procedure was implemented via a customised worklet. Early in the development process it was found that detecting delays with audio markers such as those used in REPP (Anglada-Tort et al., 2021) required multiple signal processing steps that would be challenging to implement in client-side

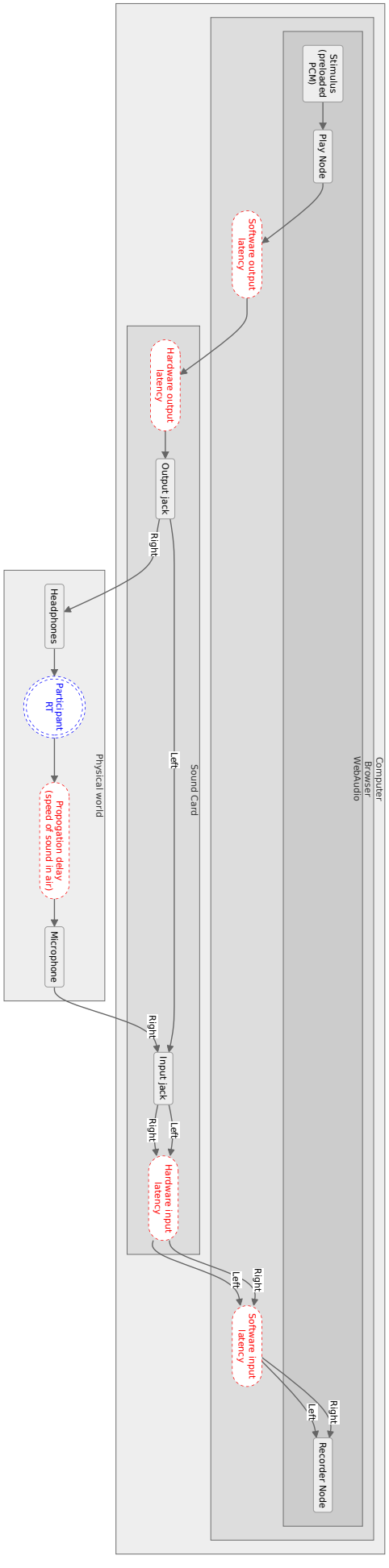

Figure 3.1: Schematic diagrams of audio-audio plugin with physical loopback. Stimulus output connected to a stereo channel of the audio input. Only the participant's response time and propogation delay due to the speed of sound in air are added to the total RTL of the loopback stimulus.

JavaScript. While this approach should not be dismissed, at the time of development it did not meet the requirements of a robust method that could be implemented in JavaScript in the available time.

A well established solution to measuring latency is found in radar and sonar technology. The distance between a measurement device and an object can be measured by calculating the delay between transmitted signal and the received signal. While there are many signal processing techniques that can be used to perform this task, the simplest way to achieve this is the frequency-modulated continuous-wave (FM-CW) technique ("Continuous-Wave Radar", 2024). This technique uses a signal referred to as a *chirp* that continuously sweeps between two frequencies. When this signal is multiplied by itself at a delay, a form of amplitude modulation occurs resulting in a signal with a prominent frequency component corresponding to delay between the signals. This is achieved by first calculating a scaling factor from the frequency range (Δf_{chirp}) and duration (Δt_{chirp}) of the signal as shown in eq. (3.1) and then using this constant to calculate the delay from the frequency detected in the multiplied signal (f_{mix}) as shown in eq. (3.2).

$$k = \frac{\Delta f_{\text{chirp}}}{\Delta t_{\text{chirp}}} \tag{3.1}$$

$$t_{\text{delay}} = \frac{f_{\text{mix}}}{k} \tag{3.2}$$

This technique is illustrated in Figure 3.3, showing spectrograms of the chirp, echo[3] and mix signals. A 200 ms delay was simulated to create an *echo* signal which was then multiplied by the chirp to produce the *mix* signal. The annotation shows that the prominent frequency in the mix signal corresponds to the intercept between the slope of the chirp and the delay in the echo. Two remaining slopes can also be seen in the spectrogram of the mix signal forming an equilateral triangle. This is the sum of the frequencies of the chirp and echo where frequencies above the Nyquist limit (half of the sample rate) are aliased.

Crucially, this method can be implemented in JavaScript using the WebAudio API. The chirp and the echo (the delayed chirp captured by the microphone) can be captured in precise synchrony within the same WebAudio context. The multiplication of the two signals can be achieved using a simple AudioWorklet and the frequency peak of the mixed signal can be detected using a FFT (Fast Fourier Transform) as implemented by the WebAudio AnalyserNode. This streamlined implementation was used in Experiment 3, while in Experiment 2 the analysis was performed using a recorder worklet and a signal processing library (Brook, 2010, February 24/2024) to perform the FFT.

As a consequence of the different implementations, the AnalyserNode version used in Experiment 3 detected the median frequency peak in a series of FFT spectra (effectively a spectro-

[3]The term echo follows conventions in the common use of the FM-CW technique in ranging tasks where the time for sound or light to be reflected (echoed) off a surface allows measurement of distance.

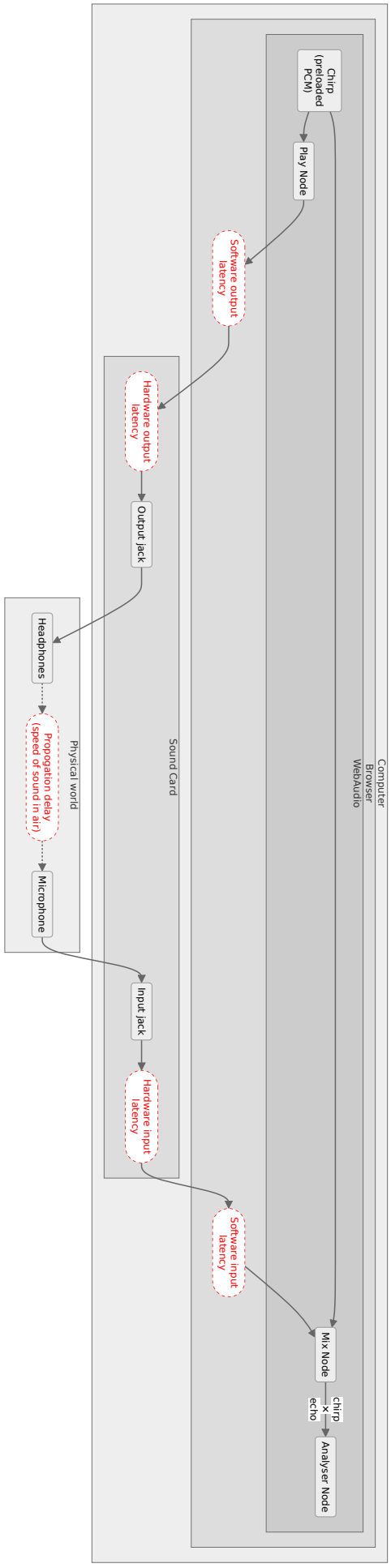

Figure 3.2: Schematic diagrams of latency-test plugin.

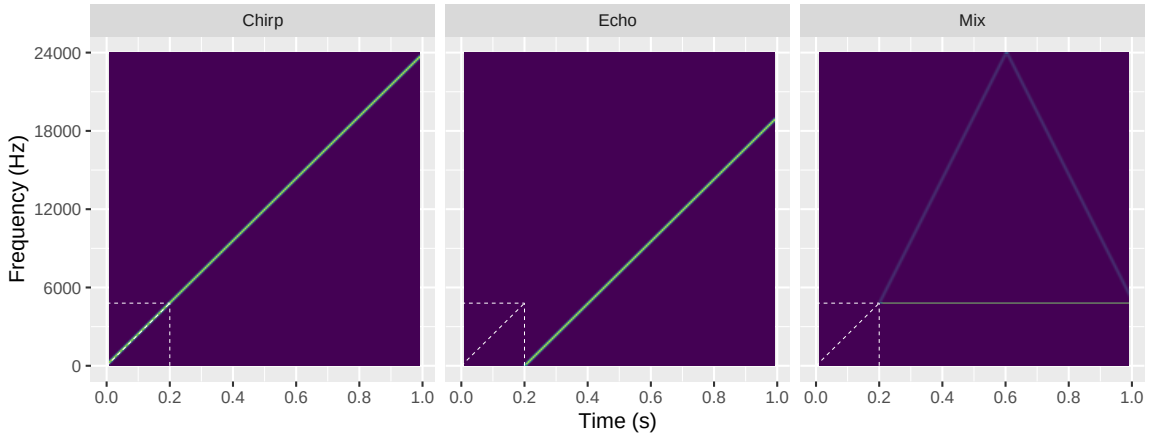

Figure 3.3: Spectrograms of chirp, echo and mix signals used in FM-CW radar technique. Annotation shows slope of chirp, delay of echo, and frequency of mix signal. The diagonal annotation aligned with the chirp signal represents the slope of k in eq. (3.1). The horizonal annotation shows the peak frequency of the mixture signal, corresponding to f_{mix} in eq. (3.2). Vertical annotation shows the latency which corresponds to t_{delay} in eq. (3.2).

gram), and the version used in Experiment 2 detected the peak in a single spectrum. The spectrogram results in lower temporal resolution as each slice of the spectrogram is shorter signal with the temporal being determined by the length of the FFT analysis and the range of the chirp. Compared to ground truth, estimates calculated from spectrogram slices should be accurate within the tolerances determined by the temporal resolution. Estimates calculated from individual spectra may be susceptible to outliers but the median of multiple trial estimates should be accurate where RTL is stable.

In order to validate these expectations, the implementation of the latency-test plugin used in the remote web-based data collection of Experiment 2 and implementation used in the lab-based data collection of Experiment 3 was performed. The validation aims to assess the precision and accuracy of the latency-test plugin. It is expected that the latency-test plugin will allow sufficiently reliable measurements to correct systematic errors in the timing of auditory responses to auditory stimuli while keeping variability within an acceptable range. The definition of *acceptable* will vary depending on the effect being studied, but based on the findings of Bridges et al. (2020), precision within 10 ms should be attainable with web-based experiment software. In lab-based data collection, precision is only expected to be limited by the tolerances of the measurement.

3.3 Methods

3.3.1 Design

There were no manipulations of the latency-test stimulus in Experiment 2. All latency-test trials were presented with a fixed frequency range ($\Delta f_{\text{chirp}} = 24$ kHz). The response was the estimated latency.

For Experiment 3, latency tests were presented in a fixed order with different FFT size and

frequency range parameters. There were two response measures: The estimated latency from latency-test trials; and the precise latency from experiment (audio-audio) trials.

Trial index	FFT Size	Δf_{chirp} (kHz)
1	1024	4
2	2048	4
3	1024	8
4	2048	8
5	1024	16
6	2048	16

Table 3.2: Experiment 3 latency test parameters. FFT size refers to the number of frequency bins used to calculate the latency. Δf_{chirp} is the frequency range of the chirp stimulus.

3.3.2 Procedure

Procedures for latency test data collection in Experiment 2 and Experiment 3 are shown in Figure 3.4. Full experiment procedures are included in the respective experiment chapters.

In Experiment 2, latency tests were initially presented as part of a screening procedure. For this initial latency test block, participants completed a microphone check where they were presented with a word on screen and asked to say the word and then listen back to a recording to confirm that the microphone was working. The next trial helped participants with integrated microphones to physically locate their microphone. In this microphone finder trial, white noise was played continuously through the participant's headphones and an on-screen instruction asked them to hold the headphones up to different areas of their computer while using a volume indicator to determine when they were close to the microphone. Once they had confirmed that they had located the microphone, they were instructed to hold the headphones up to the microphone and proceed to the first latency test trial. Latency test trials looped automatically for 10 repetitions of the same trial. The median absolute deviation (MAD) of latency test results was calculated and if less than 5 ms, participants proceeded to the first experiment block. If the MAD was not less than 5 ms they were given the option to repeat the latency test or end the experiment. Latency tests were repeated at the end of each of the 4 experiment blocks making a total of 50 latency tests.

In Experiment 3, participants completed the experiment in a lab with an external physical microphone. Headphones were placed over the microphone and participants were told not to put them on until instructed. The latency test trials were repeated 6 times with different parameters in each trial in the fixed order shown in Table 3.2. A second block of validation trials were presented between experiment trial blocks with the same presentation order as the first block.

3.3.3 Participants

Participants were recruited via the University of Glasgow School of Psychology subject pool and requests posted to social media. Participants were paid £3 for taking part in Experiment

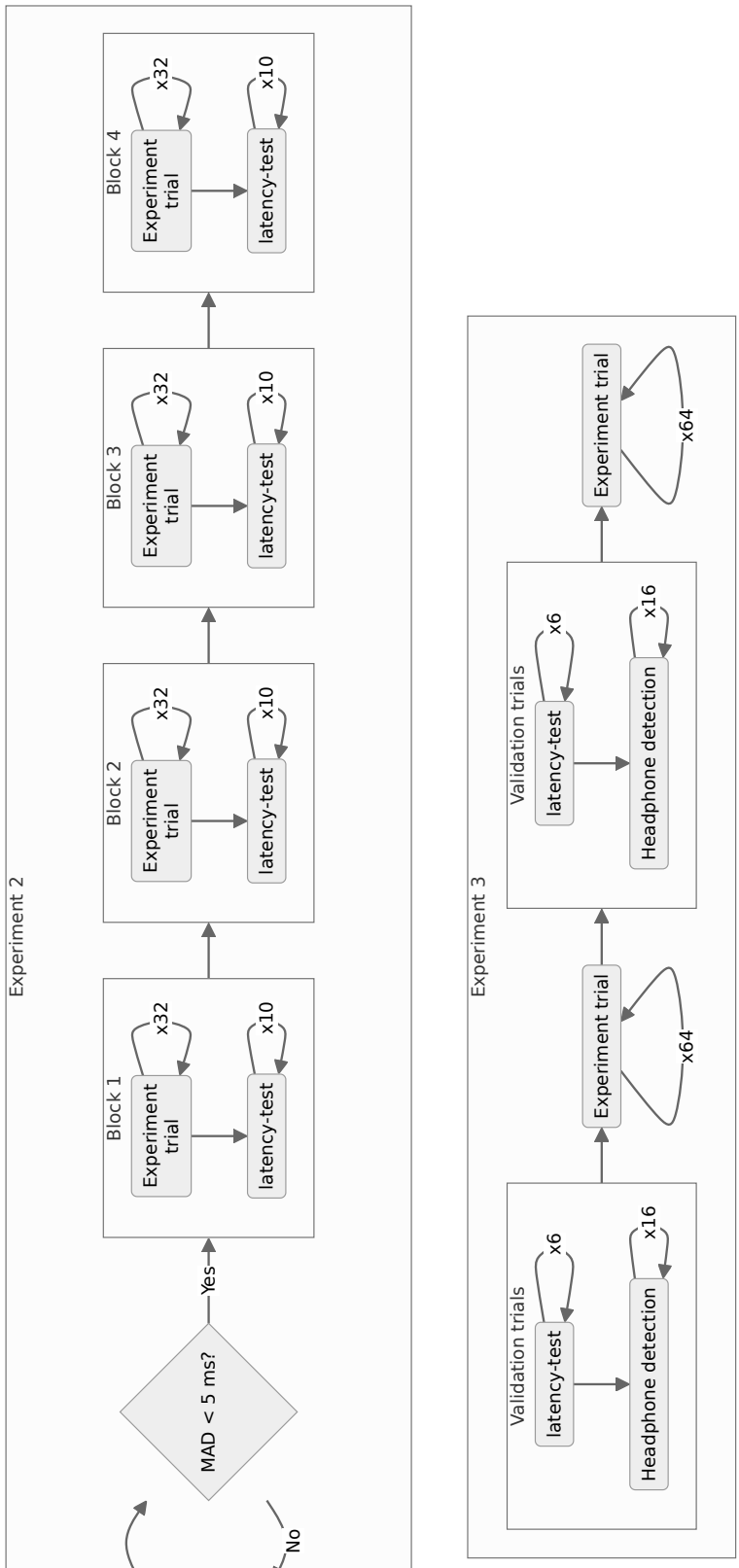

Figure 3.4: Abridged procedures for latency test data collection. Full experiment procedures are shown in their respective chapters. MAD = Median Absolute Deviation of detected latencies.

2 and £6 for taking part in Experiment 3. Participants were asked to confirm that they had not taken part in Experiment 2 to take part in Experiment 3. Ethical approval was obtained from the University of Glasgow College of Science and Engineering Ethics Committee.

A total of 64 participants took part in Experiment 2, with 58 being included in analysis after exclusions. A total of 32 participants took part in Experiment 3, with 29 being included in analysis after exclusions. Exclusions were the same as those reported in respective experiment chapters.

3.3.4 Materials

The chirp stimuli were generated using the `synth` function of the `seewave` package (Sueur et al., 2008). This function allows a chirp signal to be created by specifying a linear frequency modulation with the desired chirp range. In Experiment 2, the range was 24 kHz to match the Nyquist limit (i.e., maximum range) of a 48 kHz sample rate. As participants' audio interfaces will vary, the actual sample rate that stimuli were presented at would be lower in many cases. This would not affect latency calculations as the presentation sample rate was returned by the WebAudio API. Chirps used in Experiment 3 were created in the same way using the ranges shown in Table 3.2.

The implementation of the latency-test plugin changed between Experiment 2 and Experiment 3. The implementation used in Experiment 2 detected the peak of a single FFT spectrum of the full mixture signal. The length of the FFT was determined by the length of the recording and sample rate by rounding up to the next power of 2. This allows a high resolution calculations. For the Experiment 3 peaks were detected in FFT spectra extracted from a spectrogram. The median bin index of peaks was used to calculate the latency. This results in lower temporal resolution but makes the measure more robust to signal noise.

In Experiment 2, participants were asked to use Firefox (version $\geq$ 76) or Chrome (version $\geq$ 67). At the time of data collection AudioWorklets were not yet implemented in public releases of web browsers by Apple or Microsoft. Participants attempting to use a different browser[4] were presented with an alert indicating compatible browsers and were unable to access the experiment. In Experiment 3 all participants used the same computer in a university lab. The computer was using Ubuntu (version 20.04.6 LTS) and Firefox (version 101.0) web browser. Firefox updates, which can be automatic and frequent, were prevented by downloading a separate instance of the browser with updates disabled and using a script to ensure that the experiment was always launched in the same browser.

3.3.5 Measurement and analyses

Because Experiment 2 was remote, the accuracy of RTL estimates cannot be assessed. However, as tests were performed at 5 stages of the experiment, the variability of estimates within subject

[4]Participants' browsers were determined by the UserAgent string. A text string advertised by the browser when connecting to a server.

provides an indication of the precision of the method. Analysis of these responses consisted of plotting the variability of within-subject measurements and the cumulative distribution function of within-subject absolute deviations from the median.

The lab based data collection in Experiment 3 allowed a physical loopback. As show in Figure 3.1, the stimulus follows the same path as the participant's response. As a result, when compared to the stimulus loopback recording, the participant's response has the same RTL with the exception of the delay added by sound travelling through air from the participant to the microphone. Using this method in audio-audio experiment trials provides an accurate and precise measure of RTL by calculating the delay between the stimulus input file and the stimulus loopback output file. This calculation is performed by finding the lag associated with the maximum cross-correlation between signals.

The sub-millisecond accuracy and precision of the lab setup allowed the accuracy of the latency-test to be evaluated. As the precision of the RTL measurement was constrained by the latency-test parameters, accuracy would be defined as a measurement that is accurate at the level of single-trial precision afforded by the parameters. The range of parameters selected would also make it possible to distinguish between processing errors and out-by-one errors in the latency calculation. For example, a processing error could occur as the result of performance differences between the latency-test trials and audio-audio trials affecting all resolutions similarly. An out-by-one error would occur as the result of incorrect indexing of FFT bins.

Analysis of Experiment 3 latency measures consisted of plotting histograms of the loopback latency calculations and latency-test estimates on the same scale.

3.4 Results

3.4.1 Remote data: Experiment 2

Across all participants that completed Experiment 2 the median latency was 103.1 ms (MAD = 56.7). Across the sample, the dispersion of corrected latencies was within the 5 ms MAD screening threshold (MAD = 3.8 ms).

Within-subject variability of latency estimates are shown in Figure 3.5. Results are split by operating system and sorted by median values, showing generally lower RTLs for MacOS users compared to Windows. However, the largest median RTL was a MacOS user with a median RTL of approximately 300 ms and three additional Mac users had large interquartile ranges. Only one participant (ID = 36; Windows), used Firefox and had no deviations from the median latency estimate (111 ms). Although this is only one sample, it is consistent with experience of testing the plugin during development. The remaining participants all used Chrome.

The cumulative distribution function of within-subject absolute median deviations is shown in Figure 3.6. More than 50% of latency estimates were within 2 ms of the subject-level median across operating systems. Estimates were more precise across the Macs users until approximately 8 ms where a higher proportion of Windows were within this level of precision.

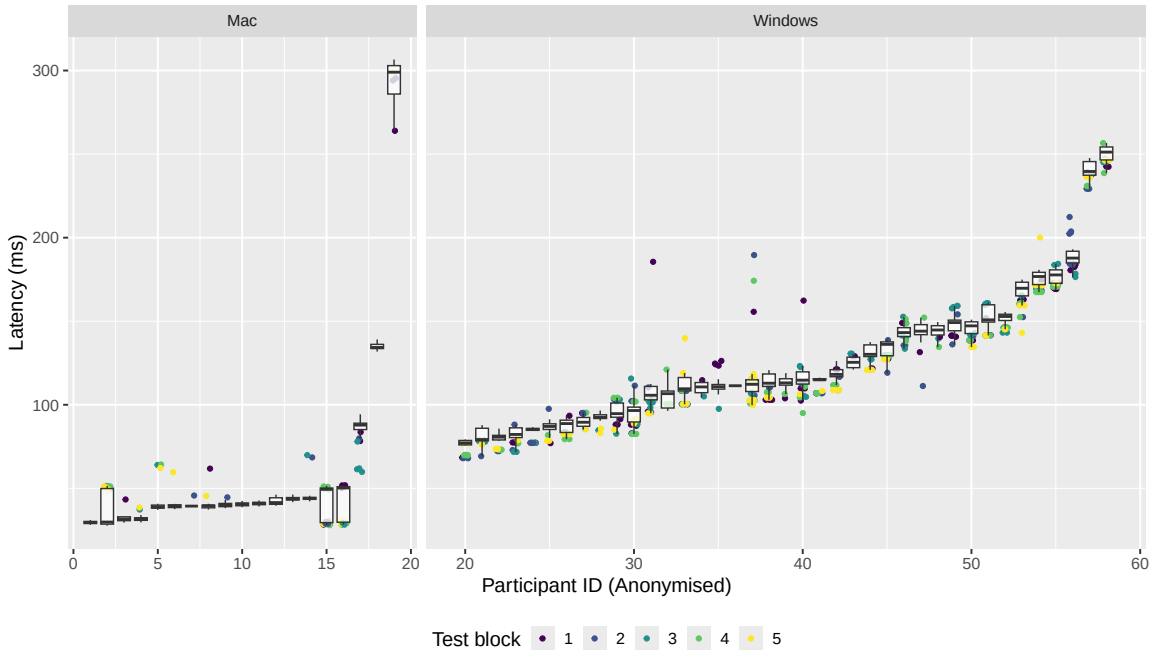

Figure 3.5: Experiment 2 latency estimates. Boxplots of latency estimates grouped by operating system and anonymised participant ID. Points show estimates with a deviation of more than 5 ms from the subject median, with colour showing the test block.

This cross-over is consistent with the wide interquartile ranges observed in four Mac users.

3.4.2 Lab-based data: Experiment 3

Across all participants that took part in Experiment 3, RTLs measured with 1 ms precision calculated from experiment trials (audio-audio) with physical loopback ranged between 63 ms and 70 ms. The distribution of these RTLs is show in the background of Figure 3.7. The foreground of the plot shows RTL estimates from latency-test trials in blue, faceted by FFT length used for analysis and frequency range of the chirp stimulus. Bin width in all panels shows the most precise combination of chirp and FFT length (bin width = 1.35 ms; $\Delta f_{\text{chirp}} = 24$ kHz; FFT length = 2048). Gaps between histogram bars provide a visual indication of the loss of precision resulting from smaller chirp ranges and shorter FFT lengths. All panels indicate that latency-test estimates are accurate within the precision of the trial parameters.

3.5 Discussion

Recent studies (Fairs & Strijkers, 2021; Stark et al., 2022; Vogt et al., 2021) have successfully detected significant effects in remote web-based experiments using speech onset measurements. However, these studies also found that response times were approximately 100 ms slower in web-based experiments than lab-based experiments. The findings from Experiment 2 presented in the current chapter found that participants' round-trip latencies (RTLs) were consistent with this magnitude of error. However, differences between operating systems and individual users' devices can result in large between-subject differences.

