OPEN ACCESS

University of BRISTOL

This electronic thesis or dissertation has been downloaded from Explore Bristol Research, http://research-information.bristol.ac.uk

Author:
Cucu, Maria Oana

Title:
Neural entrainment to acoustic edges in speech

General rights
Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

Take down policy
Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact collections-metadata@bristol.ac.uk and include the following information in your message:

• Your contact details
• Bibliographic details for the item, including a URL
• An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

Neural Entrainment to Acoustic Edges in Speech

Oana Cucu

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Life Sciences, School of Experimental Psychology. October 2019.
Word Count: forty nine thousand two hundred and eighty six

Acknowledgements

Without further ado, I thank my supervisors Conor Houghton and Nina Kazanina for helping me build up what ended up being a complete thesis.

But most of all, thanks to my friends Ollie, Ilinca, Brian, Chun and Ioana, and my mum and dad for putting up with my (numerous) existential crises, cheering me up and actually reading through and helping me out with various jumbled up pieces of my work. Wow. You guys.

Also I'd like to thank youtubers Chaz Smith for doing some quality, inspirational science and Natalie Wynn (ContraPoints) for her in-depth, layered analyses and social commentaries. Seriously, mind opening stuff.

I will dedicate this thesis to all of us out there trying to make sense of the world. It's hard.

Abstract

In the present thesis, I evaluate the role of acoustic edges in the neural tracking of speech syllables. Previous research has shown that neural oscillations exhibit phase locking to the slow temporal modulations (1-10 Hz) of the speech envelope, which is thought to correspond to the syllabic rhythm (Edwards & Chang, 2013; Ghitza, 2013; Giraud & Poeppel, 2012). It has been suggested that this is achieved through the phase resetting of ongoing neural rhythms to specific speech landmarks, such as the fine-grained spectral information placed at the onsets of syllables (Doelling et al., 2014). While some debate exists about whether entrainment occurs as a result of the phase resetting of endogenous oscillations or whether it is simply evoked activity which is temporally aligned to the rhythmic stimulus, I did not specifically investigate this distinction, but based on the present results, I suggest that further investigation into the role of syllabic landmarks in speech tracking is worthwhile nonetheless.

Experiment 1 replicated findings from Luo and Poeppel (2007), who suggested the importance of theta oscillations in tracking continuous speech. Here, we used stimuli such as natural speech sentences containing syllable-initial consonants which belonged to different phonemic categories, but we did not see differences in phase locking depending on the amount of edge provided by those phonemes. In Experiment 2, we used series of nearly-isochronous consonant-vowel syllables starting with separate phonemes and showed that syllables starting with some consonants led to less phase locking than others (lowest for sibilants, highest for stops). We also explored different edge markers based on the acoustic properties of the stimuli and, following from suggestions from other research such as Oganian and Chang (2018), considered that information which is critical for speech tracking may be found at the consonant-vowel transition of syllables. In Experiment 3, I tested this hypothesis by placing two different types of noise at various locations of “da” and “ta” syllables. We found that differences in phase locking due to the insertion of noise were the most striking at CV locations and also, the direction of change in entrainment depended on the syllable-initial consonant.

I suggest the different phonemes provide different acoustic edges for syllable tracking and that these are most prominent at the CV transition. This claim needs to be tested in the future for a variety of consonants, syllabic structures as well as for continuous speech, but could have crucial implications for the way we currently understand neural phase locking to the speech envelope.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: …… DATE: ……..

Table of Contents

List of Figures

List of Tables

This page was intentionally left blank.

1. General Introduction

In this thesis, I evaluate mechanisms of neural speech tracking, discussing existing theories and presenting evidence from three electroencephalography (EEG) studies, as well as a complementary behavioural study. In the present Introduction, I will briefly evaluate the purpose of neural oscillations in speech processing, as well as mechanisms of entrainment.

In Chapter 1, I will briefly cover the theory that the brain parses speech information by entraining to the syllabic rhythm, which is thought to be conveyed primarily by the slow fluctuations (below 10 Hz) in the acoustic envelope (e.g., Ghitza, 2013; Giraud & Poeppel, 2012). Specifically, I am interested in the notion that the endogenous neural theta rhythm (4-8 Hz) resets its phase when encountering edges in speech (e.g., Doelling, Arnal, Ghitza, & Poeppel, 2014; Gross et al., 2013). Current research has not established the nature of these edges, but suggestions have been made for a variety of syllabic landmarks, such as the acoustic content present in the onset of the syllables (e.g., Oganian & Chang, 2018), or vowel peaks (Ghitza, 2013).

In Chapter 2, I will describe an EEG experiment in which we manipulated the edges present in the syllabic onsets of continuous speech through the nature of the consonants at those locations. We measured the amount of phase locking to the low frequencies (1-10 10 Hz) of stimuli whose syllable-initial consonants were either plosives or belonged to other phonemic categories. In Chapter 3, we investigated the roles of separate syllable-initial consonants on neural entrainment to nearly-isochronous stimuli, which comprised of repetitions of consonant-vowel or vowel-only syllables. Chapter 4.A. summarizes an EEG experiment in which two syllables, “da” and “ta”, contained slight amounts of noise in their respective onsets, formant

transitions or vowel peaks, and where these syllables were presented to participants in an isochronous fashion. Differences in syllabic entrainment due to the noise present at different locations were considered and possible perceptual influences on the results were explored in a behavioural experiment outlined in Chapter 4.B. Lastly, Chapter 5 summarizes and discusses the findings of my research from the perspective of the current views regarding mechanisms of neural phase locking to the syllabic rhythm of speech.

1.1 Neural processing of speech

1.1.1. The speech envelope

Over the last few decades, neural oscillations have been shown to be involved in a variety of cognitive functions, including speech. Using a range of methods such as EEG (Di Liberto & Lalor, 2017; Di Liberto et al., 2015; Khalighinejad et al., 2017), magnetoencephalography (MEG: Howard & Poeppel, 2012; Luo & Poeppel, 2007; Peelle & Davis, 2012) and electrocorticography (ECOG: Mesgarani, Cheung, Johnson, & Chang, 2014; Nourski et al., 2009; Zion Golumbic et al., 2013), researchers have shown that neural oscillations show phase-locking to the acoustic waveform of speech and, in particular, to its slow temporal modulations present in the envelope.

The importance of temporal fluctuations in perception was initially suggested by studies investigating comprehension to altered speech, particularly those in which certain frequency ranges were removed from the acoustic signal. Importantly, a reduction in intelligibility of sentences and detection of individual speech sounds was

found more when slow temporal fluctuations (<10 Hz) of the envelope were removed as opposed to higher frequencies (e.g., Drullman, Festen, & Plomp, 1994a,b).

The low frequencies in speech are amplitude modulations of a carrier signal containing fine grained frequency information, which itself is generally over 600 Hz. Shannon, Zeng, Kamath, Wygonski and Ekelid (1995) showed that speech stimuli were still intelligible when the high frequency granularity was degraded but when the envelope was preserved. However, intelligibility was affected if low frequencies, especially below 16 Hz, were removed. Smith, Delgutte and Oxenham (2002) suggested that this may be because spectral and envelope information have different roles. By using chimaeras which combine the envelope of one stimulus with the fine-grained information of another and vice versa, the authors found that, for speech, participants understood the sentence based on envelope and not fine structure. However, the opposite was true for melody-melody chimaeras, with identification of the stimulus relying on the tune which provided the fine-grained information.

Drullman et al. (1994a) found that the most significant reductions in intelligibility could be attributed to envelopes low pass filtered at 4 Hz, or containing only information between 0 and 4 Hz, and that these reductions were progressively smaller if all the information below 8, 16 and 32 Hz was retained. The differences in performance due to low-pass filtering at 16 or 32 Hz were the smallest. Conversely, Drullman et al. (1994b) showed that the envelope high pass filtered at 4 Hz or below did not lead to changes in comprehension and that such deteriorations were observed only after high pass filtering at 8 Hz or above. Thus, frequencies between 4 and 8 Hz were revealed to play the greatest contribution to intelligibility, based on these two studies.

By reviewing a body of evidence regarding the role of slow modulations in comprehension, Edwards and Chang (2013) proposed that fluctuations between 2 and 5 Hz are those most consistently found as crucial for the perceptual detection of acoustic changes. While some of these studies do not involve speech specifically, their findings bear a strong resemblance to those from speech experiments. For example, a very early study conducted by Shower and Biddulph (1931) found that humans detected changes in the pitch of a noise signal if this was modulated at two or three cycles per second (2-3 Hz). Similarly, by altering modulations between 3 and 7 Hz in speech stimuli, Elliott and Theunissen (2009) found that participants were not able to identify whether the pitch of the acoustic signals sounded male or female.

As we shall see later in this Introduction, different frequency ranges belonging to slow envelope fluctuations are considered to correspond to the durations of different speech units, such as syllables or phrases. But before delving deeper into particular speech acoustics and their neurocognitive complements, I will first attempt to illustrate how the envelope may affect neural processing and subsequently speech comprehension.

Currently, a leading theory explaining the neural tracking of speech states that endogenous neural oscillations, particularly in the delta and theta range (2-8 Hz), reset their phase in order to match the oscillatory pattern of the slow envelope frequencies (Schroeder & Lakatos, 2009). Endogenous neural oscillations reflect cyclical patterns of excitability within local cell populations. To illustrate this, Schroeder and Lakatos (2009) describe a series of studies investigating local-field potentials in various frequency bands (alpha, delta, gamma, etc.) showing how negative voltage fluctuations of neuronal ensembles correspond to increases in firing

(high excitability) whereas positive deflections reflect states of hyperpolarisation. Importantly, by aligning high excitability neuronal states with the onset of a rhythmic stimulus, the brain is thought to be capable to predict the incoming stimulus and to select relevant information (Zoefel, ten Oever and Sack, 2018). This is particularly attractive notion from a scientific perspective because it suggests that, through neural entrainment, not only are oscillations involved in cognitive functions, but also that they play an active role in information processing. However, as we briefly summarise below, research studying neural entrainment has shown few results in favour of this theory, although more recent investigations have been promising. In the next section, we will review the evidence presented in favour or against a significant role for neuronal entrainment in processing stimuli and inquire whether this process may or may not be useful to speech tracking.

1.1.2 Processing rhythmic stimuli in the brain

In a comprehensive review of the involvement of neural oscillations in rhythmic processes, Zoefel et al. (2018) note that regular extrinsic stimulation will always trigger a repetition of phase-aligned evoked responses in the brain, in the form of steady-state potentials (see Appendix 1.1.5 and onwards), which can also be considered a form of entrainment - or, what Obleser and Kayser (2019) define as ‘entrainment in the broad sense’. Steady-state evoked responses have been observed in a variety of modalities and it has been established that there is phase coherence between these and regular stimuli. On the other hand, to show that phase resetting of endogenous activity happens on top of these responses has often been challenging.

Some studies draw evidence for the entrainment of ongoing oscillations from the appearance of imagined rhythms, or rhythms otherwise not present in the stimulus, but which can be observed in neural recordings. For example, Nozaradan, Peretz, Missal and Mouraux (2011) found sustained evoked responses in the EEG of participants when they were asked to imagine a binary or tertiary rhythm on top of a given beat, not only to the stimuli, but also to the imaginary beats. Ding, Melloni, Tian, Zhang and Poeppel (2016) presented participants with monosyllabic words (4 Hz) which constructed regular phrases (2 Hz) and sentences (1 Hz), with the latter two being unidentifiable in the speech spectrum of the stimuli. Nonetheless, they found that similar trials elicited neural responses with similar phase profiles (increased phase coherence) at the frequency of phrases and sentences. Furthermore, Zoefel and VanRullen (2015) found that EEG responses to speech-noise stimuli which do not show any low-frequency envelope fluctuations elicited oscillatory activity in the low-frequency spectrum which is phase-locked to the stimulus. This potentially indicates the entrainment of endogenous oscillations. However, one criticism of such studies is that it is possible for imaginary rhythms, as well as perceived phrases and speech regularities outside of the envelope to lead to evoked activity, which could be phase locked to the respective stimuli.

Capilla, Pazo-Alvarez, Darriba, Campo and Gross (2011) showed that responses to isochronous visual stimulation were better explained as a superposition of event-related potentials (ERPs or evoked potentials: see Appendix 1.1.4) than as entrainment of endogenous oscillations. They presented checkered patterns which were either regular or jittered and simulated the recorded responses both as superpositions of evoked activity (i.e., the ERPs pre- and post-baseline were multiplied by a Gaussian) or oscillatory entrainment (by averaging brainwaves to

jittered frequencies). What they found was that the superposition of evoked responses matched the recorded data better than the entrainment simulations. Another finding of this study was that lack of any additional neural activity post-stimulus, much like in other studies investigating steady-state responses: this is also normally taken as an argument against entrainment, as the rules of oscillation synchronisation would predict that an entrained oscillation continues even after the cessation of stimulation, albeit in a damped fashion (Zoefel et al., 2018).

However, some studies did find evidence for entrainment in some very elegant ways, in both the visual and auditory modalities. One such study by Notbohm, Kurths and Herrmann (2016) involved the apparition of an Arnold Tongue in the phase coherence profile of brainwaves observed in response to a visual stimulus oscillating in the alpha range. Entrainment between two oscillators is stronger if their frequencies match and, if this is not the case, the level of entrainment depends on both the intensity of the stimulation as well as the distance between the stimulation frequency and the oscillator's eigenfrequency (Pikovsky & Rosenblum, 2007). When entrainment is plotted as a function of both stimulation frequency and intensity, the relationship above is described by a triangular shape named the Arnold tongue (see Figure 1.1). Furthermore, an intermittent pattern of entrainment can be observed at the border of the Arnold Tongue (alternating bands of synchronisation and decoupling). This pattern is exactly what Notbohm et al. (2016) observed when investigating steady-state visual evoked potentials (SSVEPs) to a visual flicker. Entrainment (as measured by the Shannon entropy) was affected both by the light intensity of the stimulus and the distance between the stimulation frequency and the intrinsic alpha frequency, measured for each participant as the peak in power between 9 and 11 Hz.

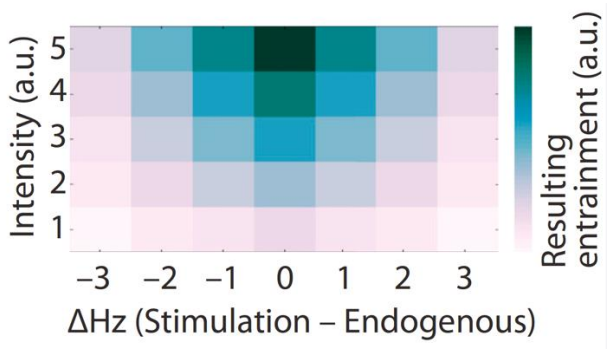


Figure 1. 1. Shorter distances between the frequency of stimulation and the eigenfrequency of endogenous neural oscillations require less intense stimulation and vice versa. The relationship between entrainment levels and the intensity and frequency of stimulation is described by a triangular shape known as the Arnold Tongue. Reproduced from Obleser and Kayser (2019).

A more recent study by Zoefel, Archer-Boyd and Davis (2018) showed the importance of ongoing oscillations in speech perception. By delivering transcranial alternating current stimulation (tACS) during an fMRI study, the authors manipulated the phase of neuronal oscillations at which a speech stream was delivered to the participants. Importantly, the phase at which the speech was delivered impacted BOLD responses to intelligible, but not unintelligible stimuli, emphasising a clear cognitive role of endogenous oscillations in speech processing. However, it still needs to be verified whether the same holds true for natural speech, with varying envelope fluctuations (Lalor, 2018). In fact, so far the existing literature does not report clear evidence in favour of entrainment when quasi-regular stimuli are used (Zoefel et al., 2018). One such study by ten Oever et al. (2014) presented participants with either isochronous or jittered auditory stimuli whose average durations were the same as those of isochronous stimuli. While isochronous stimuli led to phase coherent neural responses even when the stimulation was below threshold, this was never the case for jittered stimulation. While the stimuli in this study were not like speech, such results add further uncertainty about the validity of neural entrainment to quasi-regular stimuli, including speech.

Given the conflicting evidence, it might be useful to consider ‘entrainment in the broad sense’, as suggested by Obleser and Kayser (2019). They propose that ‘entrainment in the broad sense’ should be used whenever there is temporal phase alignment between the neural response and the stimulus, but when the process giving rise to this is unclear. They also give a compelling example of when spectro-temporal response functions (STRFs) are applied to estimate evoked impulse-type responses to speech: STRFs are linear models of the receptive fields of auditory neurons in the time-frequency domain; STRFs to speech often show a peak in frequency in the theta range despite any obvious neural oscillation at the same frequency in the raw data. It remains unclear why this happens: endogenous oscillations may be too small to be noticed. However, by considering such phase alignment as entrainment, whether in the ‘broad sense’ or not, allows us to investigate whether the neural processes in question are unique to speech processing.

In the present thesis, I will use the word ‘entrainment’ whenever there is phase consistency between the brain response and the speech or speech-like stimulus, or between brain responses to the same stimuli. This will mainly refer to ‘entrainment in the broad sense’, as Obleser and Kayser (2019) suggested. Where necessary, an explanation will be provided as to whether the obtained results were more likely to be due to synchronisation of endogenous oscillations or merely evoked activity. A summary of how ‘entrainment in the narrow sense’ may happen is also provided in the Appendix, based mainly on interpretations by György Buzsáki, whose pioneering work revealed the importance of neural rhythms in cognitive functions. In the following section, I will consider specific aspects of phase alignment of neural responses which are characteristic to speech tracking.

1.1.3 Neural windows of activity relevant for speech tracking

Speech tracking shows selectivity in terms of phase activity which depends not only on the stimulus properties, but also on the frequency window of oscillatory activity. In a now classical MEG study, Luo and Poeppel (2007) showed that the phase patterns of MEG responses were unique for each sentence which was played to participants and that consistency within responses to the same trials was observed only in the theta (4-8 Hz) range. Moreover, the power of the MEG, which measures the magnitude of the oscillations, did not differ from baseline, indicating no detectable evoked activity in response to speech stimuli. Thus, the authors point towards a primary mechanism involved in speech processing being the phase reset of endogenous theta oscillations to incoming speech stimuli.

Some researchers suggest that the brain directs specific windows of activity which are relevant in speech tracking. One experiment by Luo and Poeppel (2012) implies exactly this. The researchers created noise stimuli which were modulated at different time intervals of 25, 80 and 200 ms, respectively. Each of these correspond to frequencies of 40, 12.5 and 5 Hz. The experiment found that neural oscillations of 5 and 40 Hz were phase locked to corresponding stimuli, but not those of 12.5 Hz. This suggests that discrete oscillatory mechanisms could also be involved in speech perception, as the noise stimuli were spectrally similar to speech. Luo and Poeppel (2012) emphasise the theta and gamma ranges as particularly important for detecting acoustic properties.

Nonetheless, studies investigating speech tracking found that, in response to speech stimuli, theta (4-8 Hz) oscillations show increased temporal consistency (e.g. phase coherence), whereas gamma (25-40 Hz) show increased evoked activity (oscillatory power). (for a review see Ding & Simon, 2014). These findings were also

obtained to continuous speech stimuli (Luo & Poeppel, 2007), and sometimes in the absence of a specific task (Doelling et al., 2014). Such research indicates the primary importance of theta oscillations in speech perception, and indeed, some scientists seem to favour this theory, suggesting that other shorter temporal windows used in speech sound recognition are coordinated by lower frequency oscillations (Ghitza, 2013; Giraud & Poeppel, 2012).