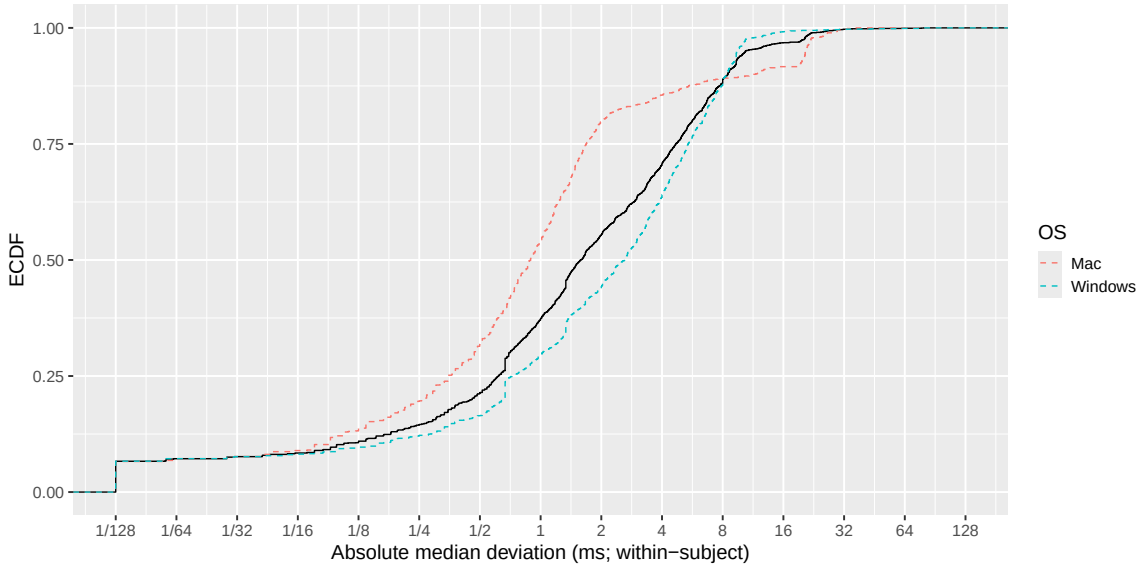

Figure 3.6: Experiment 2 latency precision. The empirical cumulative distribution function (ECDF) of absolute median deviations in milliseconds from subject median latency estimate. Deviations of 0 coded as $\frac{1}{128}$ to allow log transformation. Black line shows ECDF of all participants, with dashed lines showing ECDF for each operating system (OS).

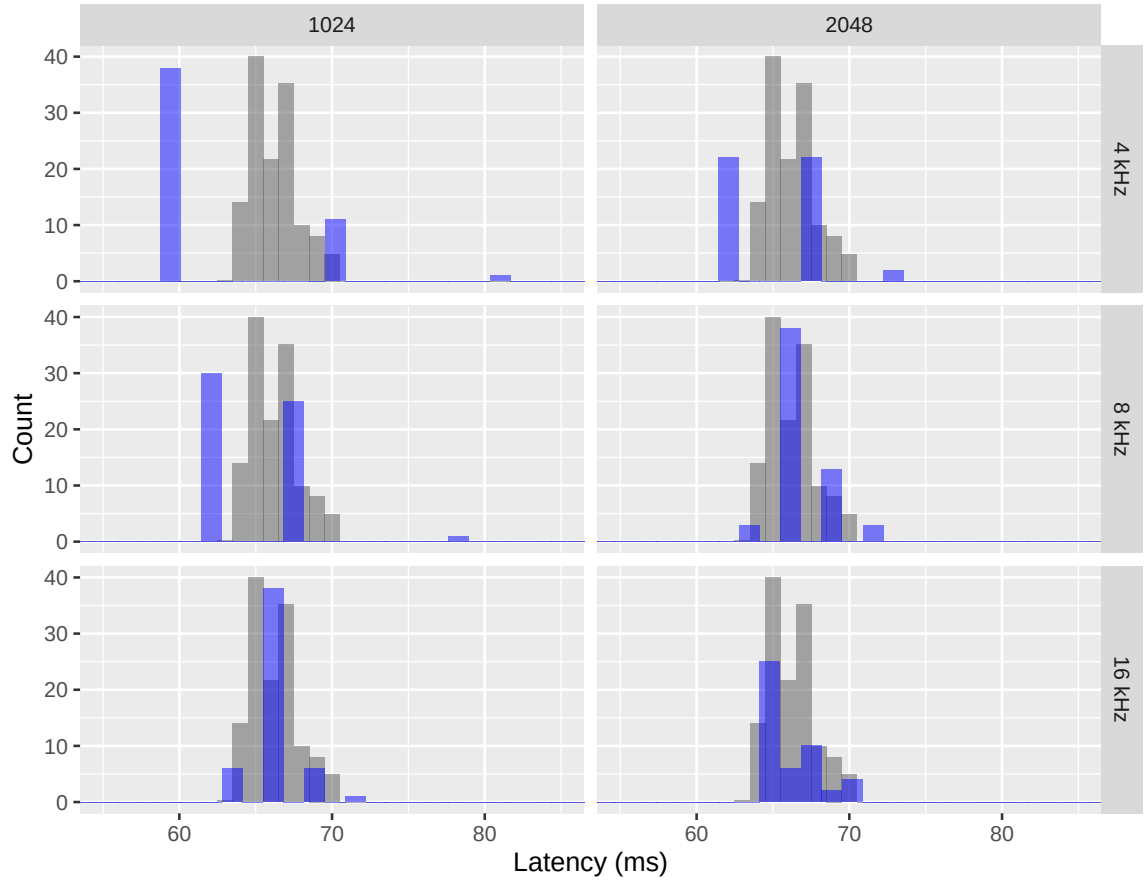

Figure 3.7: Latency test estimate distributions. Blue foreground histogram shows counts of latency estimates in latency-test trials. Grey background histogram shows scaled counts of latency measures in experiment (audio-audio) trials with physical loopback. Facet columns show the length of the FFT used to detect latency, facet rows show the frequency range of the chirp signal. Bin width for all panels is the precision of the 16 Hz chirp response measured with a 2048 bin FFT (1.35 ms).

All but three of the participants using a MacOS computer had RTLs shorter than the lowest median Windows RTL. Surprisingly, despite this, the longest participant RTL was a MacOS user. Although participants were asked to use wired headphones to take part in the study, one potential explanation would be that this participant may have instead used Bluetooth headphones to complete the study. This illustrates the importance of considering the impact of different technologies on RTLs on web-based auditory experiments. Experimenters cannot assume that reported performance of an experiment builder will be a reliable indication of performance in their own sample.

Although the audio-audio plugins allowed the timing of visual stimuli to be controlled with the high precision of the WebAudio context, this does not mean that the actual presentation of visual stimuli will be precisely aligned. Firstly, computer monitors do not share the temporal resolution of an audio signal. For example, a monitor with a 60 Hz refresh rate corresponds to a temporal precision of approximately 17 ms. Additionally, there will be an unknown delay resulting from the time it takes to process the instruction received from the AudioWorklet. With the addition of output latency, this configuration would be expected to result in the audio presentation lagging the visual presentation which would be reduced, but not eliminated, by delays in visual presentation. This would be more likely to be perceived as synchronous than visual presentation lagging audio presentation (Van Wassenhove et al., 2007). In the current implementation the intention was for participants to perceive auditory and visual stimulus as being synchronous. In the latency data collected from Experiment 2, the highest RTL was approximately 300 ms. If this was the sum of symmetrical input and output latencies, it would correspond to audio lagging video by up to 150 ms. However, if it was not symmetrical, such as in the case where a participant might use Bluetooth headphones and a wired microphone, a greater audio/visual asynchrony could occur. An important next step in web-based auditory experiment design is to determine if audio/visual asynchronous can be measured remotely without specialist equipment.

A challenge to the current evaluation is the ubiquity of Chrome among participants. Testing suggested that Firefox performed better than Chrome, but only 1 participant in Experiment 2 chose to use Firefox. At the time that Experiment 2 was conducted neither Safari nor Edge supported AudioWorkletProcessors and participants attempting to use these browsers received an error suggesting that they should use Firefox or Chrome. AudioWorkletProcessors are now supported by all major browsers, making it necessary to evaluate the performance of these browsers in future studies. Moreover, AudioWorklets are now supported on all major mobile device browsers for iOS and Android. The experiments reported here were not designed for a touch interface and have not been tested for mobile devices. However, despite clear instructions to complete the study using a desktop or laptop computer, server logs showed several errors from participants attempting to use a mobile device to take part in the experiment.

The use of the FM-CW radar technique for latency measurements would not be ideal for all experiments. The signal would be unpleasant for participants to listen to on loudspeakers, making it best suited to use between experiment blocks using headphones. In the experiments

reported here participants were required to tick a box to confirm that they had removed headphones before proceeding with the latency test in order to avoid discomfort. In experiments where stimuli are to be presented over loudspeakers to allow trial level latency measurements, the approach adopted by Anglada-Tort et al. (2021) of detecting onsets of auditory markers may be more appropriate. The audio-audio plugins presented in this chapter would allow for experimenters to use any signal of their choosing for latency measurements if offline latency measurement would be acceptable.

Chapter 4

Experiment 1

4.1 Introduction

Chapter 1 reviewed four questions: Is speech rhythmic? Can listeners entrain to speech rhythm? Does entrainment to speech rhythm affect production? And, how would we detect an effect of entrainment in speech production? The first three questions build a view of rhythm in speech as a shared temporal structure mediating speech as joint-action between speaker and listener. The final question focuses on empirical approaches to distinguishing between this joint-action characterisation of speech and alternative interpretations.

The question answering study conducted by Corps et al. (2020) addresses the theory that listeners covertly imitate the speaker to support comprehension (Pickering & Garrod, 2013) and subsequently to time responses (Garrod & Pickering, 2015). Their account is contrasted with that of M. Wilson and Wilson (2005), where an anti-phase relationship between speakers is proposed to avoid overlapping. Corps et al. predicted faster responses to faster stimuli, but they did not specifically predict responses to be in-phase. They presented questions such as "Do you have a pet horse?" at natural rates and sped rates where the sped rate was double the natural rate. Rates can be approximately inferred from reported mean trial durations and the stimulus sentence list. In the natural condition the rate was ∼278 words per minute (WPM), and ∼557 WPM in the sped condition. Expressed as the period between word onsets this would correspond to ∼216 ms in the natural condition and ∼108 ms in the sped condition. Mean response times were in excess of 400 ms in all conditions, with the main effect of the rate manipulation reported as 19 ms.

A rate of 278 WPM, for the mostly monosyllabic questions in the list, would correspond to syllable tracking in the theta range (4-8 Hz; Gross et al., 2013). In contrast, the syllable rate in the sped condition would exceed this range. The design may have benefited from evaluating responses in the slower delta range (1-2 Hz; Gross et al., 2013), which is associated with supra-syllabic periodicities in speech (Tilsen & Arvaniti, 2013) and differences in strength of neural entrainment between dyslexic participants and controls (Goswami & Leong, 2013). At this slower timescale, a participant could be expected to align responses with the pulse of the

stimulus if response times are influenced by entrainment to the stimulus.

4.2 Overview

In this first variation of the paradigm outlined in Section 1.2.2, participants were presented with full sums as on-screen text and sequentially presented individual words as auditory stimuli. A schematic of the paradigm is shown in Figure 4.1. In all trials participants were required to evaluate a complete sum by producing the word RIGHT or WRONG as appropriate.

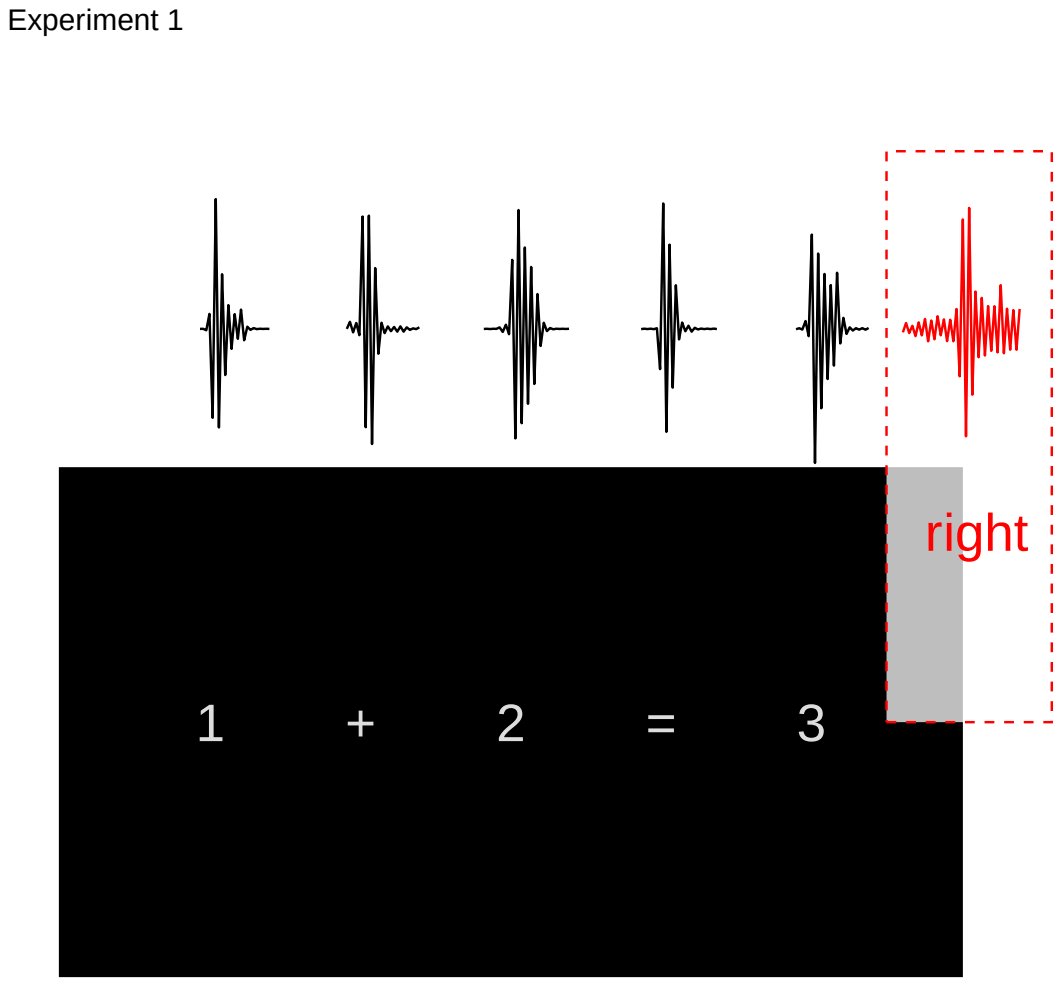

Figure 4.1: Schematic diagram of Experiment 1 paradigm. Upper panel shows auditory stimulus presented as five individual audio files. Lower panel illustrates the visual stimulus and a red circle providing feedback to participants that their response was being recorded by varying the opacity of the circle. A red circle presented below the text to indicate that the microphone was recording is ommited from the diagram. Diagram not to scale.

Both the tempo entrainment hypothesis and the rhythm entrainment hypothesis are addressed in this experiment. At the time the current experiment was conducted, the latency-test plugin had not yet been developed. Therefore, predictions were tested with time-domain analyses.

Specifically, three predictions of these two hypotheses were tested:

1. Tempo entrainment hypothesis
    - Response time prediction: Response times will be faster when trials are presented at faster tempos
    - Response dispersion prediction: Response times will be less dispersed when trials are presented at faster tempos
2. Rhythm entrainment hypothesis
    - Response dispersion prediction: The dispersion of response times will be lower when the stimulus rhythm is more regular.

4.3 Methods

4.3.1 Design

In the current experiment, a within-subject and within-item design was used. Three independent variables were manipulated by the design: tempo, rhythm, and correctness. Tempo was a pseudo-continuous variable with 13 equally-spaced levels from 60 BPM (beats per minute) to 120 BPM, with each word marking a beat. The rhythm manipulation had two levels (isochronous and anisochronous). The correctness of the presented sum was manipulated with two levels (right and wrong) which also correspond to the word participants were expected to produce in their response. The dependent variable was the time to response onset, which measured the interval between the onsets (p-centre estimates) of the final stimulus word and the participant's response.

Participants were presented with complete addition and subtraction sums which were either correct or incorrect and asked to evaluate the sum with either RIGHT or WRONG (e.g., correct, FIVE PLUS THREE IS EIGHT; incorrect, FIVE PLUS TWO IS EIGHT). The trial length was the duration from the first word onset (word onset as opposed to p-centre estimate) until 4000 ms after the final stimulus p-centre. The complete text of the sums was presented on screen for the full duration of the trial including the 4000 ms response window. The tempo and rhythmic regularity of sums was manipulated within subjects. There were 13 levels of the tempo manipulation, equally spaced between 60 BPM and 120 BPM where each stressed syllable represents a beat.

4.3.2 Participants

Participants were recruited via the University of Glasgow School of Psychology subject pool and requests posted on social media. Undergraduate students taking part were eligible for course credit. As a requirement of taking part all participants self-reported as native English speakers with normal or corrected to normal hearing and vision. Additionally, all participants were requested to use a computer with Firefox or Chrome. Data collection was stopped when 64

participants had completed the experiment. Five participant were excluded due to responding with the wrong word (e.g., CORRECT rather than RIGHT) or not at all in more than 20% of trials. Ethical approval was obtained from the University of Glasgow College of Science and Engineering Ethics Committee.

4.3.3 Materials

Sums were created to contain only monosyllabic words as far as possible. There are 66 possible sums that meet the criteria of containing only the monosyllabic digits between 1 and 10 (excluding 7) and the operators + (PLUS), − (MINUS), and = (IS). MINUS, the only disyllabic word, was produced with a reduced second syllable (/'maɪnəs/). Of these possible sums 64 were used in the experiment and the remaining two items were used as examples in the tutorial. A matched set of incorrect sums was created by permutation of correct answers, thus ensuring the incorrect sums had the same distribution of answers as the correct sums. Four lists of 64 items were created in order to counterbalance the four possible combinations of correctness and rhythm variables for each item. Trial tempos were randomly sampled with replacement from the 13 tempos independently of the list assignment. Trial order was randomised by sampling without replacement from the full trial list.

The stimuli presented to participants were assembled from word tokens segmented from natural recordings. All 66 correct sums were recorded by a female speaker of Standard Scottish English. The speaker was asked to produce the sums at a comfortable pace with separation between each word to aid segmentation. Incorrect sums were not recorded. Additionally, the speaker produced a full set of isolated numbers and operators which were used to aid segmentation.

The segmentation was performed by aligning full sums to sums concatenated from the isolated numbers and operators. Both the full sum recordings and the concatenated recordings were transformed into Mel-frequency cepstral coefficients (MFCCs) to allow alignment using dynamic time warping (DTW) based on the Manhattan distance between the two MFCCs. This allowed the joins in the concatenated sums to be mapped onto the full sums to determine the segmentation point. Segmented tokens were saved, retaining the word, source sum, and position of the word in the sum. This method is conceptually similar to that used in the alignment of rhythmic chimeras in Section 2.3.

The sum stimuli were composed of tokens that occurred in the same position in a different sum. The word IS (used for the = operator) could be sampled from a sum where it preceded any of the eight alternative answers. This avoided confounding the correctness variable with any potential effects of incongruity from splicing, as both correct and incorrect answers were preceded by a token from a sum with a different answer. For each trial a list of files was provided indicating which tokens were to be used in each position.

The presentation technique was equivalent to the serialisation method discussed in Chapter 2. However, full sums were not prepared in advance. Instead, for each trial, a list of tokens was used with file path references, presentation times and p-centre estimates. Presentation was

then controlled using a version of the audio-audio plugin discussed in Section 3.2.1.

Isochronous trials were defined as having a regular interval between p-centres and anisochronous where the intervals between p-centres were irregular. A distinction is drawn here between anisochronous and jittered rhythm where the former refers to a rhythm that is skewed away from regularity and the latter refers to adding random offsets to the timing of stimulus tokens (i.e., words or syllables). This is illustrated in Figure 4.2 with alternative approaches to producing an irregular rhythm. Jittering onsets by adding a duration sampled from a uniform distribution (top row) results in intervals between onsets fitting the distinctive triangular random uniform difference distribution. Allowing the first and last onset time to vary leads to the mean interval between onsets being free to vary creating dispersion of trial tempos as shown in the middle column of the top row of Figure 4.2. In an experiment where tempo is manipulated, this dispersion would not allow comparison between isochronous and anisochronous responses at the same tempo. Attempting to control tempo by scaling these jittered onsets to fix the position of the first and last word (middle row) leads to onsets distributed around isochronous onset times. This appears to result in an asymmetrical distribution of intervals between words. In contrast, the anisochronous rhythm manipulation presented here (bottom row) first samples the offset (i.e., jitter) of the central word p-centre from a beta distribution and the direction from a binomial distribution with equal probability of being positive or negative. The first and last onsets are fixed and the second and fourth are then interpolated with the addition a random uniform jitter. Crucially, this results in low density of the distribution of onsets at the third beat avoiding both syllable-timed (i.e., intervals between words) and stress-timed (i.e., intervals between numbers) forms of isochrony. The resulting distribution of intervals between words is a combination of two random uniform distributions of different widths.

The p-centre of each word was estimated by identifying the point of maximum rate of change in the amplitude envelope using the same method as used in Chapter 2 and illustrated in the upper panel of Figure 2.1. The timing of this onset relative to the start of the audio file was recorded for each token. All tokens were retimed using a fixed factor for each level of the tempo manipulation such that tokens to be presented at 120 BPM would be twice as fast as items to be presented at 60 BPM. For each token and tempo the calculated p-centre estimate was scaled by the retiming factor to adjust the presentation time for each token.

4.3.4 Procedure

The procedure is illustrated in Figure 4.3. Participants were presented with a tutorial video demonstrating the main experiment task. The video demonstrated allowing the experiment access to the microphone, confirming that the microphone is working in a microphone check trial, and finally a correct and incorrect example of the main experiment task. In the microphone check trial participants were asked to respond to a prompt to produce a word, listening back to the recording, and confirm that the word had been captured. They were then asked to confirm that they understood the task and that they consented to taking part before completing a short survey to collect demographic data. This was followed by the microphone check trial

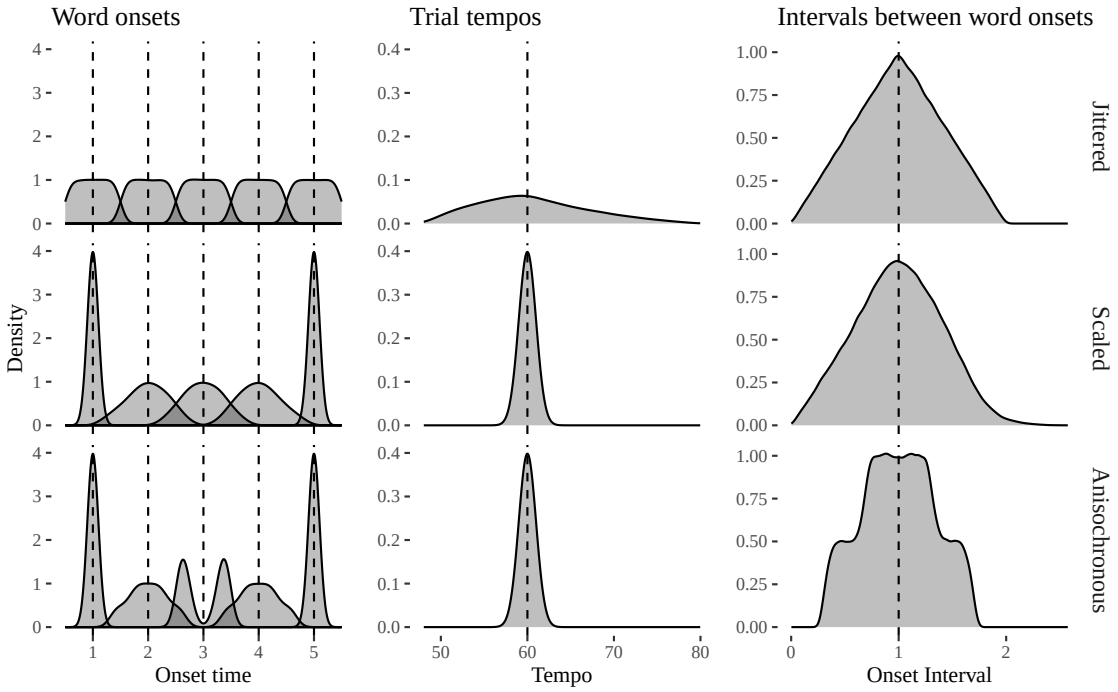

Figure 4.2: Approached to creating anisochronous stimuli. Top row shows a *jittered* approach were random uniform offsets are applied to individual words. Middle row shows the effect of normalising the tempo of the jittered approach. Bottom row shows the method used in the current experiment where a randomly signed beta distribution is used to skew the timing of the central word before jittering intermediate words. Columns show the distributions of onset times a word level; trial tempos across full stimulus items; intervals between word onsets.

demonstrated in the video and then the main experiment task trials. Once these steps were complete the browser window was set to full screen and the main experiment task began. Each trial lasted 4 seconds and required a keyboard response to advance to the next trial. Participants could also end a trial early by pressing any key after responding.

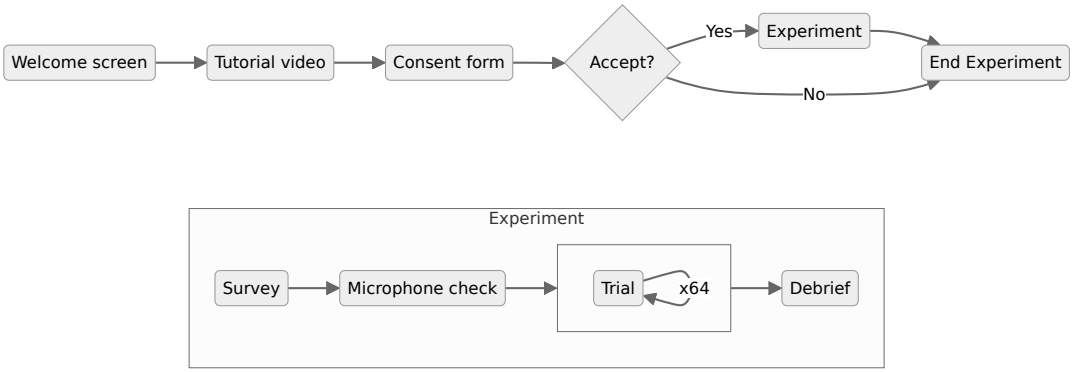

Figure 4.3: Flowchart of Experiment 1 procedure.