But what do different time scales represent, and what does the distinction between them tell us? Human speech can be divided into units of different regular temporal granularities such as phonemes, syllables, and phrases (Meyer, 2018). For example, in the sentence “Daniel carried Jasmine’s heavy suitcase”, the single phrase “Daniel carried” contains four syllables. “Dan” is one such syllable and this contains three phonemes, such as /d/, /æ/ and /n/. Phrases, phonemes and syllables are all identifiable in the slow temporal modulations of speech (see Figure 1.1), between 1 and 50 Hz, with phrases falling in ranges below 2-4 Hz, syllables between 3 and 8 Hz, and phonemes somewhere above these, up to 50 Hz (Ding et al., 2017; Meyer, 2018). It is impossible to give the exact limits of the duration intervals for all units, especially considering all languages and differences between speakers, and there is a definite overlap between all of their durations, (Edwards & Chang, 2013). However, as we shall see next, each of these different units plays an important role in speech perception.

1.1.4 Syllables

Both neural phase locking and behavioural comprehension studies emphasise syllables as essential for neural speech processing (Ding & Simon, 2014; Ghitza, 2013; Greenberg et al., 2003). This is mainly because the syllabic durations are the

most related to frequencies at which oscillations show a peak in entrainment. For example, in American English, most syllables fall between 40 and 400 ms, based on the available recorded data (such as phone dialogues from the SWITCHBOARD corpus in Greenberg et al., 2003), a range which corresponds to the rough duration of a neural theta cycle.

In general, syllables are not easy to define. One could say that they are speech sounds generally produced by the opening and closing of the vocal tract a single time. Most syllables are easy to count, such as, ‘cat’ is a single syllable word and ‘water’ is a word formed of two syllables (Davenport and Hannahs, 1998). Syllables can also be identified as the onsets and offsets of the envelope peaks. For example, Figure 1.1 shows that the number of syllables in the sentence ‘Daniel carried Jasmine’s heavy suitcase’ (10) roughly corresponds to the number of peaks which we identified (13).

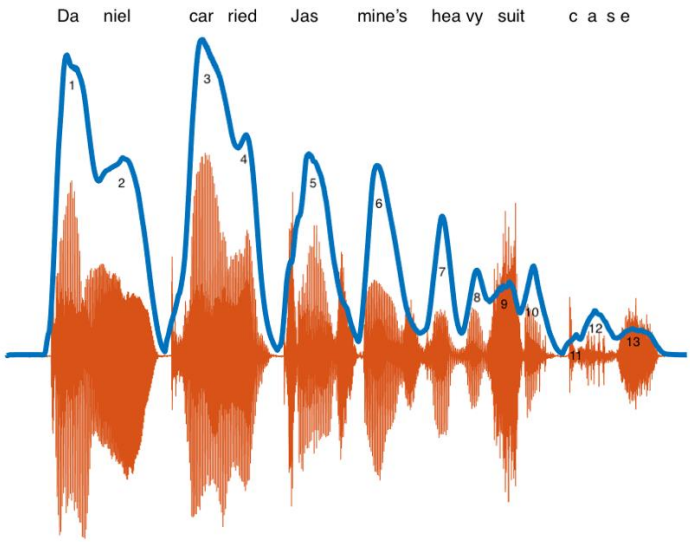


Figure 1. 2**.** Waveform of spoken sentence ‘Daniel carried Jasmine’s suitcase’ plotted in orange, and its envelope is plotted in blue. Spaces in the text above the envelope are used to represent the approximate location of the syllables or sounds with respect to the envelope. Numbers correspond to envelope peaks.

But there are exceptions to these rules: Cummins (2012) indicated that some words may consist of different numbers of syllables depending on both speaker pronunciation and the listener's interpretation. For example, the word 'naturally' can be thought of as having either three or four syllables. The same can be argued for the onsets and offsets of the envelope, not all of them belonging entirely to syllables. Figure 1.1 also helps illustrate this. The words 'Daniel' (peaks 1,2) and 'carried' (peaks 3,4) can be attributed with only one set of envelope peaks each, or be considered as containing only one syllable. Conversely, smaller envelope onsets seem to correspond mainly to individual sounds: envelope peaks 9 and 13 contain only the sound /s/ in syllables 'suit' and 'case', respectively, while peak 11 corresponds to /k/ in 'case'.

So how can we be sure that the brain tracks syllables during speech tracking? The theory formed around the time scales of speech may be able to clarify this better. Ghitza (2013) argues that the quasi-regular syllabic rhythm may have evolutionarily enabled us to dedicate oscillatory mechanisms at corresponding frequencies, in order to follow the specific durations of different speech units, which allows us to better process information. This view could somewhat be countered by the fact that phase locking to the speech envelope was seen in animals, for whom syllables or other speech units hold little significance (Nourski et al., 2009; Steinschneider et al., 2013), or given that the brain can also entrain to the envelope of noise stimuli (Luo & Poeppel, 2012), which cannot be divided into meaningful components. Of course, such findings do not imply that tracking speech is identical to tracking noise (it is not: the STRFs of noise stimuli are different from those of speech stimuli: David, Mesgarani, Fritz, & Shamma, 2009), or, in fact, that the theta

rhythm cannot be used to parse speech by tracking the syllabic rhythm, but it is not always clear how tracking the speech envelope is relevant for parsing information.

Nonetheless, evidence in favour of the idea that the theta rhythm corresponds to syllable tracking comes from numerous studies, amongst which we can enumerate experiments investigating speech compression. These have shown that progressively shortening the duration of “theta syllables” leads to proportionally less comprehension (Ghitza, 2014; Ghitza & Greenberg, 2009) and less phase locking (Ahissar et al., 2001). Moreover, intelligibility seems to be especially poor for stimuli with a compression rate above a factor of 3, or whose syllabic rates are above 9 Hz (Ghitza & Greenberg, 2009). Ghitza (2013) argues that the brain can track compressed syllables, but not if these are shorter than a single theta cycle. A study investigating the effects on target word detection of lengthened syllables complements this idea, by showing a decrease in performance with increased stimulus modification (Baese-Berk et al., 2014). The theory behind these findings is that only “theta-syllables” may convey the rhythm which allows the brain to ‘repackage’ speech information into segments (Ghitza, 2013).

Dividing incoming information into segments may also come from the processing limitations of working memory (see Logie, 2011, for a review). Working memory capacity is normally limited to three to five items which can be stored concomitantly, as suggested by various verbal or visual memorisation tasks (Cowan, 2011). However, dividing information appropriately seems to increase the amount of information which can be stored in memory. For example, people remember sequences, especially digit sequences, if these are chunked into shorter ones (Mathy & Feldman, 2012) and it has been argued that the duration of holding one of these chunks active in memory corresponds to one theta cycle (Buzsáki, 2005).

Hippocampal theta appears during information encoding in rats (Penttonen & Buzsáki, 2003). Furthermore, attention is thought to facilitate chunking mechanisms involved in working memory (Bor & Seth, 2012). In speech, attended signals lead to more entrainment of theta oscillations than non-attended ones (e.g., Besle et al., 2011; Zion Golumbic et al., 2013). For example, this has been shown by Zion Golumbic et al. (2013) using a cocktail party effect paradigm, in which participants listen to two different sound streams played to each ear, and ask to direct attention to either stream. However, Zion Golumbic et al. (2013) found that even unattended speech shows robust phase locking to neural theta oscillations. Nonetheless, the exact relationship between auditory parsing, attention and working memory has not been determined.

The quasi-regularity of the syllabic rhythm may explain why entrainment to envelope fluctuations are seen in animals, who cannot process meaning, or to speech-like noise stimuli. The modulation spectrum of speech shows that its envelope frequencies dominate in the 3-5 Hz range, and this is thought to correspond to the most frequent syllable durations (Ding et al., 2017). Furthermore, Ding et al. (2017) found similar modulation spectrum peaks across languages; speech modulation spectrums were also distinct from those of different types of music, whose modulation peaks were typically seen below 3 Hz. Consequently, the brain may dedicate oscillations of appropriate durations to distinguish between acoustic chunks of approximately equal durations, such as speech syllables.

Indeed, phase locking to the envelope in the theta range may rely on acoustic factors alone, preceding intelligibility. For example, one study comparing the phase entrainment to both normal and reversed speech found no difference between the two in the theta band, supporting the claim that comprehension is not a pre-requisite

of entrainment to syllables (Howard & Poeppel, 2010). It was also found that listening to speech in a noisy environment can lead to both deficits in intelligibility and phase locking to the acoustic envelope of the target stimulus, but this is due to degradation in the spectral resolution of the target speech signal in the presence of environmental noise (Ding & Simon, 2013a; Peelle, Gross, & Davis, 2012). An exact context of intelligibility (i.e., whether this refers to comprehension due to meaning or to amounts of signal-to-noise ratios) is nonetheless necessary when referring to the impact of top down mechanisms on speech tracking.

1.1.5 Phonemes

Slow fluctuations, albeit of higher frequency, in the speech envelope also correspond to individual speech sounds, or phonemes. The Drullman et al. (1994a,b) studies revealed that removing envelope modulations of frequencies above 8 Hz also impaired recognition of phonemes. Furthermore, recovering phonetic spectral information boosts phase locking and is also able to better predict neural responses to speech (Di Liberto & Lalor, 2017; Di Liberto et al., 2015). The windows of oscillatory activity dedicated for the perception of phonemes are thought to lie in the beta (12-30 Hz) and gamma (25-100 Hz) ranges (Ghitza, 2013; Luo & Poeppel, 2012). Hence, phonemes can be thought of as units of shorter duration than syllables which provide essential qualitative information about the speech signal.

Phonemes are the smallest units of speech which distinguish word meanings. For example, ‘can’ and ‘pan’ are different in their initial phonemes, /k/ and /p/. Syllables are made of one (e.g., the first syllable in *O-li-ver*) or multiple phonemes (e.g., second and third syllables of *O-li-ver*). Depending on the configuration of the speech organs during their articulation, phonemes can be classified into multiple

categories. While articulatory phonetics is not the main scope of this discussion, the differences in articulation pertaining to different phonemic categories give rise to acoustic properties which will be mentioned extensively throughout this thesis. Importantly, distinctions between the acoustic properties of different phonemic categories may be reflected in the speech envelope, as indicated below. The following paragraphs relating to phonemes are a summary of information found in Davenport and Hannahs (1998) and Stevens (2000).

Figure 1.2 shows a vocal tract, with its range of organs needed for articulation. Articulatory organs impose different restrictions on the air flow, which mainly originates in the lungs, travelling upwards through the trachea. This direction of air flow travel is not the only possible one, but it is the most common one across all languages. One of the most important organs which is encountered by the air flow in this direction is the larynx, where the vocal cords are. The opening between the vocal cords is called the glottis. The opening or stricture of the glottis determines whether the sound is voiceless or voiced. In the first case, the vocal cords are far apart, allowing the air to pass freely and in the latter case, the air flows through a small stricture in the glottis, creating pressure against the vocal cords and making these vibrate. The frequency at which the vocal cords vibrate is called the fundamental frequency (F0) or the pitch, and depending on how high this is one can recognise whether the speaker is a child, a man or a woman.

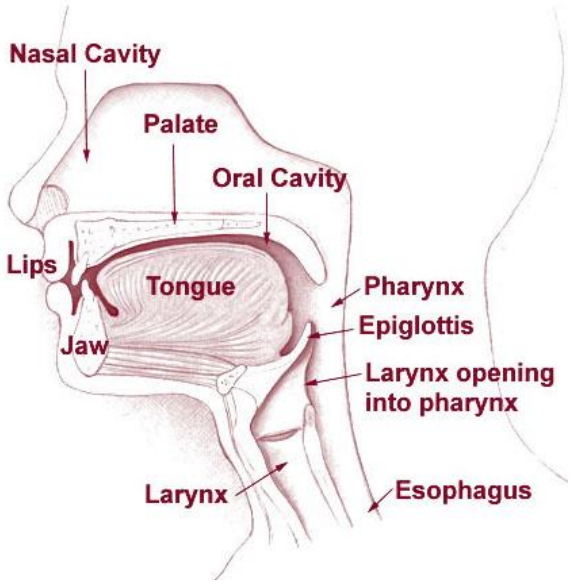


Figure 1. 3. Schematic illustration of the articulatory organs in humans. (Image downloaded from https://www.daviddarling.info/encyclopedia_of_music/V/vocal_tract.html)

Different configurations of articulators and particularly the relationships which exist between articulators in the oral tract, generate different phonemic categories. These can be active or mobile articulators, such as the lips and tongue, or passive ones, such as the pharynx wall, the roof of the mouth, the teeth and the upper lip. The vertical distance between them creates a stricture in the oral cavity which can impose different restrictions on the air flow travelling out. This relationship between the active and passive articulators is called the manner of articulation.

For example, the articulators can be pressed together to first block the air flow, in what is known as closure, but when the vocal tract subsequently opens, the air is expelled quickly. Some phonemes which exhibit a complete closure followed by a release of the air flow are called stops consonants. Most consonants exhibit full or partial restriction of the air flow, whereas vowels are another major phonetic category which are pronounced with an open vocal tract. If the vocal tract is open during pronunciation, this will resonate and lead to the apparition of harmonics of F0 in the spectrogram of the sound. This can generally be seen for vowels, where the first

three formants F1, F2 and F3 are the most important, as well as for some consonants which are called sonorants.

Consonants and vowels can also be distinguished by their position within a syllable. Consonants cannot be produced in isolation, so they will be found either at the onset (beginning) or coda (end) of a syllable. Conversely, vowels are always at the centre of syllables. In the English language, some semi-vowels or glides are considered consonants because they are only present in syllabic onsets (such as /w/ in ‘we’) (Davenport and Hannahs, 1998). However, just like vowels, the vocal tract is also open during the pronunciation of glides, albeit to a slightly smaller degree.

When pronounced together in syllables, consonants and vowels influence each other. This can be seen both in the spectrogram of the speech signal as well as in its amplitude envelope. For example, the presence of formants can be seen in the part of the spectrogram which temporally corresponds to the pronunciation of a consonant, even if the formants belong to the vowel. Furthermore, the F1 formant of a vowel shows a downward movement during the closure of a consonant, and an upwards movement during release. These changes are also reflected in the envelope, with the amplitude increasing or decreasing abruptly near consonants (Stevens, 2000).

Later in this thesis I will review the manner of articulation and voicing of different consonants, as well as the syllabic interplay between vowels and consonants, and how these affect both the envelope and neural tracking. A separation of different consonants based on voicing, manner of articulation, and constriction of the vocal tract is also provided in Table 1 of Appendix 1.2. As we shall see, the phonetic colouring of speech may be able to provide different acoustic landmarks which could facilitate entrainment to speech. But before going in more in

depth about this topic, I will first briefly mention the phrase units of speech, which can also be identifiable in the envelope.

1.1.6 Phrases and prosody

Envelope fluctuations below 4 Hz are thought to correspond to intonation patterns which reflect the boundaries between sentences or phrases (Bourguignon et al., 2013). Consequently, coherence between speech and delta oscillations is thought to reflect the processing of phrases and prosody elements. Prosody refers to suprasegmental speech structures containing stress and intonation patterns, which help identify syntactic cues (Friederici, 2004). For example, an MEG study by Bourguignon et al. (2013) found that the phase locking of delta oscillations was the highest at the frequency corresponding to the average occurrence of pauses at the end of sentences.

The fact that delta oscillations in the brain are assigned to the processing of phrasal structures or hierarchical dependencies in speech is confirmed by research showing neural coupling at the frequency level of phrases and sentences, despite the lack of prosody elements. Using monosyllabic words of equal durations, which were artificially produced so that they contained identical stress and intonation markers, Ding et al.(2016) found that phase entrainment was present in the frequency of both noun and verbs phrases. However, this was found if participants listened to words spoken in their native language, with foreign words leading to phase locking at the syllabic level only.

In the Ding et al. (2016) study, the lack of entrainment to phrases when listening to foreign stimuli suggests that, unlike syllabic and phonemic processing, phase locking in the delta range is dependent on intelligibility. This theory is

supported by studies on continuous speech. Like Howard and Poeppel (2010), Molinaro and Lizarazu (2018) showed that there was no difference in entrainment to natural and reversed speech in the theta range. However, they found that phase locking to frequencies below 4 Hz was higher to normal than to reversed speech. Together with the Ding et al., (2016), these findings suggest that delta oscillations are necessary in the processing of hierarchical structures and long dependencies between syntactic units of speech, which convey complex meaning.

However, the stress patterns contained in the slow temporal modulations of the envelope may help speech parsing, and may consequently aid entrainment to both syllabic and phonemic units. This is shown by a series of studies investigating the role of infant directed speech on entrainment (Leong, Kalashnikova, Burnham, & Goswami, 2014, 2017). The low frequency fluctuations in infant directed speech correspond to more regularly stressed syllables, and children below three years of age showed more entrainment when listening to infant than to adult directed speech, in both neural delta and theta ranges. These studies illustrate that even though the brain may operate at different time scales depending on the identification of different types of units, which may have specific roles, there is a clear interdependence between oscillations of difference frequencies.

1.1.7 Mechanisms of envelope tracking

Research has shown that envelope tracking is best when its slow temporal fluctuations are also sharp. For example, Prendergast, Johnson, & Green, (2010) showed that entrainment to the envelope of tone sequences, which were repeated between 150 and 300 ms, depended both on the duration of the tone and on the amplitude modulation of the sequence, with a sinusoidal amplitude leading to the

least amount of phase locking, and short tones presented as clicks being the most correlated with oscillatory tracking in the theta range. Such findings have led scientists to claim that acoustic edges may be responsible for parsing of continuous sounds, which additionally reflects in more successful tracking.

In speech, this has been best shown by a study by Doelling et al. (2014), who investigated the effects of envelope sharpness on theta entrainment and intelligibility. Using sequences of spoken digits as stimuli, which contained various types of alterations, Doelling et al. (2014) showed that even when envelope fluctuations were removed, playing the carrier signal with clicks added at the previous locations of the syllables recovered both entrainment and intelligibility, compared to results obtained from presenting the carrier signal alone. The waveforms of some of the different stimuli used by Doelling et al. (2014) are found in Figure 1.3. The stimuli containing clicks (Figure 1.3.E) were greater than the carrier signal stimuli (Figure 1.3.C) in a measure called sharpness, which Doelling et. al (2014) defined as the sum of first derivative of the stimulus envelope, which corresponded to the level of ascending slopes present in the amplitude fluctuations. However, the click stimuli were not sharper or more intelligible, and did not lead to more entrainment than stimuli in the control condition (Figure 1.3.B), containing only the envelope, but not higher frequency information. Doelling et al. (2014) also manipulated the original stimuli into ones containing a higher level of sharpness than the control (Figure 1.3.D) , and showed that a higher degree of sharpness indeed led to more phase locking between the theta oscillations and the stimulus envelope. Nevertheless, these stimuli were also the least intelligible, suggesting that sharpness alone does not lead to comprehension, but however confirming that intelligibility is not necessary for theta entrainment.