4.3.5 Preprocessing

Participants' spoken responses were analysed to identify the p-centre onset in the response word they produced, from which the dependent variable could be calculated, and to exclude any trials with incorrect responses.

Initially onset detection was attempted with the same method used to estimate p-centres for the stimuli recordings. This method performed poorly due to background noise and varying recording quality. Therefore response onsets were estimated using forced alignment to detect the vowel onset. Vowel onsets have been shown to be a precise (low variability) p-centre estimate, tending to closely follow the maximum rate in change of the syllable (Šturm & Volín, 2016). Additionally, as both response words share a common onset phoneme (/r/) this measurement is consistent across both trial types.

Forced alignment was carried out using the Montreal Forced Aligner (McAuliffe et al., 2017) with a custom dictionary containing only two entries: WRONG (/rɒŋ/) and RIGHT (/raɪt/). Labels for the expected response were created programmatically based on the correctness of the sum.

Alignments were visually inspected by plotting spectrograms for each trial using the phonTools (Barreda, 2015) for R (R Core Team, 2023). Spectrograms were displayed using a custom interface using the R shiny package (Chang et al., 2020), grouped by subject and word, allowing fast visual identification of mismatched responses.

Where the forced alignment or the response itself was incorrect, trials were marked for manual annotation or exclusion respectively. However, adding a silence label before the expected response and repeating the forced-alignment led to onsets being detected between the /r/ and

vowel onset of all trials where the first word was the expected response. Therefore, no manual corrections were required as they were excluded as incorrect responses. Examples of incorrect responses were CORRECT instead of RIGHT or NO instead of WRONG. Exclusions also included instances where participants gave the correct response after an incorrect response such as "corre – I mean right".

4.3.6 Model specification

In order to model both the central tendency and dispersion of responses, the `brms` packaged (Bürkner, 2017) for R (R Core Team, 2023) was used. The central tendency of the response is referred to as the `mu` parameter, or simply by the name of the response variable (`onset` in the current experiment). The dispersion of the responses is referred to as `sigma`.

Statistical tests are carried out using the probability of direction (PD) index as described by Makowski et al. (2019). The PD is expressed as a percentage and can be viewed as a Bayesian p-value, where a p-value of .05 would correspond to a one-sided PD of 95% or a two-sided PD of 97.5%. In practical terms, the PD is the percentage of model estimates from posterior draws that have the same sign. Importantly, this is not the probability of an effect in the population, but the probability (expressed as a percentage) of an effect in the available sample of model posteriors. This could lead to the criticism that a PD of 100% simply means there were not enough posteriors draws to sample an estimate in the opposite direction. To convey this all PDs of 100% or that round to 100% are render as "PD $>$ 99.9%" in this and subsequent chapters.

A further indication of the certainty of an effect is provided by the critical interval (CI), measured using the highest density continuous interval (HDCI) of the posterior distribution. Where posterior estimates are normally or otherwise symmetrically distributed, a 95% CI that excludes 0 would correspond to a PD greater than 97.5%.

4.4 Results

The `mu` parameter of the model response (labelled `onset`) was modelled with a maximal formula reflecting the study design including random intercepts and random slopes within-subject (Barr et al., 2013; Oberauer, 2022). The log-normal family was used to capture the distribution of responses.

The tempo predictor was expressed in seconds as the `period` of one complete cycle at the trial tempo, or equivalently as the mean interval between word onsets in the stimulus. For example, a 60 BPM trial would have a period of 1.0 s and a 120 BPM trial would have a period of 0.5 s. The `rhythm` variable was dummy coded with the isochronous category as the reference level (isochronous = 0, anisochronous = 1). This coding ensures that estimates of other parameters correspond to the isochronous condition. Correctness was referred to in models as `word` indicating the appropriate response the sum and deviation coded (RIGHT = -0.5, WRONG = 0.5). Participant IDs, used to model random effects, were included as a variable named `subj`. In the notation used in the `brms` package the formula for the `mu` parameter was:

Model	ELPD diff	SE diff	Sigma formula
m4	0.0	0.0	$sigma \sim period * rhythm * word +$ $(1 + period * rhythm * word \mid subj)$
m2	-11.7	15.1	$sigma \sim period + (1 + period \mid subj)$
m3	-16.9	12.7	$sigma \sim period * rhythm +$ $(1 + period * rhythm \mid subj)$
m1	-23.5	16.7	$sigma \sim (1 \mid subj)$
m0	-258.4	32.5	NA

Table 4.1: Leave-one-out (LOO) model comparisons for experiment 1. Differences between expected log predictive density (ELPD) shown where larger ELPD values indicate preferred model. Sigma parameter of formulas for models shown using syntax from brms package for R.

onset ~ period * rhythm * word + (1 + period * rhythm * word | subj)

A literature search did not return any studies assessing the appropriateness of maximal formulae for the `sigma` parameter. Therefore a model with no `sigma` formula was specified for reference along with four incremental formulae up to a maximal `sigma` formula. A leave-one-out (LOO) evaluation of these five models was conducted to ensure that responses were not better explained by a simpler formula. These models are shown in Table 4.1 with the `sigma` formulae expressed in the notation used in the `brms` package. The model with the maximal `sigma` formula (m4) achieved the highest log predictive density (ELPD) measure. Therefore, only the maximally specified model is reported here.

The current experiment tested predictions of two hypotheses: The tempo entrainment hypothesis, and the rhythm entrainment hypothesis. It was predicted from the tempo entrainment hypothesis that participants would respond faster to trials presented at faster tempos, and that there would be less dispersion of responses at faster tempos. From the rhythm entrainment hypothesis was predicted that responses would be less dispersed in isochronous trials compared to anisochronous trials.

Confirmation of the prediction that participants would respond faster to faster trials would be demonstrated by a positive effect of the `period` predictor on the `mu` parameter of the response (`onset`). The effect of `period` had a probability of 99.94% of being positive (Mean = 0.056, 95% CI [0.023, 0.088]), supporting the tempo entrainment hypothesis. Confirmation of the prediction that responses would be less dispersed in faster trials would be demonstrated by a positive effect of `period` on the `sigma` parameter of the response. The effect of `period` on `sigma` had a probability of 82.76% of being negative (Mean = -0.065, 95% CI [-0.200, 0.068]), providing evidence of an effect in the opposite direction. While the PD of this effect did not meet the 97.5% threshold, this would suggest that responses were more dispersed at faster tempos (shorter periods).

To aid interpretation of the effect of tempo on response times, as opposed to log response times, the slope of the effect of the period predictor was estimated from the posterior fixed-effect

Term	Parameter	Estimate	95% CI Lower	95% CI Upper	PD	
Intercept	mu	0.07	0.01	0.12	98.9%	*
	sigma	-1.64	-1.73	-1.55	> 99.9%	*
Period	mu	0.06	0.02	0.09	99.9%	*
	sigma	-0.06	-0.20	0.07	82.8%	
Rhythm_{aniso}	mu	0.00	-0.01	0.02	60.9%	
	sigma	0.07	0.00	0.13	97.5%	*
Word_{wrong}	mu	0.07	0.05	0.09	> 99.9%	*
	sigma	0.02	-0.07	0.11	67.3%	
Period $\times Rhythm$	mu	-0.02	-0.06	0.03	79.4%	
	sigma	0.00	-0.19	0.20	50.6%	
Period $\times Word$	mu	0.00	-0.06	0.05	54.6%	
	sigma	0.16	-0.12	0.44	87.4%	
Rhythm $\times Word$	mu	0.00	-0.03	0.02	63.2%	
	sigma	-0.06	-0.20	0.08	80.3%	
Period $\times Rhythm \times Word$	mu	-0.06	-0.15	0.02	92.7%	
	sigma	-0.35	-0.75	0.06	95.3%	.

Table 4.2: Experiment 1 model results. Subscript in main effect terms refer to the level of the variable tested in the main effect (and in contrasts) where a positive estimate is in the direction of that level. Parameter indicates the parameter of the response variable the estimate refers to where `mu` is the central tendency (mean) and `sigma` the dispersion (SD). CI = Critical Interval; * = PD > 97.5%; . = PD > 95%.

predictions of the model. Across the two levels of the correctness variable, in the isochronous rhythm condition, the mode of the slope of the effect of period from model draws was 0.12 [0.02, 0.21]. This can be interpreted as response times being 12 ms faster for every 100 ms reduction in the period between stimulus word onsets.

Confirmation of the prediction of the rhythm entrainment hypothesis that dispersion of responses would be greater in anisochronous trials compared to isochronous trials would be demonstrated by a positive effect of `rhythm` on the `sigma` parameter of the response. The effect of `rhythm` on `sigma` had a probability of 97.55% of being positive (Mean = 0.065, 95% CI [0.000, 0.130]), supporting the rhythm entrainment hypothesis.

All model terms are reported in Table 4.2 and shown in Figure 4.4.

4.5 Discussion

The response time prediction of the tempo entrainment hypothesis, that participants would respond faster to trials presented at faster tempos, was supported by the current experiment. This is consistent with previous findings. In a similar study, Corps et al. (2020) found that participants responded faster to trials presented at faster rates (by 19 ms for YES responses

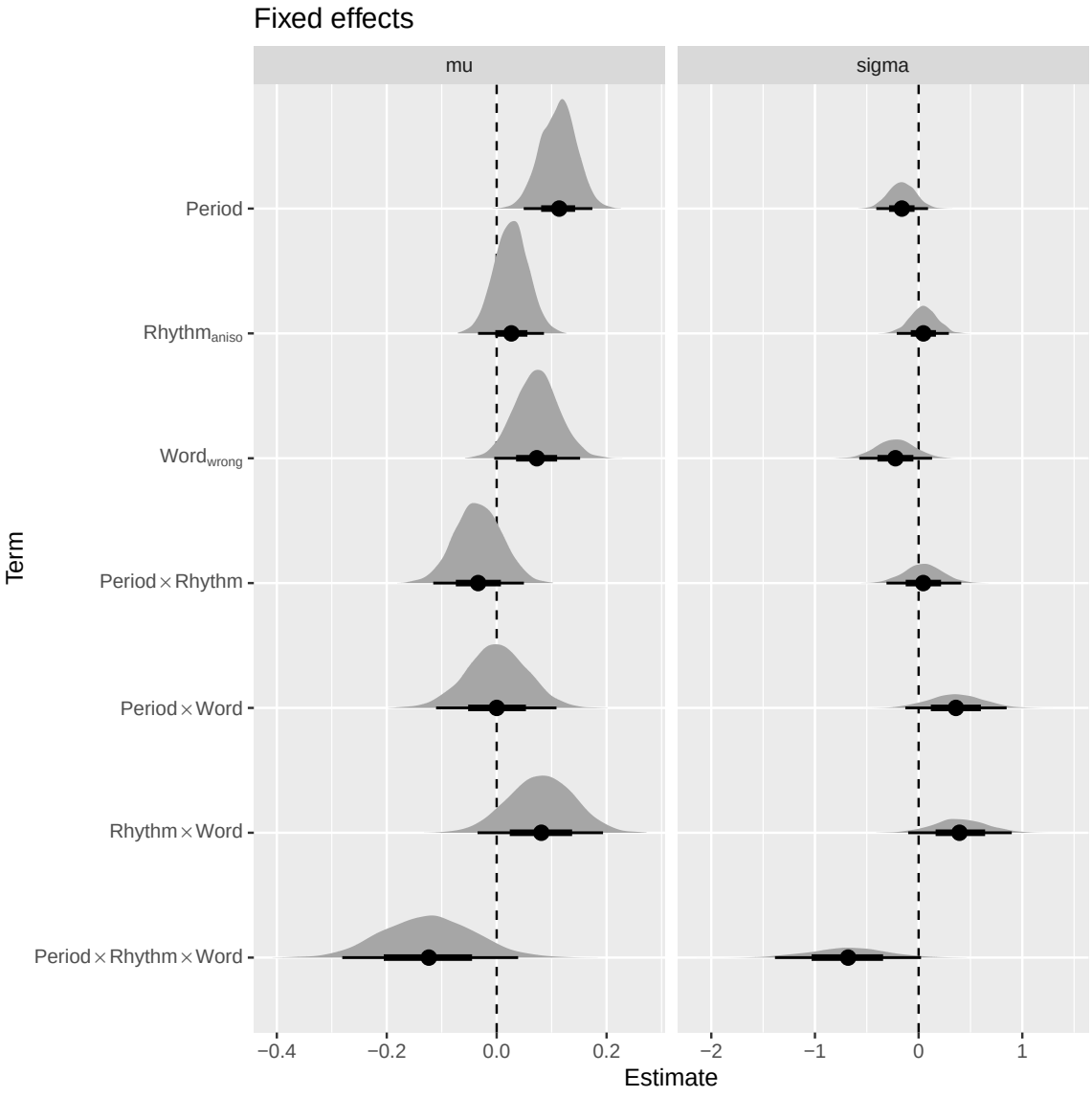

Figure 4.4: Experiment 1 posterior effects.

and 33 ms for NO responses). While these effects appear small, the rates used by Corps et al. were faster (∼278 WPM and ∼557 WPM) than in the current experiment (60 BPM to 120 BPM). Notably, the average WPM rate does not necessarily correspond to the BPM tempo, even though in the current experiment each word was a beat. The periodic BPM measure in the current experiment was intended to facilitate a strong period entrainment. The effect size of ∼12 ms / 100 ms found in isochronous trials would not be consistent with participants responding on the beat. Instead, the effect may be better explained in the time domain without a phase relationship between stimulus and response.

The current experiment tested the dispersion prediction of the tempo entrainment hypothesis that responses would be less dispersed when trials were presented at faster tempos. This prediction was not supported, with evidence instead indicating greater dispersion at faster tempos. The absence of any effect would have provided further evidence of a simple time-domain effect, however, the fact that the detected effect was in the opposite direction could be an indication of an influence of phase. A possible interpretation of this would be that participants struggled to respond on the first beat after the final stimulus onset in faster trials. This could lead to a perturbation effect where not being able to respond on the beat inhibits a response. Alternatively, participants may have attempted to respond on the subsequent beat. Either scenario could result in greater dispersion of responses in faster trials.

The dispersion prediction of the rhythm entrainment hypothesis, that response times would be more variable in anisochronous trials was supported. This finding demonstrates an extension of the rhythm effects in perception found in phoneme detection studies (Cason et al., 2015; Quené & Port, 2005).

The results of the current experiment suggest either a simple time-domain effect of tempo or that a true effect could not be detected with the current design. The latter possibility is worth exploring, given limitations of the online testing set-up used in this experiment. Specifically, response times in the current experiment would have contained a round-trip latency (RTL), which was not corrected as the experiment preceded the development of the latency-test plugin introduced in Chapter 3. The impact of this limitation on the detection of time-domain effects is minimal as RTLs would be expected to be relatively stable within participant. This stability would mean that it should be possible to capture variance due to RTLs the random subject intercept of the model. In contrast, a fixed RTL would result in different errors in a phase measurement depending on the tempo.

In addition to the accuracy of response time measures, there is another design consideration that may have affected sensitivity in the current experiment. In particular, participants were presented with a wide range of tempo and both rhythm conditions within subject and within blocks. Wynn and Borrie (2020) found effect of speech rate priming when stimulus tempos were blocked, but not when they were randomised. This suggests that prolonged exposure to tempo or rhythm conditions would benefit entrainment.

Chapter 5

Experiment 2

5.1 Introduction

In this thesis human speech is characterised as joint behaviour, where perception and production of speech is mediated by a shared temporal context of speaker and listener. The tempo entrainment hypothesis derived from this characterisation makes both weak and strong predictions. In Experiment 1 only the weaker prediction, that responses would be faster when stimuli were presented at faster tempos, was supported. The current experiment refines the design to address limitations of Experiment 1.

5.1.1 Time and frequency domains

Experiment 1 supported the time-domain effects of stimulus tempo on participants' responses found in previous studies (Wynn & Borrie, 2020; Corps et al., 2020; Jungers & Hupp, 2009). In these studies, participants respond faster or produce faster speech in response to faster stimuli. This thesis additionally aims to test a stronger prediction of a frequency-domain effect, that the dispersion of responses would be lower at faster tempos. This prediction was not supported, with results instead suggesting an effect in the opposite direction. This effect did not meet the threshold necessary to reject the null hypothesis, that trial tempo would have no influence on dispersion of responses.

The rate of a stimulus or prime can affect both the time to produce a verbal response and the rate observed in a continuous spoken response. In Jungers and Hupp (2009), participants were asked to listen to a sentence at either a fast or slow tempo and repeat the sentence. The syllable rate of responses was faster when primes were presented at faster tempos. Each sentence repetition trial was followed by a picture describing task. The syllable rate of picture descriptions was also faster following repetition trials with faster primes. Similarly, for a single word verbal response of YES or NO, Corps et al. (2020) found that participants responded faster to spoken questions presented at a faster rate.

While speech rate, as operationalised in the above studies, is closely related to tempo as manipulated here, there are important differences. A tempo of 120 beats per minute (BPM) would

refer to a strong regularity in the temporal structure with events, such as the onset of notes in music or syllables in speech, occurring either on or in relation to this regular beat. Even in highly syncopated musical rhythms, listeners can entrain to the underlying tempo structure of a stimulus (Tal et al., 2017). In contrast, a speech rate of 120 syllables per minute could refer to an average rate in the absence of any observed regularity; equivalently to the speech having an average interval of 500 ms between syllables.

For maths sums, especially isochronously retimed maths sums, a periodic definition of tempo allows for the interpretation of the phase of response.

5.1.2 Response time accuracy

Reviews of web-based experiment builders have indicated acceptable precision, as measured by variability of response time errors (Anwyl-Irvine et al., 2021; Bridges et al., 2020). However, previous studies that employed remote web-based data collection, found auditory response times in replications were slower than lab-based findings (Vogt et al., 2021; Fairs & Strijkers, 2021). The size of this difference in these studies is similar to the findings of the validation of the latency-test plugin in Chapter 3, indicating that these errors are very likely to be the result of the uncorrected systematic errors introduced by round-trip latency (RTL).

In Experiment 1, the unknown contribution of RTL to response time measures was an important limitation. Without accurate measurement of response times, the phase of a response cannot be known. For example, if a participant has a RTL of 100 ms, where the period of the beat cycle is 2π radians, this would result in an error of $\frac{\pi}{5}$ radians when measuring the phase of the response in 60 BPM trials, but an error of $\frac{2\pi}{5}$ when measuring the phase of response in 80 BPM trials. Subtracting the RTL from response times would allow the phase of response to be correctly identified. However, no RTL measurement technique had been implemented in web-based experiment software at the time of data collection for Experiment 1.

5.1.3 Trial blocks

In Experiment 1, both trial tempos and rhythm conditions were randomised within blocks. However, previous research has found that the blocking of trials can influence entrainment effects. Wynn and Borrie (2020) found effect of speech rate priming when stimulus tempos were blocked, but not when they were randomised. This suggests that prolonged exposure to tempo or rhythm conditions would benefit entrainment.

5.1.4 Visual stimuli

A further potential issue of the paradigm employed in Experiment 1 is that visual presentation of the full sum for the duration of the trial allows participants to attend to the visual stimulus rather than the auditory stimulus. The visual stimulus was intended to support the auditory presentation and engagement with the task. It was assumed that participants would read the sums at the same pace as speech stimulus. This assumption may not be met if a participant

	RT ≈ 750 ms	RT ≈ 1000 ms
Target (60 BPM)	Miss	Hit
Competitor (80 BPM)	Correct Rejection	False Alarm

Table 5.1: Example of how response times (RT) could be classified using a signal detection approach. Responses with onsets corresponding to the period of the target tempo would be hits when stimuli are presented at the target tempo, but as false alarms when presented at the competitor tempo.

reads at their own pace. An alternative approach is to use rapid serial visual presentation (Potter, 1984) to match the serialised auditory presentation of words.

The audio-audio plugin design introduced in Chapter 3 allows the presentation of visual stimuli to be scheduled using the same way as auditory stimuli.

5.1.5 Phase amplitude

With accurate response time measurement it becomes possible to measure the phase of responses to determine if the response onset falls on the next beat of the trial. For example, at 60 BPM, there would be 1000 ms between onsets in the stimulus. A response with an onset 1000 ms after the final stimulus onset would be considered to be *on the beat*. This would be seen as evidence of entrainment in the perception of speech modulating the timing of production.

Evidence of entrainment would typically be provided by continuous signals such as neural oscillations (Ding & Simon, 2014) or the amplitude envelope of produced speech (Assaneo et al., 2019; Cummins, 2009). In such cases, synchrony is not defined by the alignment of a single event such as the response onset in the paradigm employed in this thesis. Instead, analysis of continuous signals allow for various measures of coherence (Bastos & Schoffelen, 2016). However, a single point cannot be coherent with a signal.

An alternative to coherence would be to employ signal detection to determine if a response falls on the beat. For example, by introducing a competitor tempo that has low coherence with the target tempo, response times (RT) could be classified as shown in Table 5.1.

In this example the tempos 60 BPM and 80 BPM have been selected because their phase relationship allows separation between the first and second beats of the target and competitor. The appeal of a signal detection approach is that detecting *false alarms* makes it possible to distinguish between responses that happen to fall on the beat and responses that selectively fall on the beat of the trial tempo.

While signal detection may be a viable approach, it would require that responses are categorised with binary levels. This would in turn require response windows to be defined and related decisions to be made about what level of error should be considered a *hit*. Alternatively, these decisions can be avoided by borrowing from coherence techniques. Rather than using a binary response, *hit* and *miss* can be defined continuously to measure the degree of alignment with the target tempo and *correct rejection* and *false alarm* measure the degree of alignment with

the competitor tempo.

An implementation of this proposed method is presented here, where the alignment between the response and phase of the stimulus is referred to as the *phase amplitude*. This is the scaled cosine function of the response phase calculated at the trial tempo. A perfectly in-phase response would have a phase amplitude of 1 and a perfectly anti-phase response would have a phase amplitude of 0.

The phase (θ) in radians of the response was calculated using eq. (5.1) where t refers to the vowel onset time and T refers to the period of one cycle of the trial tempo in seconds.

$$\theta = \frac{2\pi t}{T} \tag{5.1}$$

As a result of the bounded nature of this measure it is possible to model responses using a beta distribution. To facilitate this, the phase amplitude (ϕ_{amp}) measure was calculated and scaled between 0 and 1 using eq. (5.2).

$$\phi_{\text{amp}} = \frac{1 + \cos(\theta)}{2} \tag{5.2}$$

By calculating the phase amplitude at both the target and competitor tempo, it is possible to reflect the advantages of the signal detection approach. Under the null hypothesis that there is no relationship between the response onset and the phase of the stimulus, the phase amplitude would be arbitrary. Furthermore, the amount of dispersion in the phase amplitude distribution will increase with faster tempos as one cycle will correspond to a shorter period of time. This means that comparing phase amplitudes at different tempos would not necessarily be informative. However, it is possible to compare trials where the period used to calculate the amplitude is fixed. For example, if an effect is present 60 BPM trials would be expected to have a higher phase amplitude than 80 BPM trials when phase amplitude is calculated at 60 BPM. Similarly, 80 BPM trials would be expected to have a higher phase amplitude than 60 BPM trials when phase amplitude is calculated at 80 BPM.

It is not expected that the phase amplitude at the target tempo is interpretable without reference to the competitor tempo. For this reason, analyses will also include *phase vector* diagrams to aid interpretation. Here, a phase vector can by though of as an arrow of length 1, at an angle determined by the phase of response. When stored as complex numbers, calculating the mean of these vector provides the circular mean of responses where the angle of the mean is the phase of response and the magnitude is a measure of the concentration of responses. Figure 5.1 compares how different response distributions would look on a phase vector diagram. If participants tend to respond on the beat, the mean phase vector is projected from the origin of the diagram to the right. The mean phase vector is scaled to match the number of beats shown in the diagram, so a mean phase vector of length 5 projected to the right would indicate that all responses were on the the beat. Shorter mean phase vectors projected to the right would

indicate that the mean of responses was on the beat but that responses are more dispersed. Mean phase vectors projected to the left of the diagram would indicate that the mean response phase is in anti-phase or off-beat. These diagrams also show individual phase vectors projected from an origin determined by the time of response. A black spiral is shown on the diagram as a guide and can be interpreted as a time-axis extending from $t = 0$ to $t = 5$.

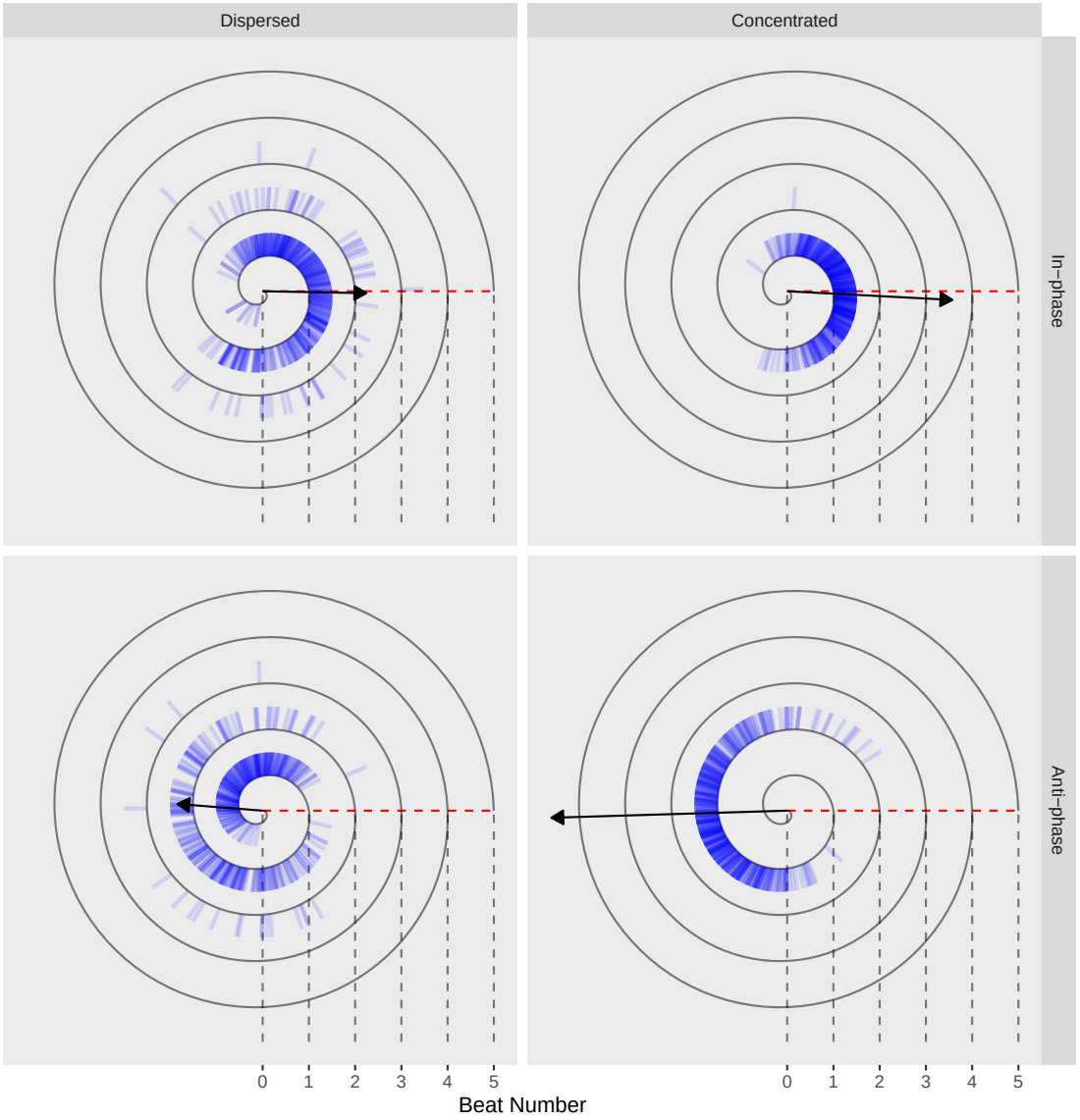

Figure 5.1: Illustrative example of how different distrubutions of response phase would appear in a phase vector diagram. The path of the black spiral represents time; blue lines represent individual response; red dashed line respresents responses that are in-phase (on the beat); arrow represents the mean phase vector scaled between 0 and 5 to match the number of beats.

5.2 Overview

The current experiment adapts the paradigm employed in Experiment 1, but blocks conditions to increase the opportunity for entrainment to occur. Firstly, the tempo manipulation was prioritised by presenting two blocks where tempo was manipulated followed by two blocks where rhythm was manipulated. In the event that participants experience technical difficulties midway through the experiment, this arrangement would maximise completeness of tempo

manipulation trials. This precaution reflects the fact that, while the latency-test plugin had been tested on development computers, it had not yet been deployed for remote web-based data collection.

Experiment 2 also replaces the static text display of Experiment 1 with RSVP, as shown in Figure 5.2. The auditory stimuli also differed from Experiment 1 in that they were concatenated in advance to simplify trial configuration. The resulting stimuli of both methods would be equivalent, with only a difference in how the serialisation of words was achieved.

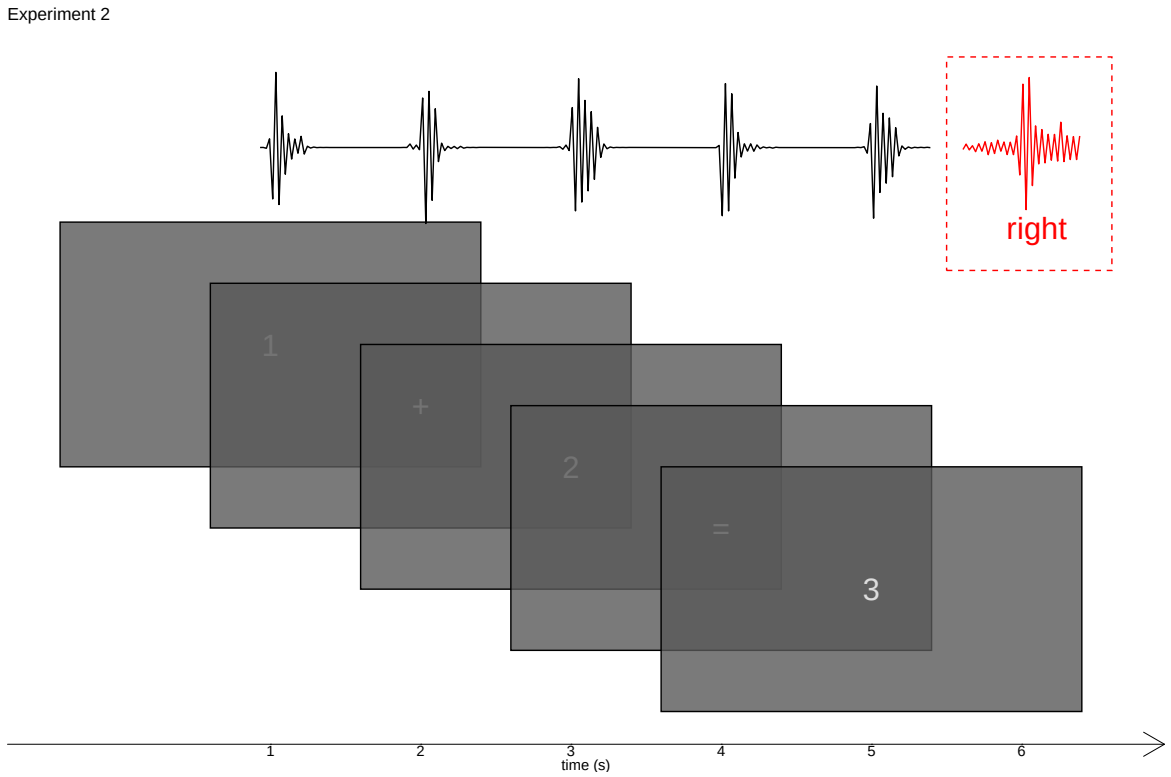

Figure 5.2: Schematic diagram of Experiment 2 paradigm. Upper panel shows auditory stimulus presented as a single audio file. Lower panel illustrates the visual stimulus presenting individual words of the sums, approximately synchronised with the perceptual word onsets. Opacity of backgrounds illustrates progression of time with actual backgrounds being solid black. Diagram not to scale.

1. Tempo entrainment hypothesis
    - Response time prediction: Response times will be faster when trials are presented at faster tempos
    - Response dispersion prediction: Response times will be less dispersed when trials are presented at faster tempos
    - Beat alignment prediction: Phase amplitude will be greater when calculated at trial tempo than at alternative tempo
2. Rhythm entrainment hypothesis
    - Response dispersion prediction: The dispersion of response times will be lower when the stimulus rhythm is more regular

5.3 Methods

5.3.1 Design

This experiment was composed of 4 blocks, where the first 2 blocks manipulated tempo and the final 2 blocks manipulated rhythm.

In the first two blocks, there were two tempos (60 BPM and 80 BPM) presented within-subjects. All trials in these blocks were isochronous. These levels were blocked such that half of participants heard the faster tempo first and the other half heard the slower tempo first. Participants were randomly assigned to one of four lists of 64 sums. In each list half of the sums had incorrect answers. Lists and tempos were counterbalanced such that there were eight pseudo-randomly allocated combinations of list and block order. The first participant was randomly assigned a combination of list and block order, and subsequent participants were randomly assigned a combination from the combinations that had the fewest completed trials.

In the final two blocks, there were two rhythm conditions (isochronous and anisochronous). All trials in these blocks were presented at 70 BPM, a neutral tempo midway between the tempo manipulations. Participants were presented with the same list of 64 sums as in the first 2 blocks, except the incorrect sums were presented with the correct answer and the correct sums with incorrect answers to avoid repeating items. Similarly to the first two blocks, the conditions were blocked and counterbalanced so that half of participants were presented with isochronous trials first and half presented with anisochronous trials first.

5.3.2 Participants

Participants were recruited via the University of Glasgow School of Psychology subject pool and requests posted to social media. Participants that completed the study were given the opportunity to claim £3 for taking part, either as a digital payment or gift voucher. As a requirement of taking part all participants self-reported as native English speakers with normal or corrected to normal hearing and vision. The data collection was stopped when 64 participants completed the experiment. Six participant were excluded due to responding with the wrong word (e.g., CORRECT rather than RIGHT) or not at all in more than 20% of trials. Ethical approval was obtained from the University of Glasgow College of Science and Engineering Ethics Committee.

5.3.3 Materials

The sums used in this experiment were created from the same 64 correct sums and 64 matched incorrect sums as used in Experiment 1. Three isochronous stimulus sets were created for the 60, 70, and 80 BPM tempos; and one anisochronous stimulus set was created for the 70 BPM tempo. As with Experiment 1, words were presented individually with timing of auditory stimuli controlled by a version of the audio-audio plugin outlined in Section 3.2.1.

In contrast to Experiment 1, sums were presented visually as individual numbers and operators.

Timing of visual presentation was cued by the audio-audio plugin as illustrated in Figure 5.2.

5.3.4 Procedure

The procedure is illustrated in Figure 5.3. In the microphone finder trial a volume indicator line allowed participants to move their headphones around their computer to find where the signal was loudest. The trial was looped until the participant indicated that they had successfully located their microphone by pressing the Y key. This trial and the following latency trials were created using the latency test plugin detailed in Section 3.2.2. After the first block of latency test trials, the median absolute deviation (MAD) of latency times was calculated and participants with a MAD of greater than 5 ms over the ten trials were shown a message suggesting possible causes of their high variation and given the option of trying again or end the experiment. The MAD rather than standard deviation (SD) was chosen as trials where detection fails will tend towards very high latencies. Subsequent latency tests did not prevent progression but all latency values and the signals used to calculate them were saved.

5.3.5 Preprocessing

As with Experiment 1, responses varied in terms of the amount of background noise and in recording quality, making them unsuitable for automatic onset detected using the same algorithm as used for stimulus creation. For this reason, the same Montreal Forced Aligner method of vowel onset detection was used as in Experiment 1.

Participants' round-trip latencies were estimated before and after each block. The median was used for the latency-test block estimate, and then mean of the two surrounding latency-test blocks were used for to correct response times in experiment trial blocks.

5.3.6 Model specification

The `brms` package (Bürkner, 2017) for R (R Core Team, 2023) was used to model all responses.

Tempo manipulation blocks The log-normal family was used to model the distribution of responses. Initially, a maximal model formula was specified for the effects of `period`, `word`, and their interaction. Using `brms` notation this gave the formula:

```
onset ~ period * word + (1 + rate * word | subject)
```

Where the section inside brackets is the random subject effect structure. However this formula resulted in large critical intervals around fixed-effect estimates. Further inspection of the model revealed that this was caused by an unintended consequence of counterbalancing the sequence of tempos (period in the model). While tempos were a within-subject variable, when block order effects are considered, the interaction of tempo and block would be a between-subject interaction as participants were only presented with one tempo in each block. Although this hidden interaction was not explicitly included in the formula, the random subject effect structure would allow random slopes for the `period` predictor to capture the variance due to block order.

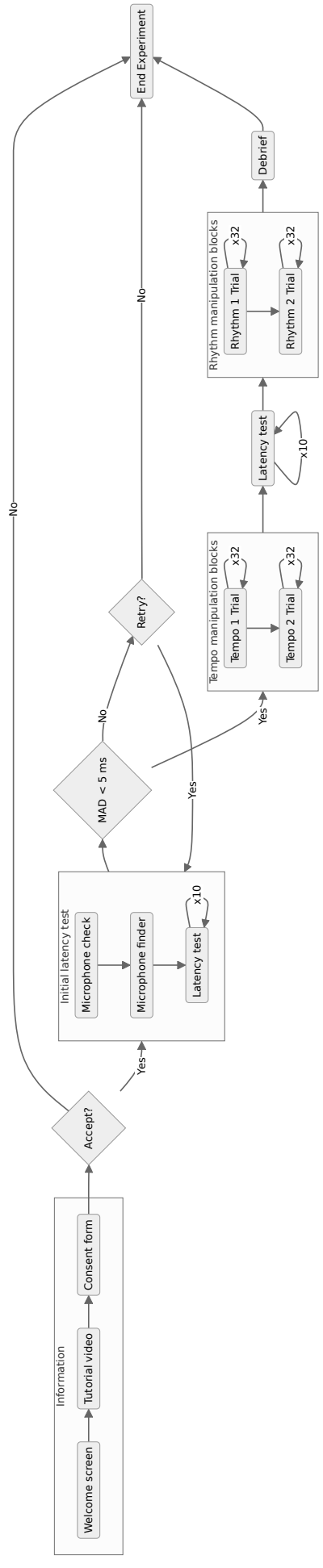

Figure 5.3: Experiment 2 procedure flowchart.

To both account for the block order effect and avoid misattributing its effect to random subject effects, `block` (first = -0.5; second = 0.5) was modelled as a fixed effect and `period` was removed from the random effect structure to give the formula:

```
onset ~ period * block * word + (1 + word | subj)
```

This notation includes all 2-way and 3-way interactions of the fixed effects.

Following the method of Experiment 1 the dispersion of the outcome was modelled with the formula:

```
sigma ~ period * block * word + (1 + word | subj)
```

Due to the simpler random effect structure compared to Experiment 1, no model comparison was performed for simpler `sigma` formulae.

In the frequency-domain model, a similar formula was used with two differences: Firstly, the tempo manipulation was labelled `rate` referring to the relative rates of the tempos (60 BPM = slow; 80 BPM = fast). The `rate` predictor was deviation coded (slow = -0.5, fast = 0.5). Secondly, two response variables were specified using the `mvbind` function (multivariate bind). These response variables were labelled `phase_amp_60` and `phase_amp_80`, referring to the phase amplitude measure at the corresponding reference tempo. In `brms` notation the formula was:

```
mvbind(phase_amp_60, phase_amp_80) ~ rate * block * word + (1 + word |
subj)
```

As phase amplitude is bound between 0 and 1, the beta distribution was used to model responses. When the shape parameters of a beta distribution are less than 1, the distribution can fit two modes at 0 and 1.

Rhythm manipulation responses The same formulae were used for both the `onset` (mu) and `sigma` responses in the time-domain analysis of rhythm blocks, except replacing `period` with `rhythm`. For example, the `onset` formula was:

```
onset ~ rhythm * block * word + (1 + word | subj)
```

The `block` predictor was also coded in the same way (first = -0.5, second = 0.5), and the `rhythm` predictor was deviation coded (isochronous = -0.5, anisochronous = 0.5). Note that this coding differs from Experiment 1 where `rhythm` was dummy coded with isochrony as the reference level in order to make effects interpretable at the isochronous level.

5.4 Results

5.4.1 Tempo entrainment hypothesis

Time domain In order to allow interpretation of the hypothesised effects, the confounding effect of block was modelled. The apparent effect of block, where participants responded faster in the second block would be evidenced by a negative effect of `block`. The effect of `block` had

			95% CI	
Block	Word	Mode	lower	upper
first	right	0.11	-0.36	0.62
	wrong	0.03	-0.38	0.62
second	right	0.12	-0.39	0.51
	wrong	0.08	-0.43	0.51

Table 5.2: Fixed-effect slope estimates for period effect in Experiment 2

a probability of 56.60% of being negative (Mean = -0.052, 95% CI [-0.800, 0.762]), providing inconclusive evidence in the expected direction. Similarly, a negative effect of the interaction of `block` and `period` would indicate a moderating effect of `block` on `period`. The effect of this interaction term had a probability of 50.98% of being negative (Mean = -0.022, 95% CI [-0.952, 0.830]), again providing inconclusive evidence in the expected direction. The wide 95% CI estimates of these effects (see Figure 5.4 for visualisation) suggest that, when controlling for tempo, experiment block was a poor predictor of response onset.

The response time prediction of the tempo entrainment hypothesis was that response times would be faster in 80 BPM trials compared to 60 BPM trials. This hypothesis would be supported by a positive effect of the `period` predictor. The effect of `period` had a probability of 99.90% of being positive (Mean = 0.081, 95% CI [0.029, 0.132]), supporting this hypothesis. As this effect estimate is expressed in log units, the mode and 95% critical interval (CI) of the slope was estimated from the fixed-effect model response expressed in seconds. These results are shown in Table 5.2 as the mode and 95% critical interval of the effect of period expressed in seconds. These effects are broken down by `block` and `word` (sum correctness). For example, the estimate for the correct sums in the second block was 0.12, 95% CI [-0.39, 0.51]. The wide 95% CI of this estimate reflects that it is drawn from posterior fixed effects across blocks. They should therefore be treated cautiously as an approximation of a more easily interpreted effect. In the example, this would be an effect of 12 ms for each 100 ms increase in the period associated with the trial tempo.

It was also hypothesised that there would be less dispersion in response times at faster tempos. This would be supported by a positive effect of `period` on the `sigma` term. The effect of `period` on the `sigma` parameter had a probability of 99.52% of being negative (Mean = -0.281, 95% CI [-0.490, -0.066]), providing strong evidence of an effect in the opposite direction. In other words, as was found in Experiment 1, participants' response times were more variable at faster tempos.

All remaining model terms are reported in Table 5.3 and shown in Figure 5.4.

Frequency domain The beat alignment prediction of the tempo entrainment hypothesis was that response onsets would be more likely to be in phase with the trial tempo than the alternative tempo. This was modelled using a beta distribution of the phase amplitude as described

			95% CI			
Term	Parameter	Estimate	Lower	Upper	PD	
Intercept	mu	-0.04	-0.11	0.03	87.9%	
	sigma	-1.38	-1.58	-1.17	> 99.9%	*
Period	mu	0.08	0.03	0.13	99.9%	*
	sigma	-0.28	-0.49	-0.07	99.5%	*
Block_2	mu	-0.05	-0.80	0.76	56.6%	
	sigma	0.49	-0.88	1.84	76.9%	
Word_{wrong}	mu	0.07	-0.02	0.16	93.8%	
	sigma	0.14	-0.21	0.51	79.0%	
Period $\times Block$	mu	-0.02	-0.95	0.83	51.0%	
	sigma	-0.51	-2.04	1.05	75.3%	
Period $\times Word$	mu	-0.04	-0.14	0.07	75.8%	
	sigma	-0.21	-0.61	0.19	85.2%	
Block $\times Word$	mu	-0.03	-0.25	0.19	59.7%	
	sigma	0.14	-1.01	1.21	59.7%	
Period $\times Block \times Word$	mu	0.01	-0.23	0.26	53.1%	
	sigma	-0.04	-1.27	1.25	52.9%	

Table 5.3: Experiment 2 time-domain results for tempo blocks. Subscript in main effect terms refer to the level of the variable tested in the main effect (and in contrasts) where a positive estimate is in the direction of that level. Parameter indicates the parameter of the response variable the estimate refers to where `mu` is the central tendency (mean) and `sigma` the dispersion (SD). CI = Critical Interval; * = PD > 97.5%; . = PD > 95%.

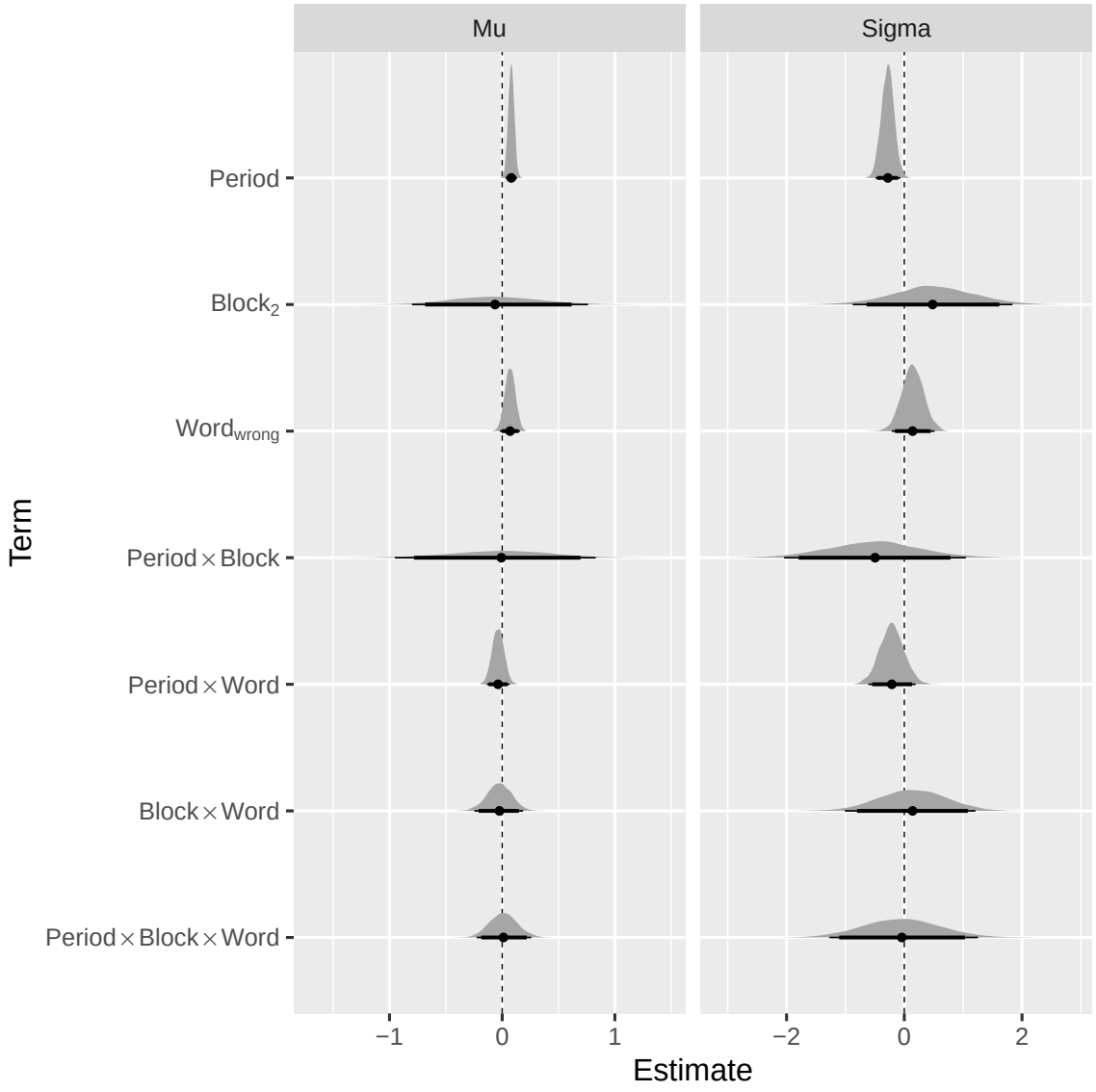

Figure 5.4: Experiment 2 time-domain model posteriors for tempo blocks.

in Section 5.1.5. Strong evidence of this requires effects to be symmetrical, such that the effect of rate is positive when phase amplitude is calculated at 80 BPM and negative when phase amplitude is calculated at 60 BPM. The effect of `rate` at 80 BPM had a probability of $>$ 99.99% of being positive (Mean = 0.227, 95% CI [0.147, 0.307]), and the effect of `rate` at 60 BPM had a probability of $>$ 99.99% of being negative (Mean = -0.205, 95% CI [-0.278, -0.131]). These effects support the hypothesis.

No effect of `block` on phase amplitude was predicted as the experiment block has no frequency-domain interpretation. However, the effect of `block` at 80 BPM had a probability of $>$ 99.99% of being positive (Mean = 0.318, 95% CI [0.238, 0.396]), and the effect of `block` at 60 BPM had a probability of $>$ 99.99% of being negative (Mean = -0.185, 95% CI [-0.259, -0.113]). These effects would suggest that participants were more likely to respond on the beat of a 80 BPM tempo in the second block and the beat of a 60 BPM tempo in the first block, regardless of the trial tempo. As this has no frequency-domain interpretation, this effect suggests further unintended consequences of the blocked design. Alternatively, time-domain effects may have aligned with the phases of reference tempos such that responses tended to fall between beats and faster or slower responses would therefore align more with the faster tempo when faster, and align more with the slower tempo when slower.

All other effects are reported in Table 5.4 and shown in Figure 5.5. Note that the symmetrical effects of the $Rate \times Word$ interaction does have a frequency-domain interpretation, that phase amplitudes were more likely to be greater at the trial tempo when the sum was incorrect. This effect was not predicted and could alternatively be explained by a coincidental time-domain effect.

To further explore potential effects of phase, the phase responses were plotted as phase vector diagrams in Section 5.1.5. Each blue line is a participant's response, projected out from the time and phase of response. Responses that occur on multiples of the trial period (i.e., on the beat) would be projected horizontally to the right of the beat markers indicated on the x-axis. The average phase of response is shown by the black arrow. This arrow is scaled between 0 and 5 to match the scale of the plot, where an arrow of length 5 would indicate that all responses were perfectly in phase. For comparison, the average phase of trials in the same block at the alternative tempo are shown in red. These plots can be interpreted in multiple ways. Firstly, in all cases the angle of the black arrow is closer to the horizontal than the red arrow, indicating that responses were more likely to be in phase when phase was calculated at the trial tempo. Secondly, in all cases the x-coordinate of the end of the arrow is greater when phase was calculated at the trial tempo. This measure is comparable to the *imaginary* component of coherence (Bastos & Schoffelen, 2016)[1]. Finally, the magnitude of the arrows can be interpreted as a measure of dispersion, with shorter arrows indicating more phase variation. In the first block, shown in the upper panels, the black arrow in the 80 BPM column represents the same responses as the red arrow in the 60 BPM column. The arrow is shorter when phase

[1]While it is conventional to show the imaginary component of a vector on the y-axis, it is shown here on the x-axis to align with the beat markers.