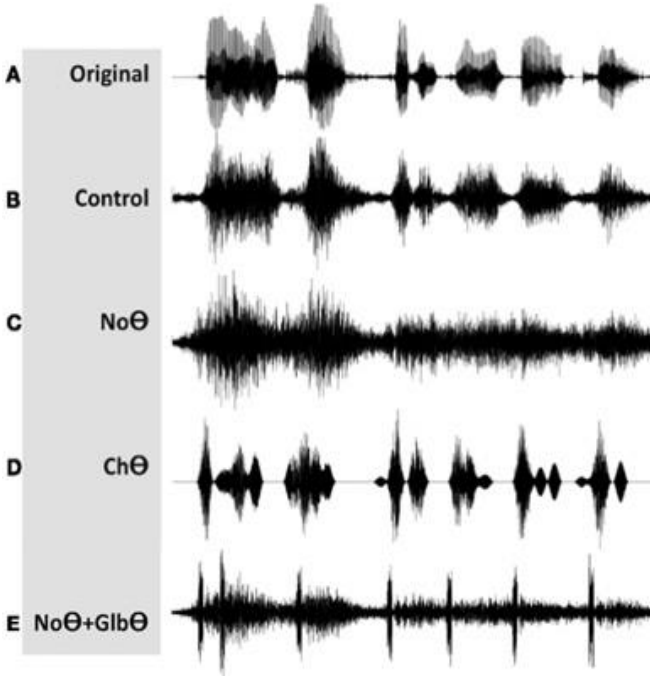


Figure 1. 4. Waveforms of four of the manipulated stimuli used by Doelling et al. (2014), as well as the waveform of the original sound (A). B**.** Waveform of control stimulus, which maintained the carrier signal, and slow modulations only between 1 and 10 Hz. C. Carrier signal only. D. Condition where the carrier signal was tightly modulated around syllabic peaks. E. The carrier signal with added clicks at syllabic locations. The order of the sharpness and elicited coherence was C<E<B<D. The order of intelligibility was D<C<E<B. Responses to original sounds were not measures. (Picture downloaded from frontiersin.org).

In natural speech, the amount of edges in a syllable could depend on its phonemic content. Syllabic rise time, or the latency measured from the onset of a syllable until its peak, differs depending on the spectral properties of the phoneme located at the syllabic onset (Goswami & Leong, 2013). For example, a syllable with a stop consonant at the onset will have a shorter rise time than one with a glide. It has been shown that children with dyslexia, who show impairments in both phonemic perception and entrainment to speech in the theta range (Alan J. Power et al., 2016; Thomson et al., 2013), are capable of detecting shorter rather than longer rise times (Goswami & Leong, 2013).

Potential landmarks for entrainment include rise times, as well as the onsets and nuclei of syllables. In a linguistic study, Greenberg et al. (2003) showed that stressed syllables differ in their amplitude fluctuations from non-stressed syllables, particularly in the nucleus and to some degree, in the onset, with the coda not being affected at all. Both Ghitza (2013) and Doelling et al. (2014) suggested syllabic onsets and vowel peaks as primary landmark candidates. A more recent ECOG

study by Oganian & Chang (2018) found that the correlation between neural oscillations and the speech envelope was the highest at the latency of the peak derivative of the envelope, and suggested that this represents a point of maximal rate of change in the acoustic properties of a syllable. Because this is generally closer to the syllabic onset, they also suggest that the latter could represent one of the main landmarks for speech entrainment.

In the experiment described in Chapter 2 of this thesis, the methods (stimulus durations and amount of repetition) and analysis (phase and power coherence between neural oscillations, or between neural oscillations and speech) are similar to the ones employed by Doelling et al. (2014). The aim of this experiment was to replicate their results with natural speech stimuli. In the second and third studies, we used isochronous stimuli, similarly to Ding et al. (2015). Lastly, in the third experiment, we aimed to identify potential landmarks for speech tracking, by introducing noise at different syllabic locations, to determine how this would affect neural phase locking to isochronous syllables.

2. Acoustic edges and neural entrainment to continuous speech

Introduction

Populations of cells in the brain oscillate; these oscillations are usually categorised into frequency bands: delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-25 Hz) and gamma (25-50 Hz). Oscillations in these bands are thought to have distinct functional roles. The neural processing of auditory and speech stimuli seems to be mainly reflected by delta, theta and gamma oscillations, depending on the level of information: theta and gamma reflect low-level, acoustic properties (Poeppel, 2003), whilst delta rhythms may be elicited by higher-order linguistic units (Ding et al., 2015).

Neural oscillations are believed to track rhythmic stimuli through a process called neural entrainment. When a stimulus is periodic, brainwaves show greater amplitude at frequencies corresponding to the presentation frequency of the stimulus, and the phase difference between the recorded neural oscillations and the target stimulus is more consistent than that to an irrelevant stimulus (e.g., Thut et al., 2011). Whereas entrainment strictly refers to the temporal alignment of endogenous oscillations and the exogenous stimulus, phase locking to the stimulus also occurs when neural activity is evoked in the presence of a rhythmic stimulus. One such classic example is the auditory steady-state response (ASSR), an evoked potential which arises in response to a series of periodic auditory clicks (Stapells et al., 1987). EEG data collected from ASSR experiments show peaks in the amplitude of neural oscillations occurring at the same frequency and in time with the presentation frequency of the stimuli. Neural phase locking to rhythmic stimuli is present in a range of modalities such as motor (e.g., Nozaradan, Zerouali, Peretz and Mouraux,

2015), visual (e.g., Herrman, 2001) and auditory (e.g. Stapells, Makeig and Galambos, 1987).

It remains unclear whether neural entrainment occurs during speech tracking and its interpretation remains problematic because it is influenced by both temporal irregularities of the stimulus and its cognitive complexities. While the ASSR is a phenomenon observed in the presence of perfectly periodical stimuli, the auditory components of speech, such as syllables or phonemes, all vary in duration. Nonetheless, both syllables and phonemes show temporal regularities (i.e., their durations fall within fixed ranges), implying that speech is a quasi-periodic stimulus (Rosen, Carlyon, Darwin, & Russell, 1992) and moreover, these regularities, especially in the syllabic domain, seem to be consistent across languages (Ding et al., 2017). Neural oscillations show consistent phase locking to the rhythms of these speech units, which has been observed in across multiple languages (Ding & Simon, 2014). Neural phase locking to speech sounds has also been found in animals, at the local field potentials of cells in the primary auditory cortex. In one study by Steinschneider, Nourski and Fishman (2013), monkeys showed consistent phase patterns to both the syllabic as well as the phonetic rhythms of human speech.

However, entrainment to speech may not be necessarily be accompanied by increased evoked activity to the stimulus. Luo and Poeppel (2007) showed that, whereas phase coherence was increased amongst responses to the same trials and additionally, phase patterns could be relied on to discriminate between intelligible stimuli, there was no increase in the amplitude of the oscillations in response to stimulation. Thus, phase coherence is considered to be the primary mechanism needed for speech tracking and comprehension, especially for the slow (4-8 Hz) speech envelope fluctuations.

It is believed that oscillatory mechanisms help the brain track regular or semi-regular stimuli, such as speech, by operating within discrete windows of activity which correspond to patterns in the incoming information (Arnal & Giraud, 2012). In fact, Luo and Poeppel (2012) claimed that the brain uses two temporal windows which allow speech tracking, by parsing the acoustic signal. They presented participants with stimuli with temporal regularities of 25, 80 and 200 ms, which corresponded to the theta, alpha, and gamma frequencies of neural oscillations, and showed that the stimuli elicited activity in the acoustic cortex only in the theta and gamma bands, but not in the alpha rhythm. However, speech stimuli triggered less consistent responses in the gamma range than they did in theta. Researchers suggest that this may be because the syllabic rhythm may be the primary rhythm involved in speech perception, to which neural oscillations in other frequency ranges couple in order to extract information relevant at their respective temporal level (Giraud and Poeppel, 2012).

Indeed, it seems that the slow oscillations of speech, or the frequency content between 2 and 10 Hz which is represented by its envelope (i.e., the contour of an oscillatory signal around its extremities) are fundamental for speech processing (Kubanek et al., 2013). ECOG research showed that neural oscillations are mostly correlated with the envelope of speech, not its higher frequency information (Nourski et al., 2009), and retaining envelope information is enough to trigger comprehension, as shown both by studies on subjects wearing cochlear implants (Rosen et al., 1992) and those with normal hearing (Shannon et al., 1995). However, the quality and strength of neural phase locking to the speech envelope in humans seem to depend on both high-level functions such as intelligibility, comprehension and attention, as

well as low level factors such as timing and the fine grain structure of the stimuli (Giraud and Poeppel, 2012; Ding and Simon, 2014).

Before considering top down influences on neural speech tracking, note that the 2-10 Hz level, corresponding to the syllabic rhythm, covers frequency ranges outside of theta at the neural level. It is possible that the limits between these various types of levels are not so strict. Amongst these, the theta range (4-8 Hz) is the most robust in terms of speech entrainment, and the one that seems to mainly respond to acoustic features (Ding & Simon, 2014). However, entrainment is lower in the theta range when speech is unattended (e.g., Kerlin et al., 2010), unintelligible or noisy (e.g., Peelle et al., 2013), or when participants listen to stimuli in a foreign language (Peréz, Carreiras, Dowens and Duñabieta, 2014). High level properties of speech such as semantic, syntactic and lexical features also affect comprehension, but they do not seem to make a difference to phase locking in the theta rhythm (Ding et al., 2015). Speech-like stimuli can still trigger phase locking between 4 and 7 Hz, even when intelligibility is affected. For example, Howard and Poeppel (2010) showed that phase coherence was still robust even when they played speech stimuli backwards, in the absence of any distinguishable linguistic structures.

Neural phase locking in the lower frequencies of speech (under 4 Hz, which at the neural level correspond to the delta rhythm) is strongly affected by top down influences, i.e., in general, it is not noticeable unless the stimulus is highly intelligible and the language is comprehended (Ding et al., 2015). It is believed that the delta rhythm is largely involved in the processing of higher syntactic units such as phrases and sentences (Ding et al., 2015), but research also suggests that this is influenced by an irregular speech rate, or by pauses introduced in speech (Kayser, Ince, Gross and Kayser, 2015), as well as prosodic information contained, for example, in

stressed syllables (Greenberg et al., 2003). Therefore, phase locking to speech in the delta range seems to arise as a consequence of both acoustic (prosodic) and top down elements (comprehended phrases or sentences).

The influence of fine-grained properties of speech on 'entrainment' is less clear. These are represented by the high frequencies in speech, generally between 600 Hz and 10 kHz, which are generally considered to correspond to acoustic-phonemic information (Rosen et al., 1992). While in general, speech is reported more unintelligible when the low frequency modulations are removed from the sounds, rather than when the fine structure alone is altered (e.g., Apoux and Bacon, 2008), there is evidence that the fine structure of speech is important to entrainment. In a psychophysical study, Zoefel and VanRullen (2015) found that when presented with speech-noise constructs with no fluctuations in sound amplitude and spectral power, but which present rhythmic variations in phonemic content (i.e., fine spectral information), participants were still able to detect clicks inserted at syllable onsets, but not clicks which were present at other syllabic locations (nucleus or coda). Moreover, it seems that despite obvious amplitude modulations, such stimuli were still able to elicit phase locking in the theta range.

It has been suggested that fine grained properties located at the onset of syllables could help phase resetting by providing an 'acoustic edge' which acts as a landmark during tracking. In a MEG study, Doelling et. al (2014) showed that noise snippets at the beginning of each syllable also led to sharp changes in the temporal fluctuations, subsequently aiding both intelligibility and entrainment. Furthermore, phase coherence increased as a result, even when envelope fluctuations were missing. This is a possible indication that syllable onsets may carry information which is crucial to the neural tracking of speech.

The fine structure conveys information about the different properties of speech, such as the formants of vowels, or the manner of articulation of consonants (Rosen Stuart et al., 1992). Furthermore, consonants vary in their degree of spectral energy fluctuations (Blumstein and Stevens, 1979), and affect both the formant transitions of the vowels (Stevens, 2000) the 'rise time' of the syllables (Goswami & Leong, 2013). These consequently affect the shape of the envelope and, possibly, the amount of 'sharpness' in the speech signal.

Consonant-related fluctuations are provided by number of articulation factors, including the manner of articulation, or the way in which the air is released through strictures of the oral tract. Depending on the extent of stricture, consonants can be divided into stops ([b], [p], [g], [k], [d], [t]), fricatives (e.g, [f], [v]), sibilants (e.g., [s], [z]), rhotics (e.g., [l], [r]), etc. At a neural level, discrete regions of the auditory cortex were found to respond to different groups of consonants, which were mainly categorized based on the manner of articulation (Mesgarani, Cheung, Johnson and Chang, 2015). Amongst these, it was the stop consonants which form the most well-defined cluster, as shown by ECOG data (Mesgarani et. al, 2015). It may be possible that the sharpness of the envelope may be one of the factors affecting consonant clustering. For example, the sudden release of spectral energy following the pronunciation of stops makes these appear 'sharper' than other consonants, which, in the brain, could lead to a tighter cluster involved in the processing of such consonants.

In addition, it was found that phonemic information seems to affect neural speech processing in both the delta and theta ranges: using temporal-response functions to create representations between the EEG activity and various properties of a continuous speech stimulus (i.e., envelope or phonemic information, or a

combination of these), Di Liberto et al. (2015) found correlations between delta and theta neural responses and speech models which included phonemic information. Specifically, these correlations were highest when temporal-response functions were modelled as a combination of the narrowband envelope of speech and phonemic properties.

In the present experiment, we investigated whether the consonant content present in the beginning of syllables would affect neural envelope tracking of continuous speech. We created sentences whose syllables began with different groups of consonants (i.e., plosives vs. other consonants), for which we calculated the degree of 'sharpness', defined in the same manner as Doelling et al (2014). We aimed to determine whether sentences would differ between each other in terms of "sharpness", based on differences concerning syllable-initial consonants, and whether these differences would reflect in the level of neural phase locking to separate "sharpness" conditions. We predicted that sentences containing higher amounts of stop consonants would be sharper than the ones in the other group and therefore elicit more phase locking, especially in the theta range, following from findings by Doelling et al (2014). Because we expected to see entrainment to the speech stimuli, we did not predict any rises in the power of the oscillations, or any detectable amplitude changes due to evoked activity, following from results published by Luo and Poppel (2007). We also manipulated the effects of intelligibility on entrainment, by using stimuli which were spoken in languages both native and foreign to the participants. Following from previous findings (Peréz et al., 2014), we expected entrainment to be stronger to stimuli spoken in participants' native language.

Methods

Subjects

Subjects were 25 adults (10 females, mean age = 27.04 years old, standard deviation = 4.56 years) who were paid £10/hour to participate in the experiment. All participants were native English speakers with minimal knowledge of Russian (e.g., Russian studied at a beginner's level, 10 years prior to the study, but no later, was considered fine). Eligibility requirements were normal hearing, right-handed and no known history of neuropsychological conditions with an emphasis on learning disabilities. Six subjects were excluded because of noisy data (i.e., too many artifacts) or poor behavioural results, leaving us with 19 subjects for the final analyses which are reported here.

Design

Our aim was to investigate the effects of edges in natural language and whether these are influenced by intelligibility. We manipulated intelligibility by having two language conditions, one native (English) and one unknown to the participants (Russian). The study followed a 2 (language) x 3 (edge) factorial design. In addition to the strong and weak edge conditions, we had mixed or filler stimuli (syllable onsets were defined by a variety of sounds, see below), which acted as a control condition.

Stimuli

Stimuli were 100 sentences spoken by a bilingual English-Russian speaker and recorded in a soundproof room using Cool Edit Pro software (Adobe Systems Inc.). The sentences were initially reviewed by 10 native English speakers who were naïve

to the study, and given plausibility and grammatical correctness scores, on a scale from one to 10. We kept the sentences which had scores between seven and eight (no score was higher than 8.5). For each language, we used 20 strong sentences which we paired with weak sentences in the sense that they had the same number of syllables, syntactic structure, intonation and stress. We also added a similar filler sentence to 10 of these pairs. Each experimental sentence was repeated four times, and a filler was repeated three times. Some examples of the stimuli can be seen below (a full list is provided in Appendix 2).

English	Peg-gy dic-ta-ted pa-pers to ty-pists.	strong
	La-rry fal-si-fied lea-ses for fe-lons.	weak
	Ba-rry im-por-ted chi-cken from trades-men.	filler
Russian	Ba-ba ka-tit ba-gazh po tro-pin-ke.	strong
	The woman is strolling luggage along the path.	
	Ljo-va lo-vit vo-ron za za-li-vom.	weak
	Leva is catching crows behind the bay.	
	Pe-tja vy-nul lis-ty iz kar-ma-na.	filler
	Petja has taken out papers from the pocket.	

We included a cough in one of the filler presentations, which always appeared 1.6 seconds after the stimulus onset and lasted 400 ms. In total there were 410 stimuli, including repetitions. For each condition, the average sentence durations were: ‘English Strong’, 2.46 seconds; ‘English Weak’, 2.6 seconds; ‘Russian Strong’, 2.52 seconds; ‘Russian Weak’, 2.57 seconds. The number of syllables in each sentence varied between 8 and 11. The mean syllable duration was 242.23 ms. All stimuli were normalised to 70 dB loudness and 100 ms of silcence was added before the onset of each stimulus, using open-source Praat software (Boersma and Weenik, 2015).

Apparatus

We used 32-channel Brain Products EEG caps (BrainProducts Ltd.) to conduct scalp activity (see channel names and configuration in Figure 2.1). The setup required inserting Electrolyte gel through indentations in the electrodes and onto the scalp, to increase conductivity. The stimuli were delivered using Presentation software (Neurobehavioural Systems, Inc.) and through a pair of Sony Stereo headphones (model MDR-XD100, Sony Europe Ltd.) placed comfortably on the participants' heads, onto the EEG cap. EEG activity was recorded at 1KHz sampling rate using actiCap equipment and Vision Recorder software (BrainProducts Ltd.).

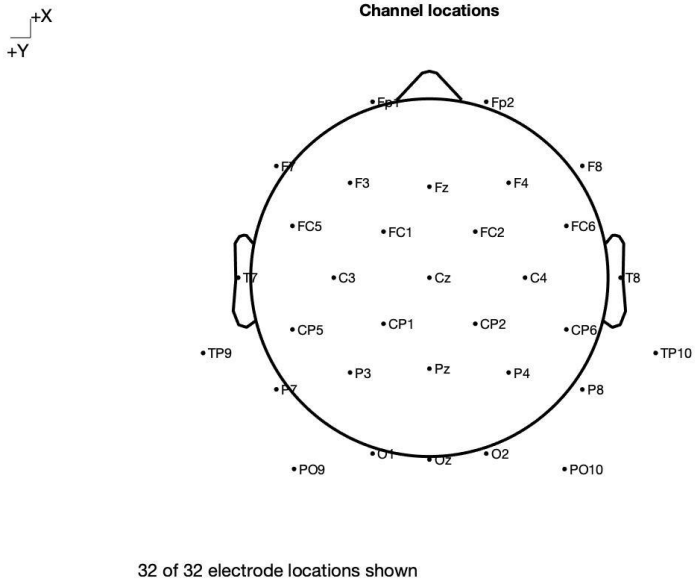


Figure 2. 1. EEG channel names and scalp configuration (read top to bottom, left to right). Plotted using EEGLAB.

Procedure

The experiment lasted approximately 1.5 hours including setup (i.e., gelling session, 40 minutes). The experimental session was 35 minutes long. Before the experiment, participants adjusted the sound volume to the level they found comfortable. During stimulus presentation, participants were asked to look at a cross presented on the screen, which prevented them from closing their eyes, and not move or blink during

the delivery of the stimulus. They were required to press the ‘Space’ bar as soon as they heard a cough. Participants were removed from the study if they correctly identified the cough on less than 50% of the trials.

The presentation of the stimuli was divided into five blocks of trials, each containing 82 stimuli. The inter-stimulus interval was 2 seconds. There was a 30 second break between each block, but participants could take longer if they wished, to minimise fatigue.

Data Analysis

EEG

All data analyses (both the acoustic analysis of the stimuli and the EEG analysis) were conducted in Matlab R2016b (Mathworks Inc.). Pre-processing of EEG data was done using EEGLAB toolbox (Delorme and Makeig, 2004). Data were low-pass filtered at 50 Hz, re-referenced using average reference, and then visually inspected for paroxistic artifacts. An Independent Component Analysis (ICA) decomposition was used to detect blinks, eye movements and ECG activity. ICA components were removed if in their topography, the power at frontal electrodes was 12 times higher than in the rest of electrodes. The EEG data were then split into 2000 ms epochs, of which the first 500 ms were removed, to avoid possible interference from event-related potential (ERP) activity, such as auditory evoked potentials (N1, P2, P3), which typically occur between 50-300 ms after the onset of sounds, but may also vary slightly outside of this range (Michalewski et al., 1986; Sur & Sinha, 2009).