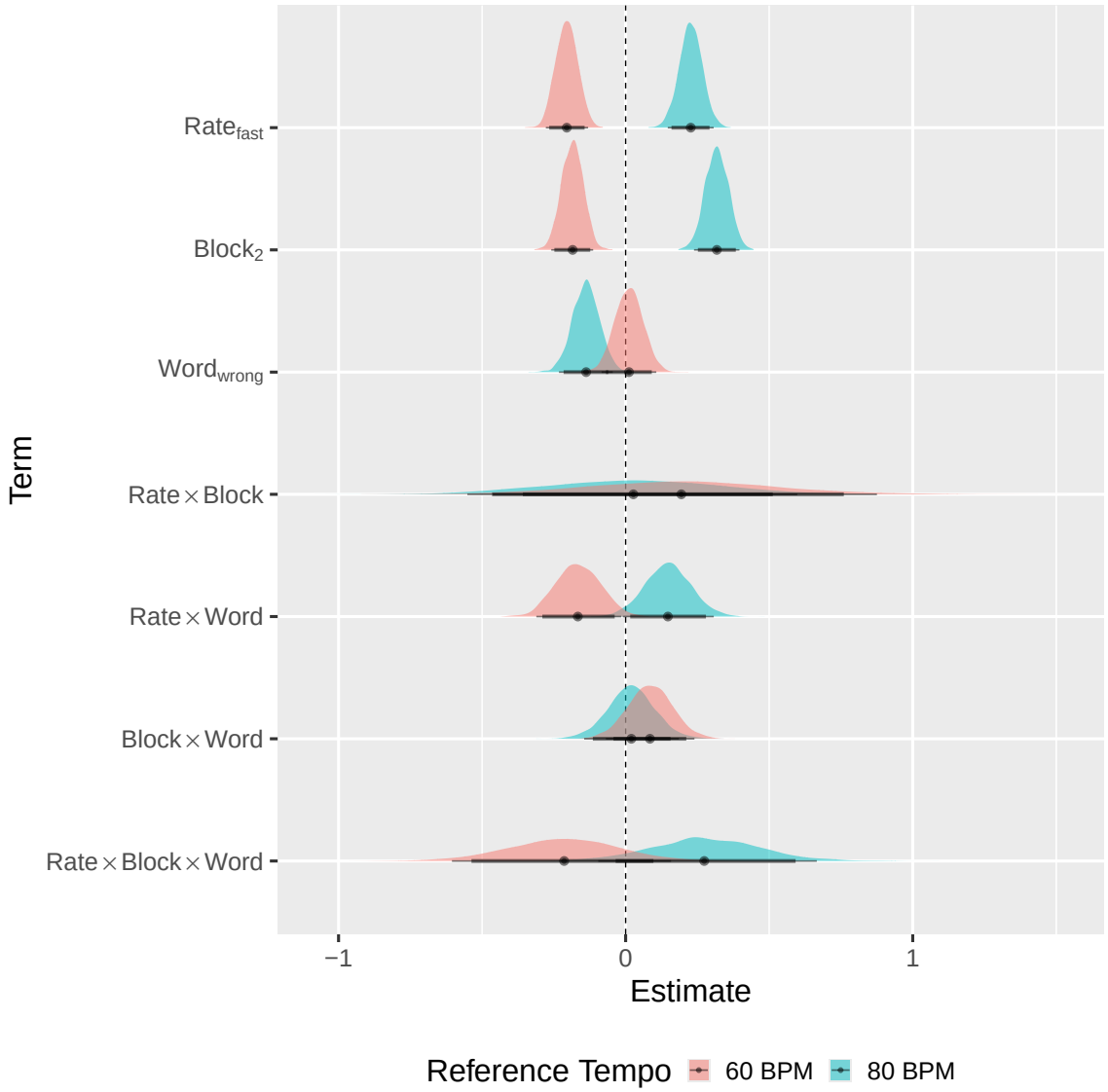

Figure 5.5: Experiment 2 frequency-domain results. Estimate on x-axis refers to the model estimate for phase amplitude model. Labels on y-axis refer to model terms where the subscript indicates the positive level. Distribution fill colours indicate the reference tempo used to calculate the phase amplitude. As the "fast" level of the rate parameter is positive, a positive estimate for the 80 BPM (i.e., fast) reference tempo is expected and a negative estimate for the 60 BPM (i.e. slow) reference tempo is expected.

Term	Ref. Tempo	Estimate	95% CI Lower	95% CI Upper	PD	
Intercept	60 BPM	0.48	0.31	0.65	> 99.9%	*
	80 BPM	0.12	-0.02	0.27	95.4%	.
Rate_{fast}	60 BPM	-0.21	-0.28	-0.13	> 99.9%	*
	80 BPM	0.23	0.15	0.31	> 99.9%	*
Block_2	60 BPM	-0.19	-0.26	-0.11	> 99.9%	*
	80 BPM	0.32	0.24	0.40	> 99.9%	*
Word_{wrong}	60 BPM	0.01	-0.08	0.11	59.3%	
	80 BPM	-0.14	-0.23	-0.04	99.6%	*
Rate $\times Block$	60 BPM	0.20	-0.46	0.88	72.2%	
	80 BPM	0.02	-0.55	0.60	53.9%	
Rate $\times Word$	60 BPM	-0.17	-0.31	-0.02	98.5%	*
	80 BPM	0.15	-0.01	0.31	96.5%	.
Block $\times Word$	60 BPM	0.09	-0.07	0.24	86.5%	
	80 BPM	0.02	-0.14	0.19	60.2%	
Rate $\times Block \times Word$	60 BPM	-0.22	-0.61	0.16	87.5%	
	80 BPM	0.28	-0.10	0.67	92.8%	

Table 5.4: Experiment 2 frequency-domain results.

is calculated at the faster tempo because the same time-domain variance corresponds to greater phase variance. Here, the short black arrow in the 80 BPM column could be an indication that participants struggled with the faster tempo when presented with it in the first block, resulting in greater variability in the phase of responses.

5.4.2 Rhythm entrainment hypothesis

The response dispersion prediction of the rhythm entrainment hypothesis was that there would be greater dispersion of response times in the anisochronous condition compared to the isochronous condition. This would be supported by a positive effect of `rhythm` on the `sigma` response parameter. The effect of `rhythm` on `sigma` had a probability of $> 99.99\%$ of being positive (Mean = 0.134, 95% CI [0.083, 0.185]), confirming this prediction. All other effects are reported in Table 5.5 and shown in Figure 5.7.

5.5 Discussion

5.5.1 Tempo entrainment hypothesis

Time domain Two predictions were made from the tempo entrainment hypothesis. The response time prediction, that response times will be faster when trials are presented at faster tempos, was supported. As with Experiment 1, the estimated effect was similar in size to that found by Corps et al. (2020).

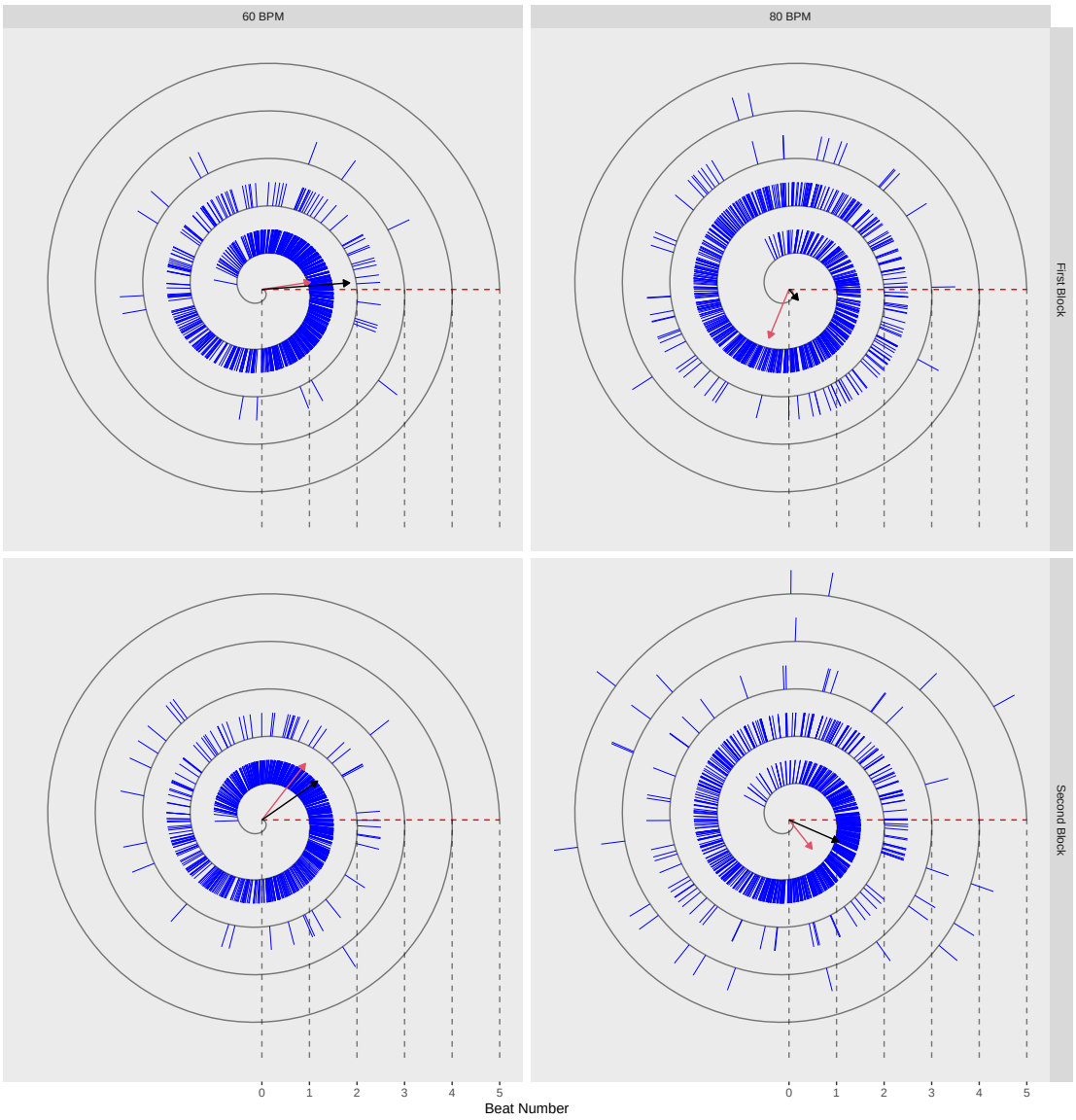

Figure 5.6: Response times expressed as phase of response at trial tempo. Blue segments show individual participant responses. Black arrow shows the average response vector (multiplied by 5 to match the limits of the plot). Red arrow shows the average response vector for trials from the altentative tempo block with phase calculated at the same labelled tempo. That is, red vector in upper right panel is average response of 60 BPM trials in the first block with phase calculated at 80 BPM. Dashed lines projected from x-axis indicate the beat associated with an in-phase response for each cycle.

			95% CI			
Term	Parameter	Estimate	Lower	Upper	PD	
Intercept	mu	-0.03	-0.09	0.02	88.1%	
	sigma	-1.59	-1.70	-1.48	> 99.9%	*
Rhythm_{aniso}	mu	-0.01	-0.02	0.01	80.5%	
	sigma	0.13	0.08	0.19	> 99.9%	*
Block_2	mu	-0.01	-0.02	0.00	92.7%	
	sigma	-0.13	-0.18	-0.08	> 99.9%	*
Word_{wrong}	mu	0.03	0.02	0.05	> 99.9%	*
	sigma	-0.04	-0.12	0.05	79.1%	
Rhythm $\times Block$	mu	0.07	-0.15	0.32	75.1%	
	sigma	0.10	-0.35	0.56	67.3%	
Rhythm $\times Word$	mu	0.00	-0.02	0.03	62.5%	
	sigma	-0.10	-0.20	0.01	96.5%	.
Block $\times Word$	mu	0.02	-0.01	0.04	93.2%	
	sigma	0.12	0.01	0.22	98.9%	*
Rhythm $\times Block \times Word$	mu	0.06	-0.01	0.13	95.7%	.
	sigma	0.15	-0.19	0.50	80.9%	

Table 5.5: Experiment 2 results for rhythm blocks. Subscript in main effect terms refer to the level of the variable tested in the main effect (and in contrasts) where a positive estimate is in the direction of that level. Parameter indicates the parameter of the response variable the estimate refers to where `mu` is the central tendency (mean) and `sigma` the dispersion (SD). CI = Critical Interval; * = PD > 97.5%; . = PD > 95%.

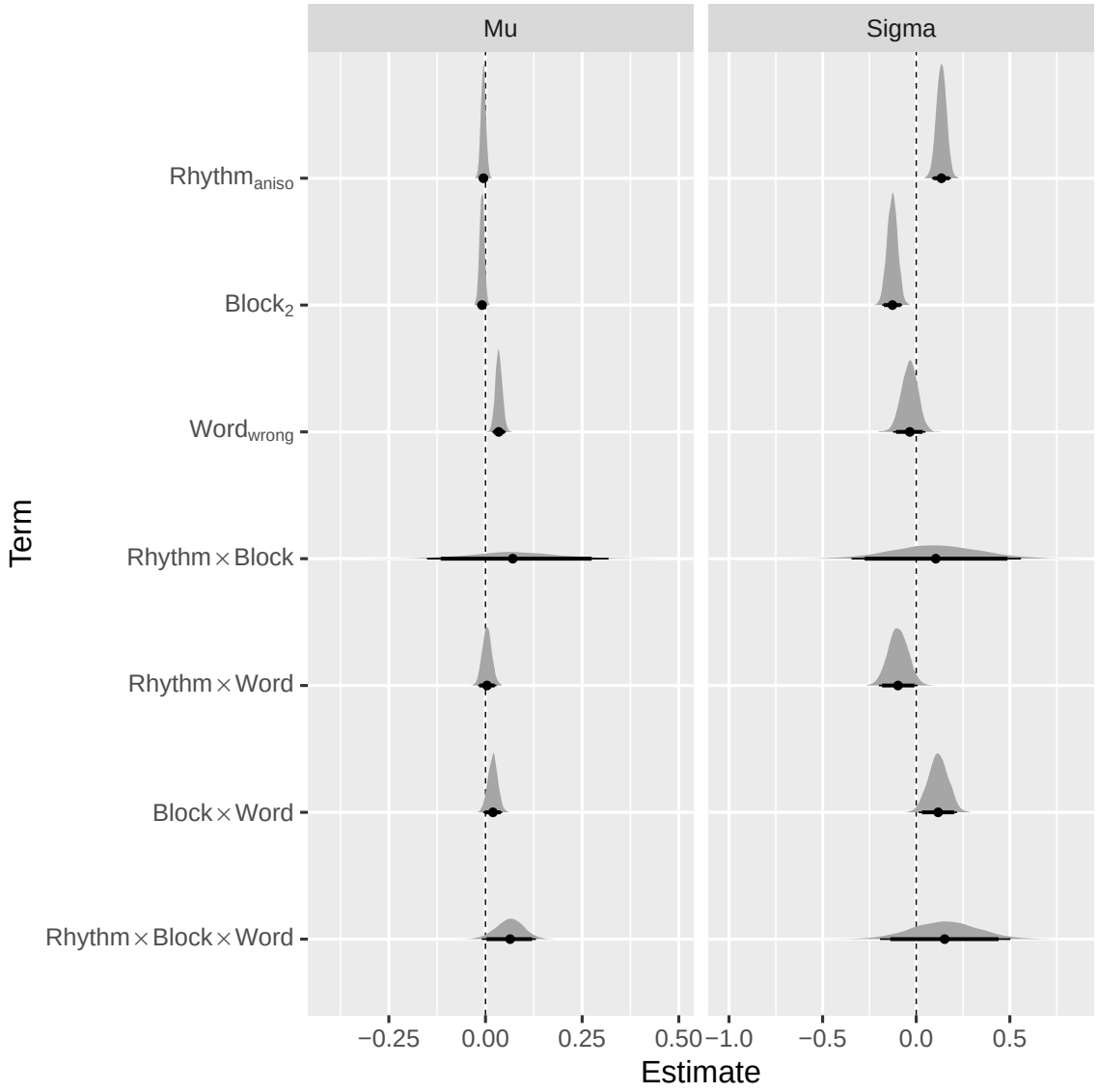

Figure 5.7: Experiment 2 model posteriors for rhythm blocks.

The response dispersion prediction of the tempo entrainment hypothesis was that response times would be less variable at faster tempos. Support of this hypothesis would indicate a strong effect of entrainment where the entire distribution of responses is affected by the stimulus tempo as opposed to just the central tendency of the distribution. As with Experiment 1, this hypothesis was not supported in the current experiment. Instead, strong evidence was found in the opposite direction. This effect could be explained by a perceived increase in the difficulty of the task at the higher tempo. However, trials at the faster tempo allowed the same amount of time to respond as the slower tempo.

Frequency domain The beat alignment prediction of the tempo entrainment hypothesis was tested using the phase-amplitude measure described in Section 5.1.5. This prediction, for the phase amplitude model, is that responses would be more likely to be in phase with the trial tempo than an alternative tempo. To make this possible, the latency-test plugin presented in Section 3.2.2 was used to calculate each participant's median round-trip latency and correct response times. In turn, this allowed the phase of the response to be measured using the phase amplitude detailed in Section 5.1.5.

Results of this phase analysis supported this hypothesis with a symmetrical effect showing that in the 80 BPM trials phase amplitudes were both higher when calculated at 80 BPM and lower when calculated at 60 BPM. Although this symmetry requirement was intended to minimise the risk of type I errors (false positives), the unexpected symmetrical effect of block suggests that these findings should be treated with caution. Further development of this method will be required to establish the type I error rate. Additionally, increasing the number of levels of the tempo manipulation would mean that coincidental alignment of time-domain effect with frequency-domain periodicities would not be found across tempo pairs. Despite the necessary caution in interpretation of phase amplitude effects, these findings were also supported by the additional phase analysis presented in Figure 5.6. Together, the evidence presented suggests that strong frequency-domain evidence of entrainment in this experimental paradigm may be found with further refinement of the experimental design.

5.5.2 Rhythm entrainment hypothesis

The rhythm entrainment hypothesis tested in the current experiment was that entrainment would be more likely in the isochronous condition compared to the anisochronous condition. This was supported by the response dispersion prediction. However, an effect of block is also apparent in the results shown in Table 5.5 and Figure 5.7.

5.5.3 Design strengths and limitations

Blocking The choice to block tempo conditions was intended to maximise exposure to each condition. This was important to eliminate the possibility that entrainment may require prolonged exposure to the trial tempo (Wynn & Borrie, 2020). However, as there was only one block for each tempo, this introduced an unintentional confound of block order. While all

participants were exposed to a first and second block, and all participants were exposed to a fast and slow tempo, the interaction of block sequence and tempo was only identifiable between subjects. This lead to two unplanned model requirements. Firstly, that it was not possible to model by-subject random slopes for the tempo predictor. Doing so would risk variability caused by this interaction being modelled as random subject-level variability. Secondly, it was necessary to model block sequence as a fixed-effect. As a consequence of this limitation it is not clear from the current experiment whether blocking tempos enhances the effect of tempo on response time.

Phase amplitude Although phase measures are common in neuroscience (Bastos & Schoffelen, 2016) and have been employed in speech entrainment studies (e.g., Assaneo et al., 2019), these measures refer to the coherence of the phase relationship between two continuous signals rather than the alignment of a response with the implied phase of a continuous signal. The term *phase amplitude* is used here to avoid implying a coherence relationship between stimulus and response.

While this approach is novel for speech production, it resembled approaches used in speech perception tasks. For example, in a phoneme detection task such as that use by Quené and Port (2005), the target syllable is presented at a point in time presumed to correspond to the phase of the prior syllables. If findings of such studies is interpreted as evidence of listeners entraining to the phase of presentation, under a model of speech as joint action this would be expected to extend to production.

As with Experiment 1, Experiment 2 found an effect of tempo on dispersion of responses. As evidence for the null hypothesis would be no effect of tempo, the unexpected direction of the effect may be explained by an alternative tempo effect. For example, faster stimuli could be perceived as being more urgent than slower stimuli. While this would be evidence of an effect of tempo it would not be evidence of entrainment as described in this thesis. This simpler time-domain influence of tempo could be that tempo was perceived as an implicit cue to respond faster.

Speaker turn In the sum evaluation paradigm employed in both the current experiment and Experiment 1, participant responses constitute a new speaker turn. In terms of functional grammar, the stimulus functions as an interrogative initiation that demands a polar response (Halliday & Matthiessen, 2013). This exchange is realised as the full clause of the stimulus and the elliptical clause of the response. As such, the participant's response is a distinct grammatical unit. Under an oscillatory model of turn-taking, while a phase relationship would be expected, this would not necessarily mean that the response is in phase (M. Wilson & Wilson, 2005). For example, an anti-phase relationship between speakers could reduce the likelihood of cross-talk in conversation.

A further potential complexity of the sum evaluation paradigm is that sums were recorded as unmarked statements but presented contextually as questions. The falling intonational contour

that would be unmarked in a statement would be a marked realisation in an interrogative (Halliday & Matthiessen, 2013). When presented as an interrogative, could be interpreted as a peremptory question requiring the participant to accept or refute. Exploration of the grammatical parameter space that could be manipulated in the sum evaluation paradigm would go beyond the scope of this thesis. Instead the intention is to employ the simplest paradigm that would test the hypotheses.

A simple alternative would be to ask participants to complete sums rather than evaluate them. In conversation, strict alignment between grammatical units and turns is not the norm. Two speakers can realise a single grammatical unit (Lerner, 1996). In a sum completion paradigm, participants may be more likely to maintain both the tempo and phase of the stimulus.

5.5.4 Conclusion

The current experiment provides further support for the tempo entrainment hypothesis. However, replication of these findings in different tempos is essential to rule out a coincidental effect of phase amplitude. It is likely that refinement of the experimental design could help to identify conditions under which participants align their responses more closely to the phase of the trial.

Chapter 6

Experiment 3

6.1 Introduction

The first two experiments set out to investigate whether participants would implicitly time responses to land on the beat of the trial tempo. In both experiments an effect of tempo was detected. However, these effects were modest, with models suggesting participants respond approximately 12 ms faster for every 100 ms reduction in the interval between word onsets in the stimulus. Furthermore, while the second experiment incorporated the latency-test plugin presented in Chapter 3, evaluating effects of tempo on the phase of response still proved challenging.

In the final of the three experiments presented here, the aim is to determine if a strong phase-locked response is observed under optimal conditions. To do so, it is necessary to address both methodological concerns and aspects of the sum evaluation paradigm.

6.1.1 Methodological concerns

Block effects In an attempt to maximise the opportunity for entrainment to occur, tempos were presented in blocks in Experiment 2. An unintended consequence of this blocking was that tempo effects were confounded with block effects. Response times tended to be faster in the second block. While this was not surprising, the impact of this on the ability to model responses was unexpected. While both tempo and block were within subject, the interaction of tempo and block was effectively between subjects. This problem was dealt with by including the experiment block and its interaction with tempo in models as a fixed effect. However, this meant that the time-domain model could not fully reflect the intended design as random within-subjects slopes of the main predictor could not be specified (Barr et al., 2013).

This limitation also had consequences for the frequency-domain model, as an unexpected effect of block was found. The inclusion of a competitor tempo in the phase amplitude method described in Section 5.1.5 was intended to distinguish between arbitrary beat alignment and selective beat alignment. However, as the order of tempos in blocks were counterbalanced, a main effect of block is not interpretable.

Both consequences of the blocked design could be addressed with additional levels of the tempo manipulation. Careful selection of tempos is necessary for effects to be meaningfully interpreted in a phase amplitude analysis. The 60 BPM and 80 BPM pair were selected because they have large phase differences at the first and second beat of both tempos. This would be true of any pair of tempos that have the same ratio. Therefore, it would be expected that a true frequency-domain effect could be detected with different tempo pairs that have the same 3:4 ratio as the 60:80 pair. Detecting an effect within pairs over multiple groups would provide stronger evidence than could be obtained from a single pair.

In order to maximise exposure to additional levels of the tempo condition, it was necessary to prioritise tests of the tempo entrainment hypothesis over the rhythm entrainment hypothesis. Compelling evidence of an effect of rhythm would require that the predicted response pattern is observed in isochronous speech. Therefore, the current experiment focuses on tempo effects and does not include a rhythm manipulation.

Remote data collection All data collection was carried out remotely in the first two experiments. This led to the technical challenges addressed in Chapter 3, and in particular the latency-test plugin that allowed for the correction of the systematic error that resulted from audio round-trip latency (RTL). Due to COVID-19 restrictions, this technique was not validated prior to data collection for Experiment 2. Therefore, the current experiment was the first opportunity to perform a validation in a lab under controlled conditions.

In addition to technical challenges, remote data collection presents practical challenges. With participants taking part at home, there is minimal control over the environment they have available or choose to use. Inspection of trial recordings during data processing revealed high levels of background noise. Furthermore, the quality of microphone and audio interfaces can vary greatly. These limitations were evident in the poor performance of the algorithm used for onset detection in stimuli when applied to participants' responses.

Remote data collection also limits the ability to ensure that participants follow instructions. For example, participants were asked to wear headphones to minimise external distractions but compliance with this was not confirmed. Detecting headphone use in web-based experiments is an ongoing area of development with a recent study demonstrating an accuracy of 80% (Milne et al., 2021). A similar approach was developed for evaluation in the current experiment. However, this is not reported as it is under development and has not yet been validated or compared to alternative approaches.

6.1.2 Paradigm

Speaker Turn The sum evaluation paradigm, where listeners hear a complete sum such as FIVE PLUS THREE IS NINE) and evaluate it as RIGHT or WRONG, may not offer the ideal conditions for speech as joint-action.

The sum evaluation paradigm may not be the ideal conditions for entrainment. In this paradigm

the response is a separate grammatical unit that constitutes a new speaker turn. A possible implication of this on the tempo entrainment hypothesis is that any entrainment that occurs during stimulus presentation could be interrupted by the end of the speaker turn. This would be consistent with the role of a phase-reset mechanism in neural entrainment (Giraud & Poeppel, 2012) where salient *edges* in the amplitude envelope correct for changes or drift in the phase of the stimulus.

An alternative possibility is that responses could be phase-locked with the stimulus across speaker turns without necessarily being in-phase. This would be consistent with oscillatory models of conversational entrainment (M. Wilson & Wilson, 2005). However, in conversation, speakers tend to minimise silence between turns (Stivers et al., 2009). Given that the current study addresses other aspects of the design, a within-turn paradigm would need to be presented in addition to the sum evaluation paradigm.

To address the potential impact of speaker turn, the sum completion paradigm is proposed. In this variation of the paradigm, participants are presented with an incomplete sum and required to answer with the appropriate number. By engaging in collaborative completion (Lerner, 1996), it would be expected that responses would be more likely to be in-phase with the stimulus than when evaluating across speaker-turns.

Establishing a baseline The detection of apparently modest effects in the first two experiments raises the question of what sort of effect should be considered strong evidence of entrainment. Responses with a strong central tendency to occur one beat after the final stimulus onset would be the strongest form of evidence of entrainment. Under conditions where a strong effect of entrainment is present, it is not obvious that participants would necessarily produce such a response. Studies that measure finger-tapping, show that participants tapping along to a beat tend to speed up when the auditory stimulus is removed (Repp & Su, 2013). Although there is evidence of high levels of accuracy under optimal conditions in synchronous speech tasks (Assaneo et al., 2019; Cummins, 2009), these studies do not tell us how well participants can coordinate their speech to another speaker's rhythm after the stimulus is discontinued.

The cognitive demands of the experimental task employed in this thesis are expected to be minimal. However, participants may struggle to meet these demands under the time pressure of faster tempos.

In order to provide a measure of optimal performance, the current experiment explicitly asks participants to respond on the beat in separate blocks of trials presented after completing the implicit trials corresponding to the paradigm employed in the first two experiments.

6.2 Overview

6.2.1 Tempos

As with Experiment 2, tempos were blocked to provide the opportunity to entrain to the tempo across trials. The ratio between tempos in blocks was the same as for Experiment 2 (3:4; 60

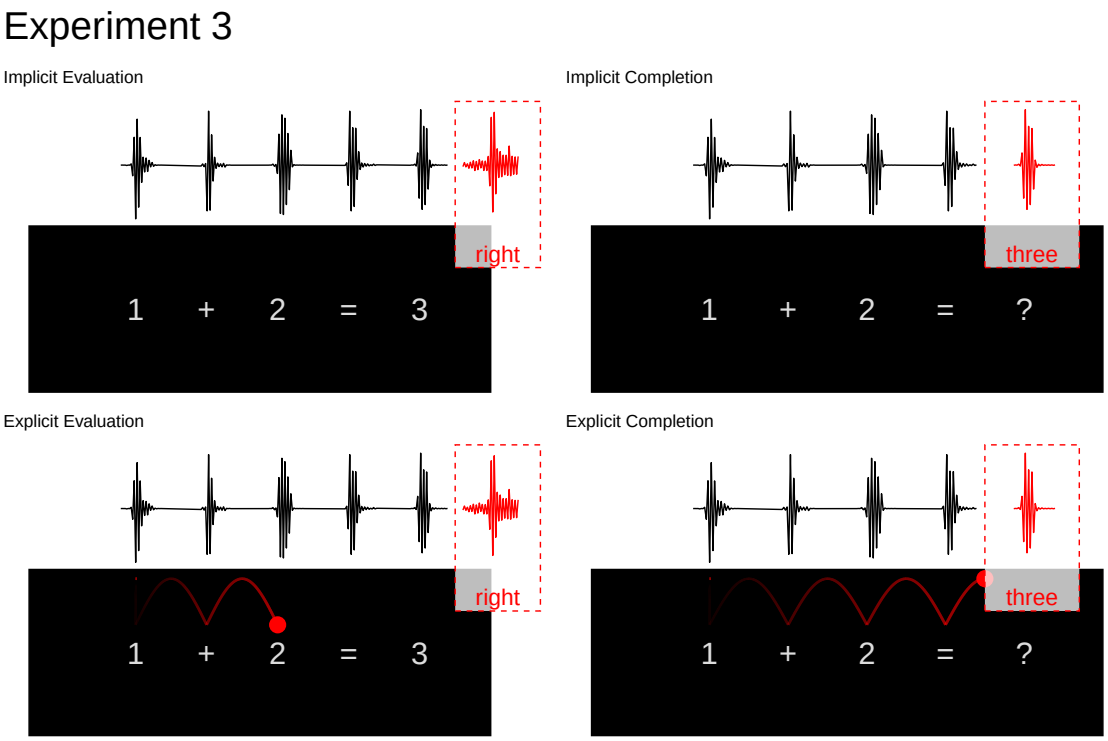

Figure 6.1: Schematic diagram of Experiment 3 paradigm. Upper panels show implicit rhythmic prime with no visual cue. Lower panels show explicit rhythmic prime with *karaoke* ball animation appearing to bounce over each number or operator of the sum in synchrony with perceptual word onsets. Tail illustrates movement and was not included in the experiment. Left panels show the evaluation condition where participants responded with RIGHT or WRONG to a complete sum. Right panels show the completion condition where participants were asked to produce the answer in an incomplete sum where a question mark acted as a placeholder in the visual presentation. Diagram not to scale.

BPM and 80 BPM) but used two pairs of tempos in a faster range as described in more detail in Section 6.3.1.

6.2.2 Implicit trials

The design employed in the previous experiments will be referred to as implicit evaluation where *implicit* refers to the absence of overt instructions or prompts to respond on the beat. As explicit instruction to respond on the beat would be expected to carry over from explicit trials to implicit trials, the implicit trial blocks were always presented first.

Implicit Evaluation The sum evaluation paradigm was used in evaluation trials with minor changes to the implementation. As illustrated in Figure 6.1, full sums were presented visually as in Experiment 1 as opposed to the rapid serial visual presentation (RSVP) design adopted in Experiment 2. This makes a strong distinction between the static presentation of the implicit condition and the dynamic presentation of the explicit condition. As with Experiment 2, audio stimuli were pre-prepared to simplify presentation.

Implicit Completion The sum completion paradigm was used in the completion trials. This used the same presentation method as the sum evaluation paradigm, with the only differences being that the auditory stimulus did not include the answer and the visual stimulus used a question mark in place of the answer.

6.2.3 Explicit trials

The explicit evaluation and explicit completion trials were the same as their implicit counterparts with the addition of a *karaoke dot* animation. A half-rectified sine function was used to give the appearance of a ball bouncing above each word in time with the p-centre estimate. When the ball reached the final position (answer or question mark) the animation continued in only the y-axis to bounce above the same value until the end of the trial. The timing of this animation was controlled using a variation of the AudioWorklet method used to control the RSVP presentation in Experiment 2. The position of the animation was continuously updated by messages from the same AudioWorklet as was used to record responses. The implementation only handles one message in each display refresh cycle to avoid any cumulative effect of processing delays.

6.2.4 Hypotheses

As there was no rhythm manipulation in the current experiment, only the predictions of the tempo entrainment hypothesis were tested. These were the same as for Experiment 2, except for a refinement of the beat alignment prediction to account for paired tempo groups with a fixed ratio. This is referred to as the phase ratio prediction.

Tempo	Tempo Group	
	Fast	Slow
Fast	140.0	105.0
Slow	122.5	91.9

Table 6.1: Two-by-two nested tempo variable structure. Rows show groups with 3:4 ratio between levels, and columns show the tempos within these groups. All tempos in BPM.

Tempo entrainment hypothesis

- Response time prediction: Response times will be faster when trials are presented at faster tempos
- Response dispersion prediction: Response times will be less dispersed when trials are presented at faster tempos
- Phase ratio prediction: Phase amplitude will be greater when calculated at trial tempo than at alternative tempo within tempo pairs

6.3 Methods

6.3.1 Design

This experiment was a two-by-two-by-four factorial design, with the independent variables prime (implicit vs. explicit), task (completion vs. evaluation), and tempo (four levels). Trials where the task was *completion* has an additional correctness variable with two levels (right vs. wrong). Due to the complexity of this design the two-by-two component of the design was partitioned out into four data sets for modelling. These were implicit-completion, implicit-evaluation, explicit-completion and explicit-evaluation.

The tempo factor had a nested two-by-two structure as shown in Table 6.1 where columns show the top level tempo groups and rows show tempos within groups. In each group there was a 3:4 ratio between the slow and fast tempos (rows of Table 6.1). The fast tempo of the slow group was selected as the mid-point between the tempos of the fast group to maximise the spread of tempos. In phase effect models, groups and sub-groups are modelled as separate variables.

6.3.2 Participants

Participants were recruited via the University of Glasgow School of Psychology subject pool and requests posted on social media. Participants received £6 for taking part in the study. As a requirement of taking part all participants self-reported as native English speakers with normal or corrected to normal hearing and vision. The data collection was stopped when 32 participants completed the experiment. Four participants were excluded from the implicit blocks due to responding with the wrong word (e.g., CORRECT rather than RIGHT) or not at all in more than 20% of trials. Five participants were excluded from the explicit blocks for the same reasons. Across both task the total number of participants included was 29. Ethical

approval was obtained from the University of Glasgow College of Science and Engineering Ethics Committee.

6.3.3 Materials

Auditory stimuli were created from the same files as for the previous experiments. The retiming of audio stimuli was refined to utilise the methods presented in Chapter 2. For each word, in each sum position, an average amplitude envelope was created from the set of tokens. All tokens were then retimed using the continuous method, to align with this averaged envelope. This increased the uniformity of tokens and simplified calculations used to generate full stimulus items.

6.3.4 Procedure

Participants were shown to the room by the experimenter and asked to sit at a computer. They were given a printed information sheet along with a brief verbal description of the task. After consent was given the experimenter left the room.

There were three main sections to the session as shown in Figure 6.2. The first of these, labelled validation trials, included the latency test trial and a stereo hearing task [1]. This was followed by the implicit and explicit sections that made up the main experiment.

For all participants the first half of the experiment was made up of implicit trials. Half of the participants were presented with completion trials in the first and fourth quarter of experiment and the other half of participants were presented with the evaluation trials in these positions. Controlling the sequence in this way maximised the number of trials between repetitions of items.

6.3.5 Preprocessing

Auditory response data quality was high relative to the remotely collected recordings. This allowed onsets to be detected with the same algorithm as used to detect stimulus onsets in Chapter 2. To confirm the accuracy of annotations, a click was added to the responses at the detected onset point. The experimenter listened to these modified responses and marked trials where the click did not occur during the expected response. The algorithm correctly identified the response in all trials where the participant responded with the expected word. Checks were also grouped by expected response to allow trials with the incorrect response to be removed. Trials where participants responded with the wrong word (e.g., CORRECT) were excluded. Where more than 20% of trials were excluded due to incorrect responses, the participant was excluded from the analysis of that task.

For two participants, responses were not captured due to a recording error. One participant was excluded from the explicit trials due to consistently responding in time with the answer of the sum in the evaluation task.

[1] This task is not reported here as it is currently under development

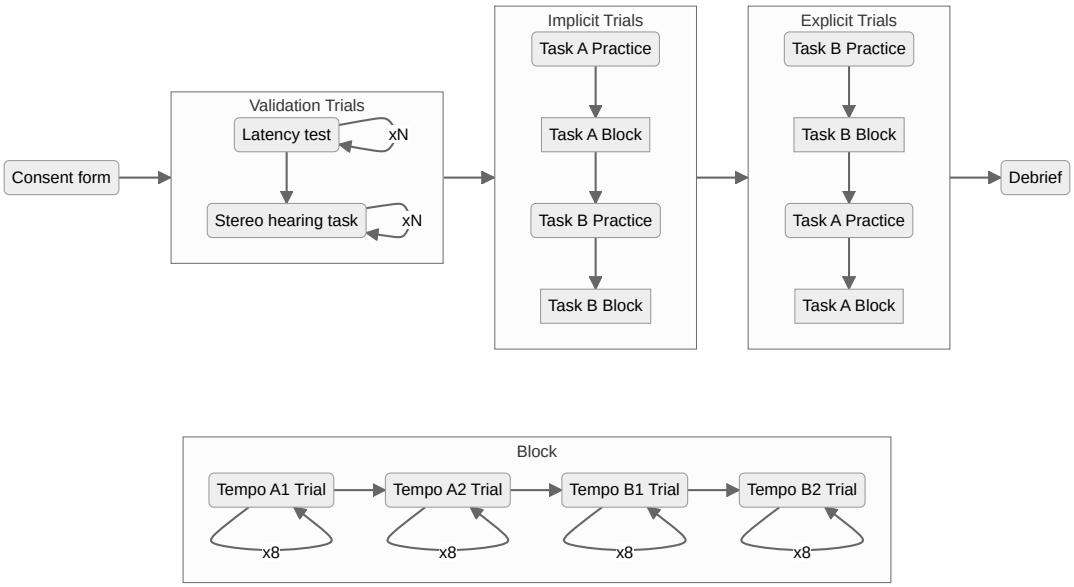

Figure 6.2: Flowchart of Experiment 3 procedure.

6.3.6 Model specification

Time-based The experiment produced four data sets that were analysed separately. These were implicit-completion, implicit-evaluation, explicit-completion and explicit-evaluation. The `brms` package (Bürkner, 2017) for R (R Core Team, 2023) was used to carry out analyses.

The main predictor in all models was the `period` between word onsets in the trial. For the evaluation task the effect of correctness was also modelled and is referred to in the model as `word`. The primary outcome for time-domain models was the time to response onset (p-centre), the central tendency parameter is referred to as `onset` in the model. The dispersion of the response time is also modelled and represented by the term `sigma` in the formula. To avoid ambiguity, the `onset` term from the model is referred to as `mu` in figures and tables to distinguish it from `sigma`. A maximal model random effect structure was used (Oberauer, 2022; Barr et al., 2013) for both the `onset` and `sigma` formulae. The log-normal family was used to model the distribution of responses.

Hypotheses were tested using the probability of direction (PD) and 95% critical interval (CI) obtained from the `hypothesis` function in the `brms` package. The PD is a two-tailed measure of the proportion of the distribution of posterior model estimates that are greater than or less then zero.

Frequency domain The phase amplitude measure was also modelled using the `brms` package. A multivariate response variable was used to model the phase amplitude calculated at both the fast and slow tempo for the group. The predictors were `rate`, `rate_group`, `word` and their interactions. The random effect structure included random intercepts and random slopes for `word` in the evaluation trials; and random intercepts in the completion trials. The beta distribution was used to model the bounded distribution of the phase amplitude measure.

6.4 Results

6.4.1 Time-domain results

In each of the four datasets it was expected that participants would respond faster to trials presented at faster tempos. Effects were tested using the probability of direction where a threshold of 97.5% was considered sufficient to accept the hypothesis. All effects of period reached this threshold. In the implicit evaluation condition the effect of period had a probability of $>$ 99.99% of being positive (Mean = 0.863, 95% CI [0.560, 1.167]). The effect of period in the implicit completion had a probability of $>$ 99.99% of being positive (Mean = 1.007, 95% CI [0.751, 1.257]). In the explicit evaluation condition the effect of period had a probability of $>$ 99.99% of being positive (Mean = 1.659, 95% CI [1.531, 1.792]). In the explicit completion condition the effect of period had a probability of $>$ 99.99% of being positive (Mean = 1.565, 95% CI [1.470, 1.657]).

It was also hypothesised that the variability of response times would be lower at faster tempos. All estimates were in the opposite direction of the hypothesised effect, but did not reach the 97.5% threshold. Results of these tests are shown in Table 6.2 in addition to tests of all independent variables. Distributions of model posteriors are shown in Figure 6.3.

6.4.2 Frequency-domain results

In each of the datasets it was expected that responses would be more likely to occur on the beat of the trial tempo. This was tested using the phase-amplitude measure described in Section 5.1.5. In the implicit evaluation condition the effect of rate (sub-group) on phase amplitude calculated with the slow reference tempo had a probability of 80.41% of being positive (Mean = 0.069, 95% CI [-0.087, 0.224]) and with the fast reference tempo had a probability of $>$ 99.99% of being negative (Mean = -0.318, 95% CI [-0.479, -0.159]). This effect is symmetric in the expected directions but only reaches the 97.5% threshold for the fast reference tempo. In the implicit completion condition the effect of rate (sub-group) on phase amplitude calculated with the slow reference tempo had a probability of $>$ 99.99% of being positive (Mean = 0.334, 95% CI [0.184, 0.483]) and with the fast reference tempo had a probability of $>$ 99.99% of being negative (Mean = -0.768, 95% CI [-0.930, -0.609]). This effect is symmetric in the expected directions but only reaches the 97.5% threshold for the fast reference tempo. In the explicit evaluation condition the effect of rate (sub-group) on phase amplitude calculated with the slow reference tempo had a probability of $>$ 99.99% of being positive (Mean = 1.173, 95% CI [1.026, 1.325]) and with the fast reference tempo had a probability of $>$ 99.99% of being negative (Mean = -1.696, 95% CI [-1.862, -1.534]). This effect is symmetric and reaches the 97.5% threshold in the expected directions. In the explicit completion condition the effect of rate (sub-group) on phase amplitude calculated with the slow reference tempo had a probability of $>$ 99.99% of being positive (Mean = 1.389, 95% CI [1.255, 1.526]) and with the fast reference tempo had a probability of $>$ 99.99% of being negative (Mean = -2.290, 95% CI [-2.453, -2.123]). This effect is symmetric and reaches the 97.5% threshold in the expected directions. All other terms are

			95% CI			
Term	Type	Estimate	Lower	Upper	PD	
implicit evaluation						
Intercept	mu	-0.85	-1.05	-0.65	> 99.9%	*
	sigma	-1.34	-1.89	-0.82	> 99.9%	*
period	mu	0.86	0.56	1.17	> 99.9%	*
	sigma	-0.22	-1.19	0.78	68.6%	
period:$word_{wrong}$	mu	0.09	-0.23	0.42	70.5%	
	sigma	-0.79	-2.10	0.50	88.0%	
$word_{wrong}$	mu	-0.04	-0.21	0.14	65.8%	
	sigma	0.58	-0.12	1.29	94.4%	
implicit completion						
Intercept	mu	-0.97	-1.15	-0.79	> 99.9%	*
	sigma	-1.42	-2.30	-0.52	99.9%	*
period	mu	1.01	0.75	1.26	> 99.9%	*
	sigma	-0.80	-2.31	0.71	85.5%	
explicit evaluation						
Intercept	mu	-1.44	-1.51	-1.35	> 99.9%	*
	sigma	-1.70	-2.37	-1.02	> 99.9%	*
period	mu	1.66	1.53	1.79	> 99.9%	*
	sigma	-0.86	-2.02	0.28	93.0%	
period:$word_{wrong}$	mu	-0.09	-0.21	0.03	92.7%	
	sigma	-0.99	-2.59	0.65	88.7%	
$word_{wrong}$	mu	0.06	-0.01	0.12	96.4%	.
	sigma	0.55	-0.34	1.44	89.6%	
explicit completion						
Intercept	mu	-1.42	-1.48	-1.36	> 99.9%	*
	sigma	-2.08	-2.79	-1.39	> 99.9%	*
period	mu	1.57	1.47	1.66	> 99.9%	*
	sigma	-0.68	-1.87	0.53	87.4%	

Table 6.2: Experiment 3 time-domain model results: Subscript in main effect terms refer to the level of the variable tested in the main effect (and in contrasts) where a positive estimate is in the direction of that level. Parameter indicates the parameter of the response variable the estimate refers to where `mu` is the central tendency (mean) and `sigma` the dispersion (SD). CI = Critical Interval; * = PD > 97.5%; . = PD > 95%.

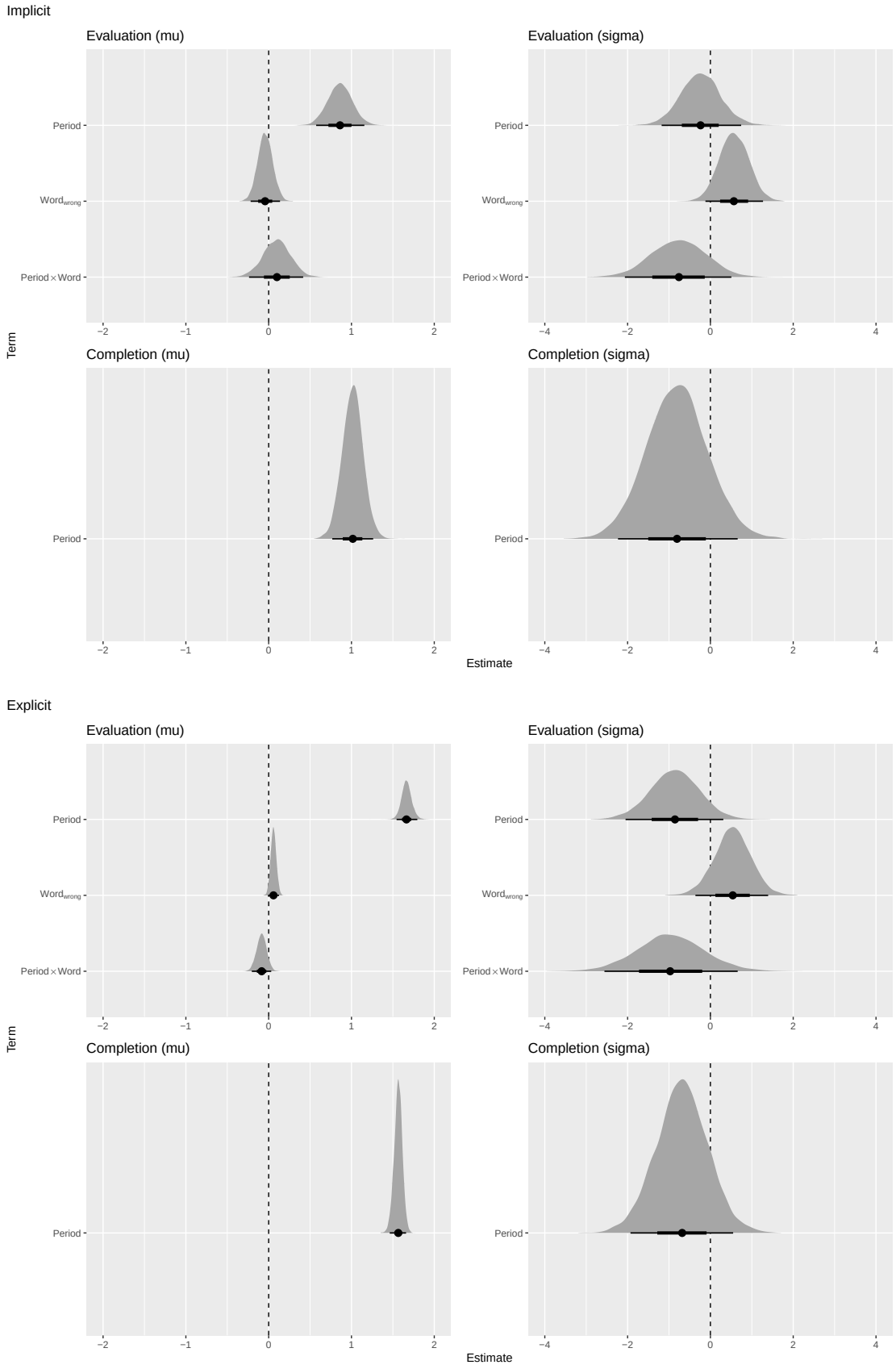

Figure 6.3: Posterior effects of time-domain models.

Term	Ref. Tempo	Estimate	95% CI Lower	95% CI Upper	PD	
Implicit Evaluation						
Intercept	fast	0.03	-0.12	0.18	68.9%	
	slow	0.20	-0.01	0.41	96.7%	.
Rate_{slow}	fast	-0.32	-0.48	-0.16	> 99.9%	*
	slow	0.07	-0.09	0.22	80.4%	
Rate $\times word$	fast	-0.17	-0.50	0.16	84.8%	
	slow	0.15	-0.16	0.46	83.1%	
Rate $\times RateGroup$	fast	0.24	-0.09	0.56	92.5%	
	slow	-0.04	-0.36	0.27	60.0%	
Rate $\times RateGroup \times word$	fast	-0.08	-0.72	0.58	59.1%	
	slow	0.51	-0.12	1.17	94.4%	
RateGroup_{slow}	fast	0.02	-0.14	0.18	57.1%	
	slow	0.10	-0.06	0.26	89.1%	
RateGroup $\times word$	fast	-0.18	-0.50	0.15	85.4%	
	slow	0.11	-0.20	0.43	76.1%	
word_{wrong}	fast	-0.03	-0.21	0.14	65.5%	
	slow	0.09	-0.09	0.26	84.8%	
Implicit Completion						
Intercept	fast	0.06	-0.11	0.22	76.4%	
	slow	0.65	0.45	0.84	> 99.9%	*
Rate_{slow}	fast	-0.77	-0.93	-0.61	> 99.9%	*
	slow	0.33	0.18	0.48	> 99.9%	*
Rate $\times RateGroup$	fast	-0.09	-0.41	0.22	71.3%	
	slow	0.15	-0.15	0.45	83.6%	
RateGroup_{slow}	fast	0.10	-0.06	0.25	88.9%	
	slow	0.03	-0.12	0.19	67.1%	

Table 6.3: Experiment 3 phase effect results for implicit trials. Subscript in main effect terms refer to the level of the variable tested in the main effect (and in contrasts) where a positive estimate is in the direction of that level. Parameter indicates the parameter of the response variable the estimate refers to where `mu` is the central tendency (mean) and `sigma` the dispersion (SD). CI = Critical Interval; * = PD > 97.5%; . = PD > 95%.