Further time-frequency analyses were done using custom Matlab scripts, for data recorded at all 32 channels. These were done for each language, but only for the ‘weak’ and ‘strong’ conditions and not the ‘filler’ condition. Our analysis

resembles the one employed by (Luo & Poeppel, 2007). First, we took the spectrogram of the EEG signal and that of each sound file. We computed spectrograms in 100 ms steps for frequencies between 1 and 40 Hz, with a frequency resolution of .5 Hz below 10 Hz and 1 Hz above. We used these in our calculations for the phase coherence between the EEG signal and that of the sound envelope, or the Cerebro-acoustic coherence, as well as that between individual trials, or the Inter-trial phase coherence, using equation 2.1:

$$Cphase_{ij} = \left(\frac{\sum_N \cos(\theta_{ij})}{N}\right)^2 + \left(\frac{\sum_N \sin(\theta_{ij})}{N}\right)^2 \quad (2.1),$$

where θ_{ij} is the phase difference between two trials (i and j) of the same stimulus, or the phase difference between the the EEG and its corresponding stimulus, and N is the number of time points of one trial.

We measured the power covariance between different EEG trials only, and not between the EEG signal and the sound, like in equation 2.2:

$$Cpower_{ij} = \frac{\sqrt{\frac{\sum\left(A_{ij}^2 - \overline{A}_{ij}^2\right)^2}{N}}}{\overline{A}_{ij}^2} \quad (2.2),$$

Where A_{ij} is the magnitude of the product of the spectrograms of trials i and j, $\overline{A}_{ij}$ is the magnitude averaged over time, i, j are different trials, and N is the length of time. The power and phase coherence take values between 0 and 1, with large values indicating a high amount of phase coherence, but a small amount of power coherence (Luo & Poeppel, 2007). We computed the phase coherence between EEG trials and their corresponding sound envelopes (which were downsampled to 1000 Hz to match the frequency resolution of the EEG trials), and then, 100 times, to random sounds. The phase and power coherence between random trials were also

computed 100 times. We then subtracted the coherence to random sounds from that to actual sounds, and to random trials from that to corresponding trials, which gave us measures of de-noised coherence. Like Luo and Poeppel (2007), we named these phase and power dissimilarity functions, as given by equations 2.3 and 2.4:

$$Phase - dissimilarity = \frac{\sum_{j=1}^{J} Cphase_{intra}}{J} - \frac{\sum_{j=1}^{J} Cphase_{inter}}{J} \quad (2.3),$$

$$Power - dissimilarity = \frac{\sum_{j=1}^{J} Cpower_{inter}}{J} - \frac{\sum_{j=1}^{J} Cpower_{intra}}{J} \quad (2.4),$$

Where J is the number of trials.

Positive differences in the dissimilarity functions indicated that the coherence was higher between trials to the same stimulus, and negative differences suggested that the coherence was higher between trials to different stimuli. A positive difference would thus be more indicative of successful tracking.

The phase and power dissimilarity functions were also applied to the baseline of each trial, for the purpose of comparing neural activity during rest with that during experimental stimuli. A baseline was extracted for the duration of 1500 ms prior to each trigger, after which a spectrogram was applied in the same fashion as for experimental trials. The intra-trial phase and power coherence were taken between baselines to the same stimuli, and the inter-trial measures, between baselines to different stimuli. We did not expect these to differ between each other, but, if entrainment to our stimuli was present, we thought it might be possible for this to be present in the inter-stimulus interval, and therefore show in the time-frequency analyses of the baseline activity. Therefore, the phase and power coherence differences were also taken for the baselines, further allowing for a direct comparison between EEG responses and neural oscillations at rest.

Sounds

To compute sharpness, we used a cochlear filter, in the same manner as Doelling et al. (2014). Each stimulus was filtered into 32 frequency bands, between 80 and 8000 Hz. The envelope of each frequency band was taken as the absolute value of the Hilbert transform, after which we computed the sum of the narrowband envelopes. Sharpness was calculated as the sum of the positive first derivative of the summed envelope. As a separate measure, we also normalised this to the total amplitude of the envelope. We argued that, because our stimuli were natural, they may not differ in the total amount of sharpness, as the artificial stimuli used by Doelling et al. (2014) did, but they would differ in the amount of normalised sharpness.

Results

Stimuli

We first investigated whether the stimuli differed in their amount of sharpness depending on condition and whether our 'strong' stimuli indeed scored higher in their amount of sharpness, relative to the 'weak' ones. In terms of sharpness as defined by Doelling et al. (2014), there was no difference between conditions. This may be due to the fact that, while Doelling et al. (2014) used highly artificial stimuli, with highly distinct envelopes across conditions, we used natural ones. Doelling's sharpness values may have not only been influenced by the steepness of the syllabic slopes, but also by the maximum amplitude of the syllabic peaks, which were consistently higher in 'sharper' conditions. On the other hand, natural stimuli may have more similar syllabic peaks across sentences which are spoken at a similar overall loudness, and with the amplitude of the peaks being most likely affected by stress and intonation, more so than by the constitutive phonemes of the syllables.

Therefore, Doelling's sharpness may not necessarily reflect the properties of syllabic slopes in a natural context, but one that is normalised to the envelope's overall amplitude might, as this would take into consideration similarities in sound intensity across sentences.

A two-way ANOVA conducted on the normalised sharpness values of our stimuli indicated that there was an effect of edge ($F_{1,20}$ = 62.26, p<.001), but no effect of language, and no significant interaction between the two factors. Subsequent t-tests revealed that 'strong' stimuli indeed had higher normalised sharpness values than the 'weak' ones, in both English (t(19) = 7.88, p<.001), and Russian (t(19) = 9.65, p<.001), but that there was no difference between the two languages. The differences in normalised sharpness may be due to the fact that the total amplitude of the envelopes of weak stimuli was higher than those of strong stimuli (t(19) = -2.93, p<.01).The differences in normalised sharpness between all conditions can be seen in Figure 2.2.

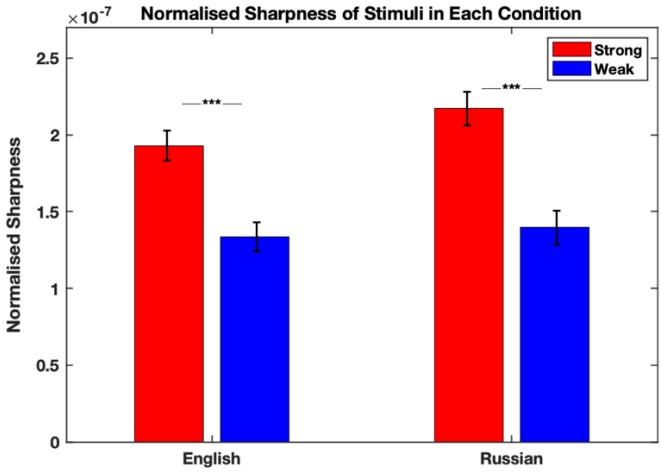


Figure 2. 2. Sharpness values are plotted as a function of language. Red bars indicate 'strong edge' conditions and blue bars indicate 'weak edge' conditions, for each language. Error bars indicate ± standard error of the mean. Stars represent significant differences between conditions joined by horizontal lines: three stars correspond to p<.001.

Figure 2.3 shows an example of a pair of weak and strong sentences in English. The number of syllables generally corresponds to the number of peaks in the envelope. One can see that the while the 'weak edge' sentence seems to contain one peak which is higher than all of the peaks in the 'strong' condition, the 'strong

edge' sentence seems to have more high peaks, which are also higher than the 'weak edge' sentence's second highest peak. The syllables in the 'strong edge' sentence also seem narrower, implying that they have steeper slopes, and are therefore sharper.

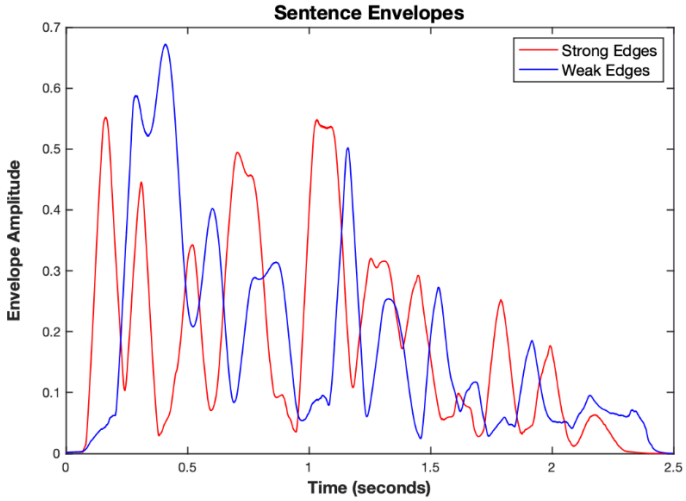


Figure 2. 3. Sound envelopes of a pair of weak ('Sarah reviewed seven Physics lessions') and strong ('Bobby compared bargain data packets') sentences, in the English condition. The amplitude of the envelope is plotted as a function of time. Peaks above 0.15 amplitude, which follow very low troughs in the envelope (i.e., lower than 0.2 amplitude) are likely to indicate separate syllables.

EEG

To follow from results by Doelling et al (2014), we investigated the phase coherence between the EEG signal and the sound envelope, or the cerebro-acoustic coherence. Specifically, we first took the coherence of the EEG response and the spectrogram of the stimulus which corresponded to the EEG trial in question. Subsequently, we computed the phase coherence between the EEG and a random stimulus, and subtracted the latter from the former. This method is also reported in Peelle, Gross, & Davis (2012). Figure 2.4 illustrates spectrograms of an EEG trial and its corresponding sound envelope, which both show increased power at frequencies below 15 Hz. Even though slight inflections, both positive and negative, can be noticed at frequencies below 10 Hz, and in the vicinity of 20 Hz, the cerebro-acoustic coherence difference did not show significant peaks at any of the

frequencies of interest, it was not different from baseline, and there were no significant differences between conditions.