Term	Ref. Tempo	Estimate	95% CI Lower	95% CI Upper	PD	
Explicit Evaluation						
Intercept	fast	0.31	0.10	0.51	99.8%	*
	slow	1.03	0.84	1.22	> 99.9%	*
Rate_{slow}	fast	-1.70	-1.86	-1.53	> 99.9%	*
	slow	1.17	1.03	1.33	> 99.9%	*
Rate $\times word$	fast	0.12	-0.18	0.42	78.9%	
	slow	-0.18	-0.47	0.12	88.7%	
Rate $\times RateGroup$	fast	-0.31	-0.61	-0.02	98.1%	*
	slow	0.31	0.03	0.60	98.5%	*
Rate $\times RateGroup \times word$	fast	-0.13	-0.76	0.47	66.0%	
	slow	-0.31	-0.88	0.25	86.1%	
RateGroup_{slow}	fast	0.02	-0.13	0.17	60.2%	
	slow	-0.06	-0.20	0.08	81.1%	
RateGroup $\times word$	fast	0.06	-0.24	0.36	64.6%	
	slow	-0.04	-0.32	0.25	59.9%	
word_{wrong}	fast	-0.18	-0.35	-0.02	98.7%	*
	slow	0.08	-0.08	0.23	84.0%	
Explicit Completion						
Intercept	fast	0.31	0.07	0.54	99.4%	*
	slow	1.25	1.08	1.43	> 99.9%	*
Rate_{slow}	fast	-2.29	-2.45	-2.12	> 99.9%	*
	slow	1.39	1.25	1.53	> 99.9%	*
Rate $\times RateGroup$	fast	-0.34	-0.62	-0.07	99.1%	*
	slow	0.07	-0.17	0.32	71.9%	
RateGroup_{slow}	fast	-0.02	-0.16	0.11	60.7%	
	slow	-0.10	-0.23	0.02	94.5%	

Table 6.4: Experiment 3 phase effect results for explicit trials. Subscript in main effect terms refer to the level of the variable tested in the main effect (and in contrasts) where a positive estimate is in the direction of that level. Parameter indicates the parameter of the response variable the estimate refers to where `mu` is the central tendency (mean) and `sigma` the dispersion (SD). CI = Critical Interval; * = PD > 97.5%; . = PD > 95%.

reported in Table 6.3, Table 6.4, and Figure 6.4.

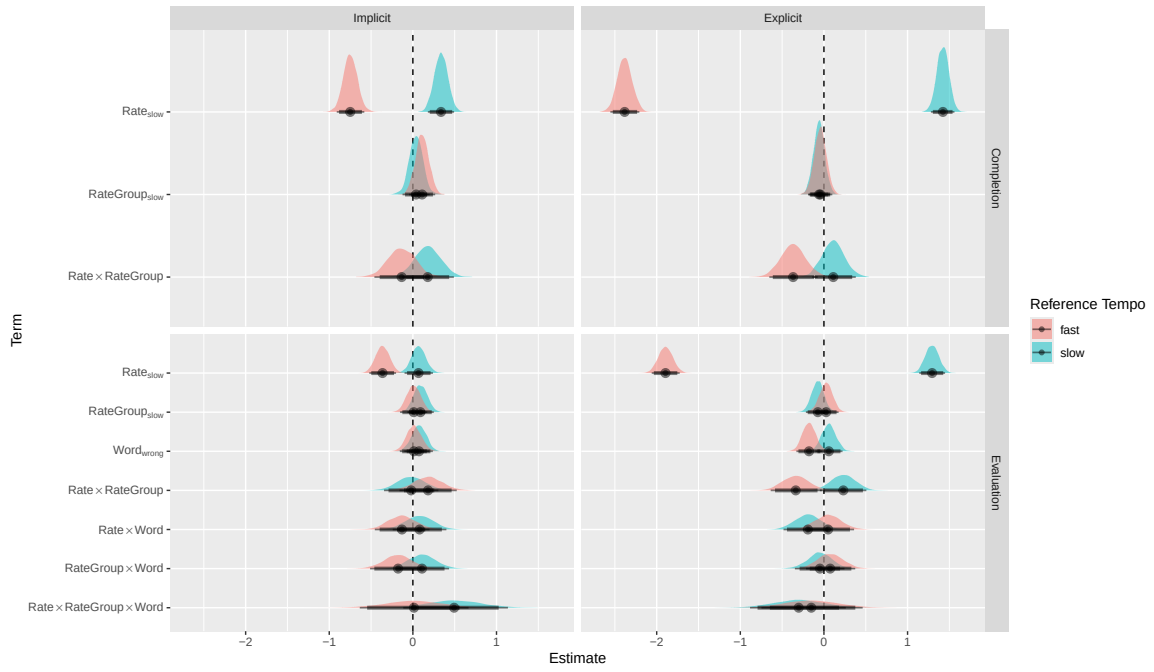

Figure 6.4: Posterior distributions of model estimates for phase amplitude models. Reference tempo refers to the tempo use to calculate the phase amplitude. True tempo effects would be expected to be symmetrical.

6.4.3 Exploratory analyses

As effects of time-domain models are expressed in log units, fixed-effect slopes were estimated from population-level model predictions. These estimates are shown in Table 6.5. These estimates can be interpreted as the unit increase in response onset per unit increase in stimulus period. In Experiment 1 and Experiment 2, where all trials were implicit evaluation, this estimate was approximately 12 ms / 100 ms.

Condition	Task	Word	Mode	CI (95%)
Explicit	Evaluation	right	0.99	[0.89, 1.12]
Explicit	Evaluation	wrong	0.94	[0.85, 1.06]
Explicit	Completion	-	0.88	[0.83, 0.95]
Implicit	Evaluation	right	0.58	[0.34, 0.82]
Implicit	Evaluation	wrong	0.62	[0.36, 0.87]
Implicit	Completion	-	0.64	[0.48, 0.81]

Table 6.5: Modes and 95% critical intervals (CIs) of model fixed-effect slopes estimates expressed in seconds.

In order to visualise observed and modelled responses, they were combined in Figure 6.5. The background of each panel shows the phase of the trial tempo, where black would be in-phase and white would be anti-phase. In the explicit condition, observed response onsets (shown in red) fall predominantly in the black band. The modelled distributions shown in highest density continuous interval (HDCI) slabs appear to capture these observations with a slight lag. In the implicit conditions a similar pattern can be seen, but with greater dispersion of distributions.

Higher densities at the modes of the implicit completion condition suggest a stronger effect in completion trials compared to evaluation trials as also shown in Table 6.2 and Table 6.5.

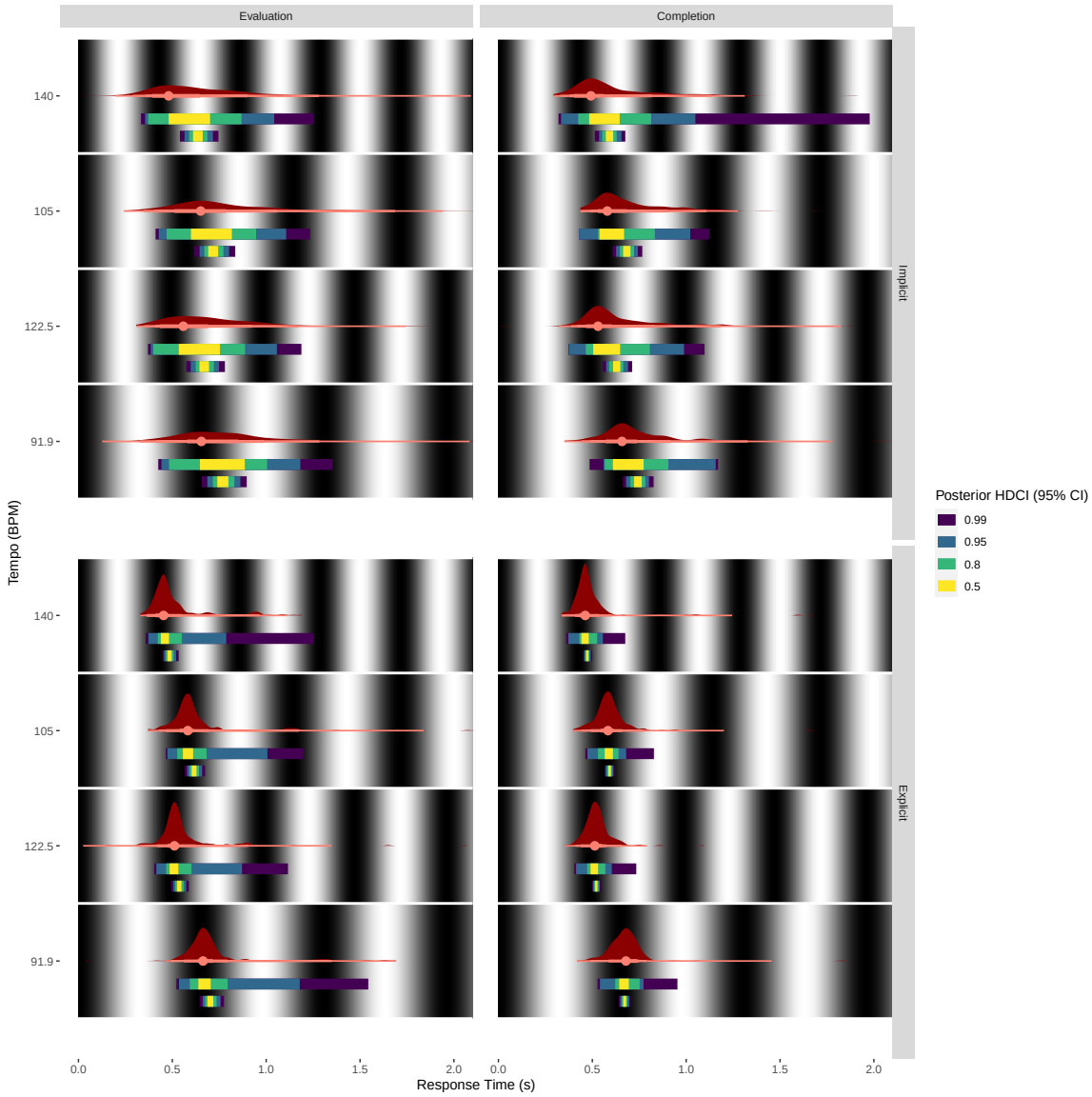

Figure 6.5: Experiment 3 Posterior model fits. Columns show the two task conditions (completion and evaluation). Rows show the two presentation conditions (implicit and explicit). Bands on plot backgrounds show phase of corresponding tempo where black areas would correspond to in-phase responses and white to anti-phase responses. Each sub-plot, from top to bottom, shows the distribution of observed responses (red); the HDCI of predictions from model posterior with random effects, the HDCI of predictions from model posterior without random effects. HDCI = Highest Density Continuous Interval. Point in distribution of observations shows mode.

An alternative representation of the phase of response is to express response onsets as phase vector diagrams. Figure 6.6 shows individual responses as segments extending out from what would be a vector where both the magnitude (modulus) and angle (argument) are derived from the response onset time. The black vector arrow is derived from a variation of this method where the modulus is given a fixed value of 1 and the mean of the complex valued vectors is calculated. This is equivalent to connecting each vector end-to-end and dividing the vector projected from the origin to the final vector by the number of observations. For convenience this

was then multiplied by 5 to match the scale of the plot. For comparison, a red vector arrow is shown which indicated the same measure for the response onsets in the alternative trial tempo. Interpretation of these plot would be similar to interpretation of coherence measures such that a horizontal vector with a magnitude of 5 would indicate all responses being perfectly in phase. Where phase is random the vector would be expected to tend toward the origin of the plot. Alternatively the x coordinate of the vector would be equivalent to the "imaginary" component of coherence (see Bastos & Schoffelen, 2016, for further detail). Notably, in all panels of the plot the x coordinate of the black vector is greater than the x coordinate of the red vector.

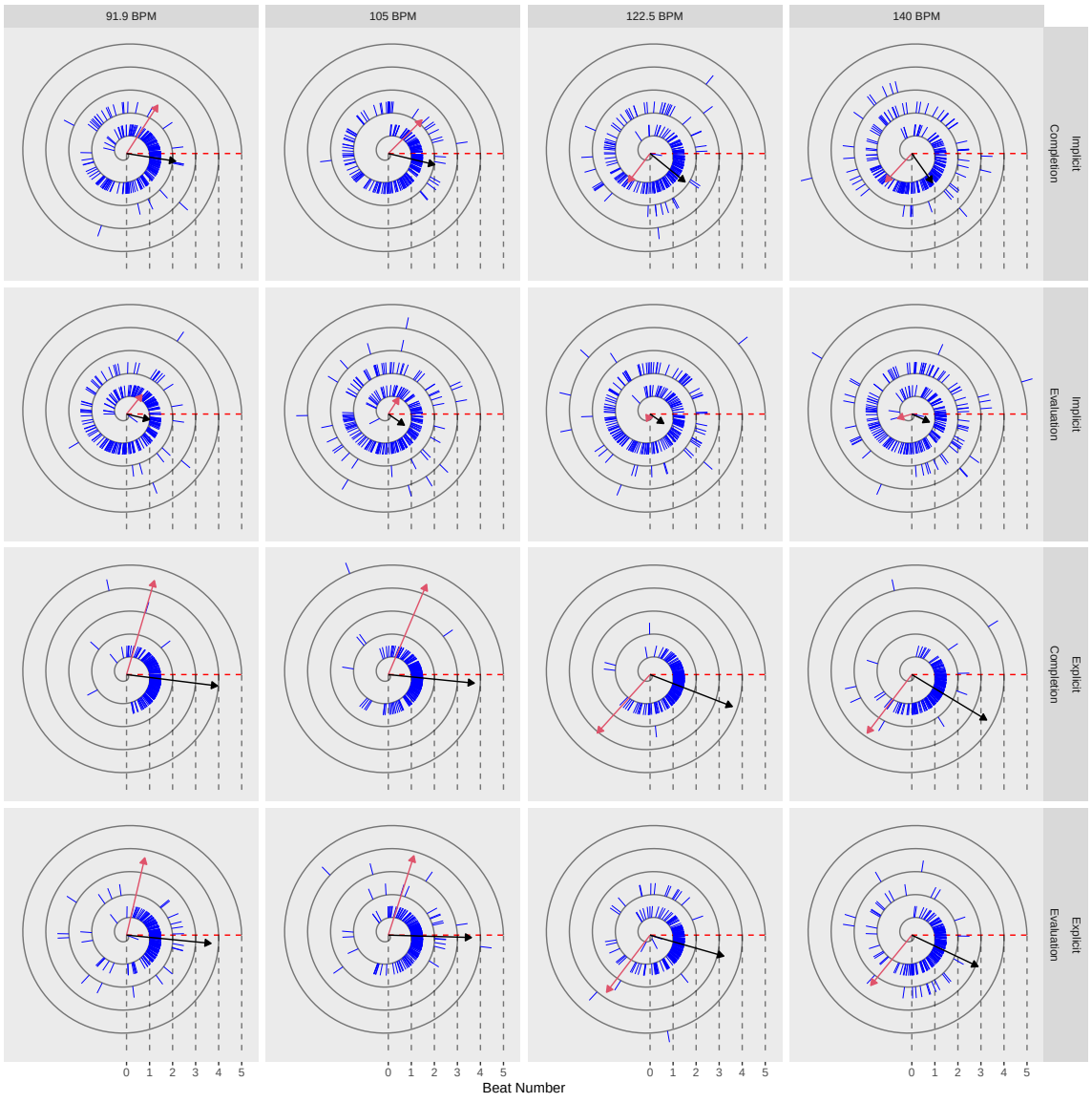

Figure 6.6: Response times expressed as phase of response at trial tempo. Blue segments show individual participant responses. Black arrow shows the average response vector (multiplied by 5 to match the limits of the plot). Red arrow shows the average response vector for trials at the alternative tempo calculated using the same reference period (e.g., the red vectors in the 105 BPM column are responses from 140 BPM trials with phase calculated at 105 BPM). Dashed lines projected from x-axis indicate the beat associated with an in-phase response for each cycle.

6.5 Discussion

Following the previous two experiments, there were remaining concerns about the interpretation of findings. While the time-domain effect appeared modest yet robust, the frequency-domain evidence of Experiment 2 required confirmation due to design limitations. This experiment aimed to address those limitation to test the stronger prediction of an effect of the phase of the trial tempo. Two modifications of the paradigm were also introduced, in the explicit tasks, to determine a measure of optimal performance, and in the completion tasks, to determine if a strong effect would require the response to be within speaker-turn. This discussion first addresses findings of these new tasks before considering how the lab-based data collection may have impacted on findings.

6.5.1 Explicit vs. implicit

The explicit tasks were successful in establishing a reference for optimal task performance. This is evident across time-domain, frequency-domain and exploratory analyses. Generally participants were able to respond close to the beat. In the explicit conditions, the mode of the fixed effect approaches a 1:1 relationship with the trial period, with the 95% critical interval including 1 in the explicit evaluation trials. However, at faster tempos responses lag behind the beat slightly. This is most apparent in the phase-vectors in Figure 6.6 where the mean phase moves clockwise from the horizontal in the faster tempos.

Notably, the distinctive distributions of explicit trials shown in Figure 6.5 highlight qualitative differences in the responses. The fastest tempo shows a right-skewed distribution, suggesting that participants found it more challenging to respond on the beat of the fastest trials. Conversely, in the slowest tempo a slight left-skew [2] can be seen suggesting that participants may have found this tempo too slow.

Compared to the near-optimal performance in the explicit tasks, what we see in the implicit conditions is more dispersed responses, but with modes showing a tendency to respond close to the first beat following the final stimulus onset.

6.5.2 Completion vs. evaluation

The finding that responses were less consistent in the evaluation task than the completion task for the explicit trials would suggest that responding on the beat across speaker turns is less intuitive than completing an utterance in phase with the speaker. This can also be seen in the implicit condition where the magnitude of the vector is smaller for evaluation than completion and there is a higher density of participant level vectors near the centre of the plot. This observation is not apparent from the time-domain model estimates shown in Table 6.2. However, the phase amplitude model failed to meet the threshold in the implicit evaluation condition. Furthermore, the magnitude of estimates in the explicit condition was nominally

[2] This left-skew is visible around the centre of the distribution but an additional right-skew is visible in the tail.

greater than those for the implicit condition in both the implicit and explicit trials. While the current experiment was not designed to contrast these conditions, taken together these findings suggest that a frequency-domain effect is less likely to be observed across speaker turns such as in the evaluation trials.

6.5.3 Lab-based

Effect sized in the current lab-based study were larger than observed in the previous two web-based studies (∼60 ms compared to ∼12 ms). This may be explained in part by the increased precision of the physical loopback increasing power. As discussed in Chapter 3, the use of a loopback allowed accurate response time measures with millisecond precision. However, Chapter 3 also demonstrated excellent precision of RTL measurements. Other differences between the web-based and lab-based experiment may have had more of an impact. In-person data collection allowed for control of the participant's environment and provided a distraction-free room. In the prior web-based data collection, participants may have been distracted by uncontrolled factors such as interruptions from people entering the room; notifications or alerts from phones and other devices; or background noises.

Chapter 7

General Discussion

7.1 Research goals and key findings

In Chapter 1, four questions were asked concerning the effects of rhythm in speech. Firstly, asking if speech is intrinsically rhythmic. Secondly, whether listeners entrain to speech rhythm. Thirdly, whether entrainment affects production. Finally, how the transfer of perceptual entrainment to production could be detected. Regarding the first two questions, a review of the literature suggested that while there is sufficient evidence that rhythm in speech can be perceived and tracked at a cortical level, it is less clear how speech rhythm might be functionally defined. Literature related to the third question found active interpretations of neural entrainment where predictability in the temporal structure of speech may facilitate communication as collaborative joint-action.

These research questions led to two complementary hypotheses: The tempo entrainment hypothesis and rhythm entrainment hypothesis. The tempo entrainment hypothesis predicted that the tempo of heard speech affects the timing of spoken responses. The rhythm entrainment hypothesis predicted that the regularity of heard speech affects the regularity of spoken responses. Together these hypotheses predict that entrainment and the strength of entrainment in perception, will affect the time to respond and the phase of response. To address these predictions, a new *maths sums* paradigm was introduced in which participants responded verbally to spoken maths sums presented with temporally defined rhythms.

Because this paradigm required rhythm to be manipulated in a controlled manner, it was necessary to consider the potential impact of different speech retiming methods. In Chapter 2, the terms *serialisation*, *point-to-point*, and *continuous* were defined to help researchers describe key differences in the way retimed speech stimuli are produced. Analyses of the degree of temporal distortion each method provided reassuring findings, suggesting that, in the grammatically predictable case of maths sums, all three methods could be used to make stimuli more temporally predictable. The serialisation method best describes the retiming used in this thesis, however, continuous retiming was used at word level in the final experiment. In this case the continuous retiming primarily served to minimise the retiming factor calculated by the

`get_serial_anchors()` discussed in Section 2.2.2.

As a result of the COVID-19 pandemic, the associated restrictions and uncertainty around in-person research, it was necessary to perform data collection remotely. Chapter 3 reviews the particular challenges of remote web-based data collection in speech research and presents methods of correcting systematic errors and maximising precision. Crucially, these methods allowed more sophisticated analyses of responses in Experiment 2 and Experiment 3, addressing the effect of stimulus timing on the phase of response.

The first two experiments, which were conducted remotely, provided support for the tempo entrainment hypothesis. Experiment 1, which preceded the round-trip latency (RTL) method, provided evidence in the time domain, showing that participants responded faster to faster tempos. Experiment 2, benefiting from more accurate response time measurement, found a similar time-domain effect and additionally provided evidence of frequency-domain effect. This frequency-domain effect was detected using a method that combined principles of signal detection with techniques used in coherence measures.

In addition to measurement improvements, Experiment 2 also differed from Experiment 1 in terms of design. Experiment 2 presented tempo and rhythm manipulations in a counterbalanced block design, with 2 tempos and 1 (isochronous) rhythm in the first two blocks and 1 tempo and 2 rhythms in the final two blocks. This design may have been successful in achieving its aim of providing greater exposure to conditions. However, an unintended consequence of this design was that it confounded experiment block sequence with manipulations within subjects.

Experiment 3 was conducted in person in a lab after COVID-19 restrictions had eased. This opportunity was taken to reassess the approach taken in the first two experiments and address uncertainties. Concerns about the block design were addressed by adding additional levels of the tempo manipulation in a paired configuration that allowed the phase relationship between tempos, within pairs, to be distinguished from the simple time-domain relationship (i.e., faster or slower) between tempos. In addition to these differences, two new paradigm variations were introduced. The first paradigm variation added an explicit prompt to respond "on the beat" along with an animation providing an additional visual cue to establish a measure of how well participants were able to align responses to the beat. This task addressed the interpretation of the previous results that effect sizes were not consistent with a phase-locked effect, where the magnitude of the tempo manipulation would be expected to correspond with the magnitude of the response time effect when expressed in the same units. By explicitly asking participants to respond on the beat, it was possible to test whether such a deterministic prediction could be modelled, and facilitate comparison of effect sizes in the implicitly primed responses from prior experiments. The second new variation of the paradigm asked participants to complete sums, rather than evaluate them. This variation more closely aligns with the proposal of speech as joint-action.

As a result of the additional paradigm variations and the finite nature of sums meeting the stimulus requirements, the rhythm entrainment hypothesis was not tested in Experiment 3.

The explicit trials demonstrated that participants were able to align responses to the beat, but also showed participants lagging slightly behind the beat at faster tempos in both the completion and the evaluation task. In the time-domain analysis, when expressed in the same units as the tempo manipulation (i.e., seconds), the modes of fixed-effect slope estimates ranged between 0.88 and 0.99. In the implicit tasks, the tempo entrainment hypothesis was supported in both time-domain and frequency-domain analysis. Responses in the four paradigm variations were analysed in separate models, meaning hypothesis testing of the contrasts between paradigms was not possible. However, exploratory analyses were performed by visualising both time-domain and frequency-domain distributions of responses. In implicit responses, these visualisations demonstrated more concentrated modes near the beat in the time domain and stronger phase alignment in the frequency domain.

Chapter 3 also reported a validation, using latency estimate responses to assess precision in the remote data collected in Experiment 2, and accuracy in the lab-based data collected in Experiment 3. This validation successfully demonstrated that the method achieves the aim of allowing accurate measurement, and thus correcting, of systematic latency errors. Moreover, the precision achieved in remote data collection was impressive, even by the standards of lab-based evaluations of alternative web-based software.

The empirical findings of the three experiment chapters supported the tempo entrainment hypothesis. Participants were not only more likely to respond faster to faster tempos, but were also more likely to respond closer to the beat of the trial than a competitor. Tests of the rhythm entrainment hypothesis were supported in both Experiment 1 and Experiment 2.

The explicit conditions appended to the end of Experiment 3 provide a valuable reference point for optimal task performance. They confirm that participants are able to respond on the beat when explicitly asked to do so. With the hindsight of having conducted all three experiments, this baseline measure now seems essential to the interpret ability of results. Even with the explicit instruction to respond on the beat, and an additional visual stimulus to guide responses, participants struggled to respond on the beat. In the fast tempos of both rate groups, the mode of responses and the average response vector lagged behind the beat.

7.2 Comparison with other studies

The closest experimental paradigm to the maths sum paradigm is the yes/no question paradigm used by Corps et al. (2020). Specifically, this design resembles the sum evaluation paradigm in that participants are asked a question with a binary response, such as "Do you have a pet horse?". However, findings of the experiments reported here are not simply a replication. There are many important differences between the maths sum paradigm and yes/no question paradigm, but crucially, the maths sum paradigm measured phase of response. This was possible due to the suitability of maths sums for isochronous retiming. Similarly, previous findings in studies where participants produce continuous speech responses (Jungers & Hupp, 2009; Wynn & Borrie, 2020) are consistent with the findings reported in this thesis but would also be

consistent with simpler time-domain interpretations of effects and did not allow frequency-domain interpretations to be tested.

The rhythm manipulations in Experiment 1 and Experiment 2 were similar to those employed by Quené and Port (2005) in a phoneme detection task where words were, to use the terminology suggested in Chapter 2, the auditory presentation was serialised. The inconclusive findings in Experiment 1 and the cautiously interpreted support in Experiment 2 suggest that further investigation would be required to confirm that the two paradigms detect the effects of the same underlying mechanisms. The less ambiguous tempo effects found in all three experiments of this thesis support an oscillatory entrainment interpretation of the findings of Quené and Port (2005), that participants were faster to respond in isochronous trials due to periodic modulation of attention. Another consideration would be that the differences between anisochronous speech, as employed in this thesis, and the jittered speech employed by Quené and Port (2005) may affect responses. In jittered speech, the mean interval between stimulus onsets is free to vary, which in turn allows trial tempos to vary. This would not be acceptable in experiments, such as Experiment 1, where both rhythm and tempo are manipulated within blocks.

This thesis supports the findings of synchronous speech paradigms, that participants can both implicitly (Assaneo et al., 2019) and explicitly (Cummins, 2009) demonstrate entrainment to heard speech in the speech they produce. By demonstrating this in a paradigm where responses are not concurrent with the stimulus, the current findings are consistent with a sustained entrainment effect after stimulus offset as found by Kösem et al. (2018).

In a previous study, isochronous retiming was found to improve intelligibility of speech in noise compared to anisochronous speech, but not compared to naturally produced speech (Aubanel et al., 2016). Due to the atypically predictable nature of maths sums, the difference between isochronous retimings and original recordings were minor. As demonstrated in the analysis presented in Chapter 2, simply normalising the mean tempo of sums resulted in clear peaks in the FFT and ACF analysis. The *rhythmic chimera* method demonstrated in the same chapter would allow for further investigation of the more complex temporal aspects of rhythm in continuous speech. For example, creating rhythmic chimeras that cross the timing of speech produced in a live synchronous speech task with speech produced in a solo reading task, could help to identify the features of synchronously produced speech that facilitate synchronisation between a live speaker and a recording (Cummins, 2009).

7.3 Evaluation of the maths sum paradigm

The apparent simplicity of the maths sum paradigm was revealed to be deceptive over the course of this thesis. Over three experiments, tempo was manipulated as a pseudo-continuous variable, a target/competitor pair, and as nested groups of interleaved target/competitor pairs; rhythm was manipulated within block, between blocks, and not at all; auditory stimuli were presented along with static text, individually as words in RSVP, and accompanied by a karaoke inspired animation; participants were asked to evaluate the sum in one task, and to complete the sum

in another; the response was primed implicitly, and explicitly; data collection was conducted remotely, and in a lab. Had the findings been less consistent, this would be an extensive parameter space to explore. However, findings and the differences between experiments were consistent with expectations. For example, stronger effects in the final experiment reflected refinements of the design and the greater control over the participant's environment afforded by the lab setting.

Important strengths of the maths sum paradigm are that both stimulus and response are predictable, objectively defined, and (with the exception of MINUS) monosyllabic. These strengths lead to the limitation that the possible set of stimuli meeting these requirements are finite. Fortunately, this would not be true if the requirement that numbers are monosyllabic were to be relaxed, or that only the + and − operators are used. However, this would make the cognitive demands of the task greater.

These constraints of the maths sum stimuli are also indicative of a limitation of ecological validity. In studying an atypically rhythmic speech form, the paradigm does not address the need for more naturalistic evidence of entrainment (Alexandrou et al., 2020).

Alternative stimulus types with predictable responses could also be explored. For example, *knock-knock* jokes would allow dynamic interaction between participants while maintaining predictability for at least the second speaker in each trial. This, would also address the limitation that isochrony is a highly constrained definition of speech rhythm and allow other forms of prosodic predictability to be explored, such as intonational patterns in the example of knock-knock jokes (Day-O'Connell, 2010).

Across all three experiments the same source recordings of a female speaker of Standard Scottish English were used. This thesis does not consider how potential sociolinguistic aspects of communication could affect responses. Exploration of effects of different speakers or exploring non-temporal speech manipulations (e.g., fundamental frequency; f0), in combination with a review of known effects of these features, may open the paradigm up to further applications. A specific prediction that could be drawn from the literature reviewed here is that responses to a live speaker would be expected to differ from responses to a recording (Cummins, 2009).

7.4 Is it *entrainment proper*?

Meyer et al. (2020) proposed a distinction between extrinsic and intrinsic forms of entrainment, where the latter allows for a more flexible interpretation of the former. This thesis did not measure neural entrainment directly, but rather explored expectations of the type of stimulus driven entrainment that Meyer et al. refer to as *entrainment proper*. In doing so, this thesis asks whether frequency-domain terminology such as neural *oscillators*, quasi-*periodic*, or quasi-*isochronous* used in neuroscience literature express necessary conditions of behavioural forms of entrainment.

The findings of Experiment 2, and the more robust findings of Experiment 3 suggest that

frequency-domain influences do affect the timing of spoken responses. However, this finding could be scrutinised further in future research. An example is the finding mentioned in Chapter 1, that retiming neural tracking of jittered speech can allow detection of tracking using frequency-domain methods (Jin et al., 2018). In the context of the findings of this thesis, we may ask if an effect was detected in the frequency domain because of periodic stimulus tracking in the frequency domain, or because of time-domain tracking of a periodic stimulus.

As isochronous speech is not typical of natural speech, even periodic oscillatory theories must account for variation through mechanisms such as flexible oscillators (Ghitza, 2013) or phase-reset (Giraud & Poeppel, 2012). As a result, strictly periodic interpretations of neural entrainment theories are not necessary.

In the context studied here, neural entrainment is a potential underlying mechanism supporting speech as joint-action. The findings presented in this thesis are consistent with the theory that speaker and listener engage both perception and production processes (Pickering & Garrod, 2013) to predict the content and also timing (Garrod & Pickering, 2015) of each other's speech.

7.5 Conclusion

This thesis found consistent effects of the tempo of heard speech on the timing of spoken responses. Significantly, effects were detected in the frequency domain, contributing strong evidence for an oscillatory mechanism bridging perception and production in speech. This evidence is consistent with the characterisation of speech as joint-action, where the speaker's production and listener's perception share a common temporal underpinning.

The methodological contributions, developed to enable careful stimulus creation and accurate response measures, provide tools to address the wider research questions addressed here. To say that speech *is* rhythmic based on the findings presented here would fail to recognise the richness of human speech. However, this thesis demonstrates that speech *can* be rhythmic, and that entrainment to this rhythm can transfer to production.

Bibliography

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh University Press.

Aldarmaki, H., Ullah, A., Ram, S., & Zaki, N. (2022). Unsupervised Automatic Speech Recognition: A review. *Speech Communication*, *139*, 76–91. https://doi.org/10.1016/j.specom.2022.02.005

Alexandrou, A. M., Saarinen, T., Kujala, J., & Salmelin, R. (2020). Cortical entrainment: What we can learn from studying naturalistic speech perception. *Language, Cognition and Neuroscience*, *35*(6), 681–693. https://doi.org/10.1080/23273798.2018.1518534

Anglada-Tort, M., Harrison, P. M. C., & Jacoby, N. (2021). *REPP: A robust cross-platform solution for online sensorimotor synchronization experiments* (preprint). Neuroscience. https://doi.org/10.1101/2021.01.15.426897

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, *53*(4), 1407–1425. https://doi.org/10.3758/s13428-020-01501-5

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398. https://doi.org/10.1016/j.tics.2012.05.003

Arvaniti, A. (2009). Rhythm, Timing and the Timing of Rhythm. *Phonetica*, *66*(1-2), 46–63. https://doi.org/10.1159/000208930

Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., & Poeppel, D. (2019). Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nature Neuroscience*, *22*(4), 627–632. https://doi.org/10.1038/s41593-019-0353-z

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*. https://doi.org/10.3389/fnhum.2016.00430

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barreda, S. (2015). *phonTools: Functions for phonetics in r*. manual.
R package version 0.2-2.1.

Bastos, A. M., & Schoffelen, J.-M. (2016). A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls. *Frontiers in Systems Neuroscience*, *9*. https://doi.org/10.3389/fnsys.2015.00175

Beith, A. (2023). *RetimeR: Tools for retiming speech in R* (Version 0.1.2).

Beith, A., Barr, D. J., & Smith, R. (2024). Preserving prosody in temporal distortions of speech. https://doi.org/10.17605/OSF.IO/9VC7J

Beith, A., Barr, D. J., & Smith, R. (in press). Preserving prosody in temporal distortions of speech. In L. Meyer & A. Strauss (Eds.), *Rhythms of Speech and Language: Culture, Cognition, and the Brain*. Cambridge University Press.

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer* (Version 6.3.10). http://www.praat.org/

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-Reading Enables the Brain to Synthesize Auditory Features of Unknown Silent Speech. *The Journal of Neuroscience*, *40*(5), 1053–1065. https://doi.org/10.1523/JNEUROSCI.1101-19.2019

Breska, A., & Deouell, L. Y. (2016). When Synchronizing to Rhythms Is Not a Good Thing: Modulations of Preparatory and Post-Target Neural Activity When Shifting Attention Away from On-Beat Times of a Distracting Rhythm. *Journal of Neuroscience*, *36*(27), 7154–7166. https://doi.org/10.1523/JNEUROSCI.4619-15.2016

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Brook, C. (2024, June 25). *Corbanbrook/dsp.js*. Retrieved June 26, 2024, from https://github.com/corbanbrook/dsp.js

Bürki, A., Ernestus, M., Gendrot, C., Fougeron, C., & Frauenfelder, U. H. (2011). What affects the presence versus absence of schwa and its duration: A corpus analysis of French connected speech. *The Journal of the Acoustical Society of America*, *130*(6), 3980–3991. https://doi.org/10.1121/1.3658386

Bürkner, P.-C. (2017). **Brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01
R package version 2.17.0.

Byrd, D., & Krivokapić, J. (2021). Cracking Prosody in Articulatory Phonology. *Annual Review of Linguistics*, *7*(1), 31–53. https://doi.org/10.1146/annurev-linguistics-030920-050033

Cason, N., Astésano, C., & Schön, D. (2015). Bridging music and speech rhythm: Rhythmic priming and audio–motor training affect speech perception. *Acta Psychologica*, *155*, 43–50. https://doi.org/10.1016/j.actpsy.2014.12.002

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2020). *Shiny: Web application framework for r*. manual. https://CRAN.R-project.org/package=shiny
R package version 1.7.1.

Continuous-wave radar. (2024, June 9). In *Wikipedia*. Retrieved June 26, 2024, from https://en.wikipedia.org/w/index.php?title=Continuous-wave_radar&oldid=1228134314
Page Version ID: 1228134314.

Corps, R. E., Gambi, C., & Pickering, M. J. (2020). How do listeners time response articulation when answering questions? The role of speech rate. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(4), 781–802. https://doi.org/10.1037/xlm0000759

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, *37*(1), 16–28. https://doi.org/10.1016/j.wocn.2008.08.003

Cummins, F. (2012a). Looking for Rhythm in Speech. *Empirical Musicology Review*, *7*(1-2), 28–35. https://doi.org/10.18061/1811/52976

Cummins, F. (2012b). Oscillators and Syllables: A Cautionary Note. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00364

Cummins, F., Grimaldi, M., Leonard, T., & Simko, J. (2006). The chains corpus: Characterizing individual speakers. *Proc of SPECOM*, *6*(2006), 431–435.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and speech*, *40*(2), 141–201. https://doi.org/10.1177/002383099704000203

Day-O'Connell, J. (2010). "Minor third, who?": The intonation of the knock-knock joke. *Speech Prosody 2010-Fifth International Conference*. https://doi.org/10.21437/SpeechProsody.2010-218

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dilley, L. C., & Pitt, M. A. (2010). Altering Context Speech Rate Can Cause Words to Appear or Disappear. *Psychological Science*, *21*(11), 1664–1670. https://doi.org/10.1177/0956797610384743

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

Driedger, J., & Müller, M. (2014). TSM Toolbox: MATLAB Implementations of Time-Scale Modification Algorithms. *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 249–256.

Fairs, A., & Strijkers, K. (2021). Can we use the internet to study speech production? Yes we can! Evidence contrasting online versus laboratory naming latencies and errors (S. Sulpizio, Ed.). *PLOS ONE*, *16*(10), e0258908. https://doi.org/10.1371/journal.pone.0258908

Gallois, C., Ogay, T., & Giles, H. (2005). Communication Accommodation Theory. In W. B. Gudykunst (Ed.), *Theorizing about intercultural communication* (pp. 121–148). Sage.

Garrod, S., & Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00751

Ghitza, O. (2013). The theta-syllable: A unit of speech information defined by cortical function. *Frontiers in psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00138

Gilbert, B. (2020). *Jspsych-audio-audio-response.js*. Retrieved December 15, 2023, from https://github.com/becky-gilbert/jsPsych/blob/audio-response/plugins/jspsych-audio-audio-response.js

Branch: audio-response

Commit 02ab3f3.

Giorgino, T. (2009). Computing and Visualizing Dynamic Time Warping Alignments in *R* : The **dtw** Package. *Journal of Statistical Software*, *31*(7). https://doi.org/10.18637/jss.v031.i07

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Goswami, U., & Leong, V. (2013). Speech rhythm and temporal structure: Converging perspectives? *Laboratory Phonology*, *4*(1). https://doi.org/10.1515/lp-2013-0004

Grillo, N., Santi, A., & Turco, G. (in press). Duration and the prosodic disambiguation of nested structure. In L. Meyer & A. Strauss (Eds.), *Rhythms of Speech and Language: Culture, Cognition, and the Brain*. Cambridge University Press.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P. G., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLOS Biology*. https://doi.org/10.1371/journal.pbio.1001752

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2013). *Halliday's Introduction to Functional Grammar* (4th ed.). Routledge. Retrieved November 7, 2018, from http://ebookcentral.proquest.com/lib/gla/detail.action?docID=1323310

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, *88*, 70–86. https://doi.org/10.1016/j.jml.2016.01.001

Henke, L., Roll, M., Beier, E., & Meyer, L. (in press). Cognitive and neural constraints on timing and rhythm in language. In L. Meyer & A. Strauss (Eds.), *Rhythms of Speech and Language: Culture, Cognition, and the Brain*. Cambridge University Press.

Henry, M. J., & Herrmann, B. (2014). Low-Frequency Neural Oscillations Support Dynamic Attending in Temporal Context. *Timing & Time Perception*, *2*(1), 62–86. https://doi.org/10.1163/22134468-00002011

Hyman, J. M. (1983). Accurate Monotonicity Preserving Cubic Interpolation. *SIAM Journal on Scientific and Statistical Computing*, *4*(4), 645–654. https://doi.org/10.1137/0904045

Jacoby, N., & McDermott, J. H. (2017). Integer Ratio Priors on Musical Rhythm Revealed Cross-culturally by Iterated Reproduction. *Current Biology*, *27*(3), 359–370. https://doi.org/10.1016/j.cub.2016.12.031

Jin, P., Zou, J., Zhou, T., & Ding, N. (2018). Eye activity tracks task-relevant structures during speech and auditory sequence perception. *Nature Communications*, *9*(1), 5374. https://doi.org/10.1038/s41467-018-07773-y

Johnson, K. (2004). Massive reduction in conversational american english. *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*, 29–54.

Jungers, M. K., & Hupp, J. M. (2009). Speech priming: Evidence for rate persistence in unscripted speech. *Language and Cognitive Processes*, *24*(4), 611–624. https://doi.org/10.1080/01690960802602241

Kello, C. T. (2003). Patterns of timing in the acquisition, perception, and production of speech. *Journal of Phonetics*, *31*(3-4), 619–626. https://doi.org/10.1016/S0095-4470(03)00017-2

Kentner, G., Franz, I., Knoop, C. A., & Menninghaus, W. (2023). The final lengthening of pre-boundary syllables turns into final shortening as boundary strength levels increase. *Journal of Phonetics*, *97*, 101225. https://doi.org/10.1016/j.wocn.2023.101225

Kim, K. S., Wang, H., & Max, L. (2020). It's About Time: Minimizing Hardware and Software Latencies in Speech Research With Real-Time Auditory Feedback. *Journal of Speech, Language, and Hearing Research*, *63*(8), 2522–2534. https://doi.org/10.1044/2020_JSLHR-19-00419

Kolly, M.-J., Boula De Mareüil, P., Leemann, A., & Dellwo, V. (2017). Listeners use temporal information to identify French- and English-accented speech. *Speech Communication*, *86*, 121–134. https://doi.org/10.1016/j.specom.2016.11.006

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., & Hagoort, P. (2018). Neural Entrainment Determines the Words We Hear. *Current Biology*, *28*(18), 2867–2875.e3. https://doi.org/10.1016/j.cub.2018.07.023

Lerner, G. H. (1996). On the "semi-permeable" character of grammatical units in conversation: Conditional entry into the turn space of another speaker. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and Grammar* (1st ed., pp. 238–276). Cambridge University Press. https://doi.org/10.1017/CBO9780511620874.005

Ligges, U., Krey, S., Mersmann, O., & Schnackenberg, S. (2023). *tuneR: Analysis of music and speech*. manual. https://CRAN.R-project.org/package=tuneR
R package version 1.4.3.

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541
R package version 0.13.1.

Mareüil, P. B. D., & Vieru-Dimulescu, B. (2006). The Contribution of Prosody to the Perception of Foreign Accent. *Phonetica*, *63*(4), 247–267. https://doi.org/10.1159/000097308

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386
Software version 1.0.1.

Meyer, L., Sun, Y., & Martin, A. E. (2020). Synchronous, but not entrained: Exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, *35*(9), 1089–1099. https://doi.org/10.1080/23273798.2019.1693050

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, *83*(5), 405–408. https://doi.org/10.1037/0033-295X.83.5.405

Novembre, G., & Iannetti, G. D. (2018). Tagging the musical beat: Neural entrainment or event-related potentials? *Proceedings of the National Academy of Sciences*, *115*(47), E11002–E11003. https://doi.org/10.1073/pnas.1815311115

Oberauer, K. (2022). The Importance of Random Slopes in Mixed Models for Bayesian Hypothesis Testing. *Psychological Science*, *33*(4), 648–665. https://doi.org/10.1177/09567976211046884

Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife*, *5*, e14521. https://doi.org/10.7554/eLife.14521

Peelle, J. E., & Davis, M. H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00320

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Potter, M. C. (1984). Rapid serial visual presentation (RSVP): A method for studying language processing. *New methods in reading comprehension research*, *118*, 91–118.

Pulvermüller, F. (2018). Neural reuse of action perception circuits for language, concepts and communication. *Progress in Neurobiology*, *160*, 1–44. https://doi.org/10.1016/j.pneurobio.2017.07.001

PyTSMod: An open-source Python library for audio time-scale modification. (Version 0.3.5). (2022). Retrieved October 21, 2022, from https://pytsmod.readthedocs.io/en/0.3.5/

Quené, H., & Port, R. F. (2005). Effects of Timing Regularity and Metrical Expectancy on Spoken-Word Perception. *Phonetica*, *62*(1), 1–13. https://doi.org/10.1159/000087222

R Core Team. (2023). *R: A language and environment for statistical computing*. manual. Vienna, Austria, R Foundation for Statistical Computing. https://www.R-project.org/ Software version 4.2.1.

Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition*. PTR Prentice Hall.

Rathcke, T. (in press). The p-centre effect and the domain of beat perception in speech. In L. Meyer & A. Strauss (Eds.), *Rhythms of Speech and Language: Culture, Cognition, and the Brain*. Cambridge University Press.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 978–996. https://doi.org/10.1037/a0021923

Repp, B. H., & Su, Y.-H. (2013). Sensorimotor synchronization: A review of recent research (2006–2012). *Psychonomic Bulletin & Review*, *20*(3), 403–452. https://doi.org/10.3758/s13423-012-0371-2

Riecke, L., Formisano, E., Sorger, B., Başkent, D., & Gaudrain, E. (2017). Neural Entrainment to Speech Modulates Speech Intelligibility. *Current Biology*.

Rimmele, J. M., Gross, J., Molholm, S., & Keitel, A. (2018). Editorial: Brain Oscillations in Human Communication. *Frontiers in Human Neuroscience*, *12*, 39–39. https://doi.org/

10.3389/fnhum.2018.00039
MAG ID: 2789600288.

Salomatin, A. (2018). *An example of a recorder based on AudioWorklet API. · GitHub*. GitHub. Retrieved January 22, 2024, from https://gist.github.com/flpvsk/047140b31c968001dc563998f744

Seeger, C. (1958). Prescriptive and descriptive music-writing. *The Musical Quarterly*, *44*(2), 184–195. https://doi.org/10.1093/mq/XLIV.2.184

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193–247. https://doi.org/10.1007/BF01708572

Smith, R., & Rathcke, T. (2017). Glasgow Gloom or Leeds Glue? Dialect-Specific Vowel Duration Constrains Lexical Segmentation and Access. *Phonetica*, *74*(1), 1–24. https://doi.org/10.1159/000444857

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87–90. https://doi.org/10.1038/416087a

Sperber, D., & Wilson, D. (1987). Précis of Relevance: Communication and Cognition. *Behavioral and Brain Sciences*, *10*(04), 697. https://doi.org/10.1017/S0140525X00055345

Stark, K., Van Scherpenberg, C., Obrig, H., & Abdel Rahman, R. (2022). Web-based language production experiments: Semantic interference assessment is robust for spoken and typed response modalities. *Behavior Research Methods*, *55*(1), 236–262. https://doi.org/10.3758/s13428-021-01768-2

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

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

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave: A free modular tool for sound analysis and synthesis. *Bioacoustics-the International Journal of Animal Sound and Its Recording*, *18*, 213–226. https://www.tandfonline.com/doi/abs/10.1080/09524622.2008.9753600
R package version 2.2.0.

Tal, I., Large, E. W., Rabinovitch, E., Wei, Y., Schroeder, C. E., Poeppel, D., & Golumbic, E. Z. (2017). Neural Entrainment to the Beat: The "Missing-Pulse" Phenomenon. *Journal of Neuroscience*, *37*(26), 6331–6341. https://doi.org/10.1523/JNEUROSCI.2500-16.2017

Tilsen, S. (2019). Space and time in models of speech rhythm. *Annals of the New York Academy of Sciences*, nyas.14102. https://doi.org/10.1111/nyas.14102

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

Tilsen, S., & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *The Journal of the Acoustical Society of America*, *124*(2), EL34–EL39. https://doi.org/10.1121/1.2947626

Turk, A., & Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Laboratory Phonology*, *4*(1), 93–118. https://doi.org/10.1515/lp-2013-0005

Turnbull, R. (2018). Patterns of probabilistic segment deletion/reduction in English and Japanese. *Linguistics Vanguard*, *4*(s2), 20170033. https://doi.org/10.1515/lingvan-2017-0033

Valbret, H., Moulines, E., & Tubach, J. (1992). Voice transformation using PSOLA technique. *Speech Communication*, *11*(2-3), 175–187. https://doi.org/10.1016/0167-6393(92)90012-V

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. https://doi.org/10.1016/j.neuropsychologia.2006.01.001

Venables, W., & Ripley, B. D. (2000). *S programming*. Springer Science & Business Media.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van Der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . Vázquez-Baeza, Y. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vogt, A., Hauber, R., Kuhlen, A. K., & Rahman, R. A. (2021). Internet-based language production research with overt articulation: Proof of concept, challenges, and practical advice. *Behavior Research Methods*, *54*(4), 1954–1975. https://doi.org/10.3758/s13428-021-01686-3

Wilson, D., & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics*, *38*(10), 1559–1579. https://doi.org/10.1016/j.pragma.2005.04.012

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, *12*(6), 957–968. https://doi.org/10.3758/BF03206432

Wynn, C. J., & Borrie, S. A. (2020). Methodology Matters: The Impact of Research Design on Conversational Entrainment Outcomes. *Journal of Speech, Language, and Hearing Research*, *63*(5), 1352–1360. https://doi.org/10.1044/2020_JSLHR-19-00243

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Ragni, A., Valtchev, V., Woodland, P., & Zhang, C. (2015). *The HTK Book (version 3.5a)*. Cambridge University Engineering Department.

Zellner, B. (1994). Pauses and the temporal structure of speech. In *Zellner, B.(1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41-62). Chichester: John Wiley.* (pp. 41–62). John Wiley.

Zhang, Y., Zou, J., & Ding, N. (2023). Acoustic correlates of the syllabic rhythm of speech: Modulation spectrum or local features of the temporal envelope. *Neuroscience & Biobehavioral Reviews*, *147*, 105111. https://doi.org/10.1016/j.neubiorev.2023.105111

Zhou, H., Melloni, L., Poeppel, D., & Ding, N. (2016). Interpretations of Frequency Domain Analyses of Neural Entrainment: Periodicity, Fundamental Frequency, and Harmonics. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00274

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase Entrainment of Brain Oscillations Causally Modulates Neural Responses to Intelligible Speech. *Current Biology*. https://doi.org/10.1016/j.cub.2017.11.071