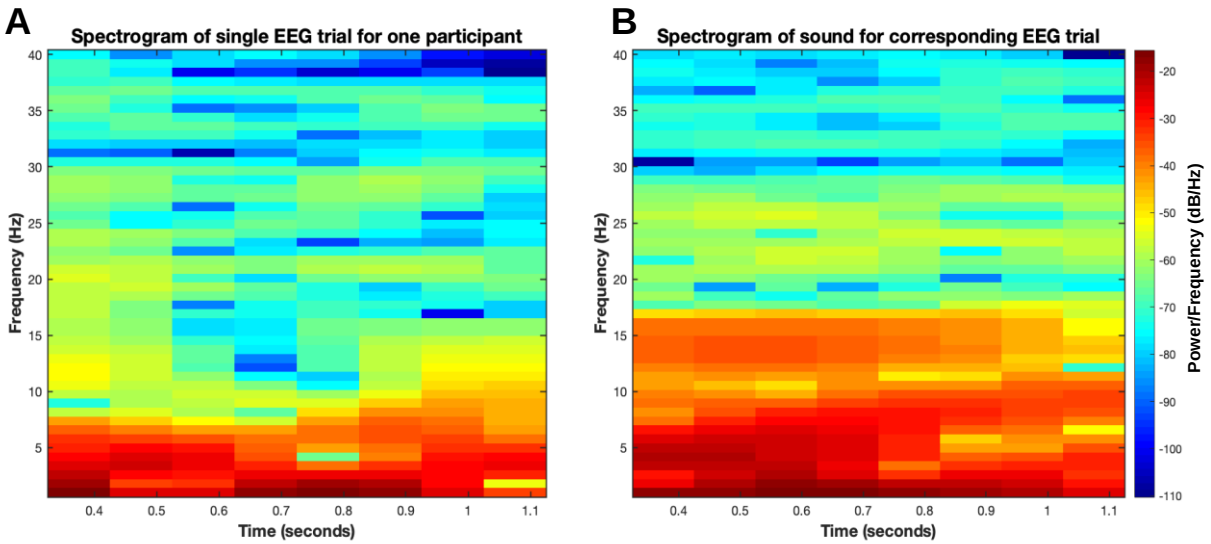


Figure 2. 4**.** A. Spectrograms of an EEG trial for one participant (subject number 2). B. Spectrogram of the stimulus envelope which was attended during the corresponding EEG recording (an 'english strong' sentence). Both spectrograms show increased power in the lower frequencies (below 15 Hz), with the EEG spectrogram showing more power at frequencies below 10 Hz. The time vector is plotted at latencies between 0.375 and 1.125 seconds, with a resolution of 100 ms.

The power and phase coherence differences, calculated between the same or different EEG trials, also elicited little significance. Unlike the cerebro-acoustic coherence, the phase coherence difference to EEG trials shows several peaks in the low-frequencies, specifically, between 2 and 12 Hz, as can be seen in Figure 2.5.B. The difference is also positive, implying that the phases were more consistent between responses to the same stimulus than between those to different stimuli, which suggests some degree of stimulus tracking, possibly to the syllabic rhythm. In the following analyses, we compared the phase coherence difference across three frequency ranges: delta (1-4 Hz), theta (4-8 Hz) and alpha (8-12 Hz). The phase coherence difference averaged over conditions was significantly higher than baseline activity, in the delta ($t(18) = 2.51$, $p<.05$), theta ($t(18) = 3.98$, $p<.001$) and alpha frequency ranges ($t(18) = 2.58$, $p<.02$). Figure 2.6 depicts these differences, showing

that they were most significant in the theta range, which is consistent with previous research emphasising syllabic tracking.

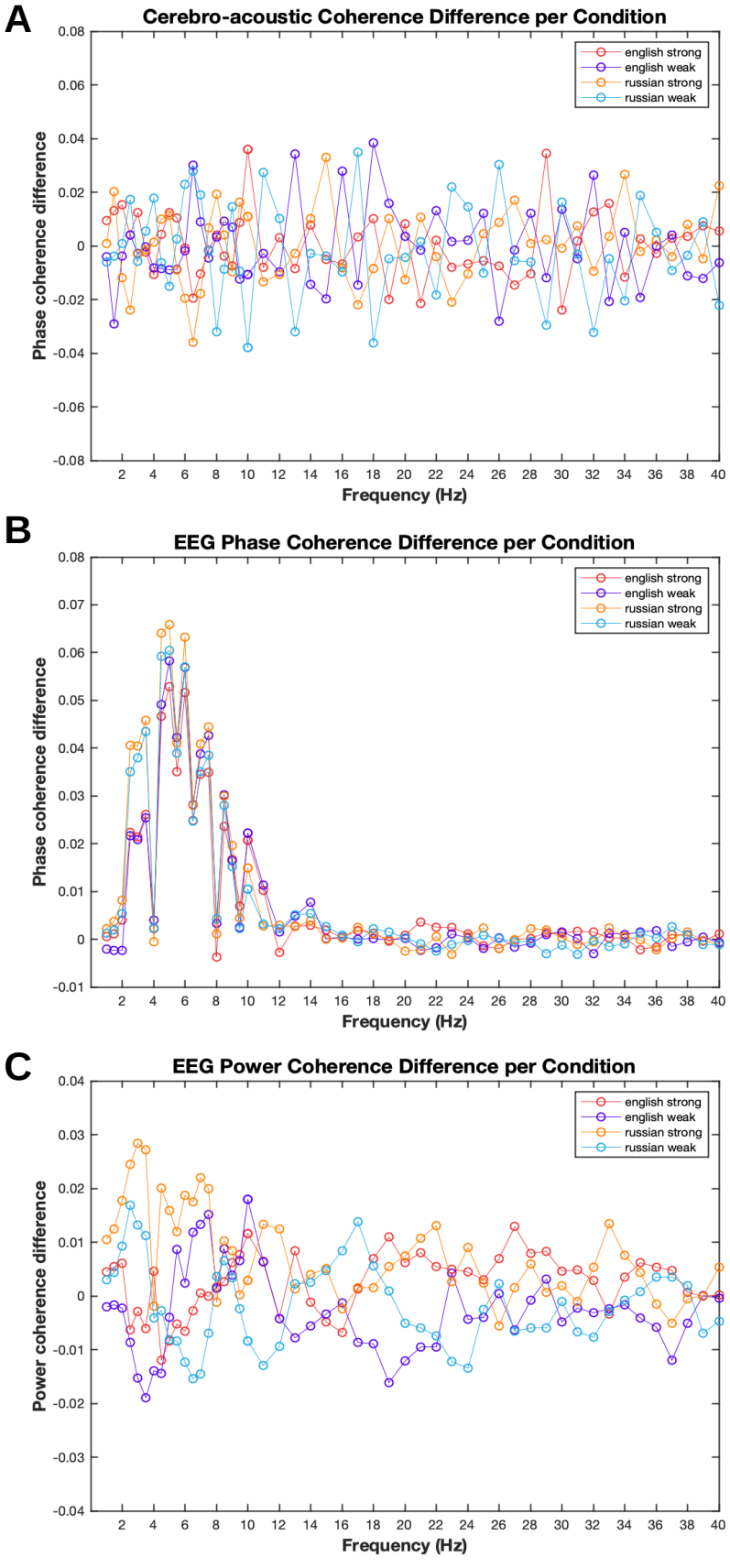


Figure 2. 5. Phase and power coherence differences, between EEG trials to the same sound and EEG trials to different sounds, are plotted for each condition. A. Cerebro-acoustic coherence difference, calculated as the phase coherence between EEG responses and stimulus envelopes. There are no significant peaks in any of the frequency ranges B. Phase coherence difference between EEG trials, indicating the consistency of the phase difference of the oscillations of the same trials versus that of random trials. This shows peaks in the frequency range 1-10 Hz, which are significantly greater than the values above 10 Hz. The phase coherence difference between 1-10 Hz is also higher than baseline phase coherence difference. C. Power coherence difference between EEG trials, indicating the magnitude of the responses to the same stimuli versus that of responses to random stimuli. Just like in **A**, no significant peaks can be found here.

When considering conditions separately, only the ‘Russian Weak’ condition elicited significantly more phase coherence than baseline in the delta range ($t(18) = 2.31$, $p<.05$), while the ‘English Strong’, ‘Russian Strong’, and ‘Russian Weak’ conditions had higher coherence than baseline in the theta range ($t(18) = 2.63$, $p<.05$; $t(18) = 3.04$, $p<.01$; $t(18) = 3.7$, $p<.01$, respectively). In the alpha range, only the ‘strong edge’ conditions elicited more phase coherence compared to baseline (‘English Strong’, $t(18) = 2.3$, $p<.05$; ‘Russian Strong’, $t(18) = 2.44$, $p<.05$). However, when the phase coherence difference was compared between conditions, there were no language or edge effects, and post-hocs did not show any significant differences between individual conditions, in any of the investigated frequency ranges (see Figure 2.7). Subsequent t-tests revealed that the phase coherence in the theta range was also significantly greater than delta ($t(18)=6.58$, $p<.001$, Bonferroni-corrected) and alpha ($t(18) = 11.76$, $p<.001$, Bonferroni-corrected), but that there was no difference between phase coherence in delta to that in the alpha range.

Unlike the phase coherence, the power coherence difference is not always positive (see Figure 2.5.C). Also, while it seems that its values fluctuate the most in the low frequency range, the power coherence difference is never significantly distinct from baseline power (Figure 2.8), either when averaged or when comparisons are made for individual conditions. The conditions are also not significantly different from each other (Figure 2.9), and there are no differences between the different frequency ranges in terms of power coherence. These results are in line with findings from Luo & Poeppel (2007), in the sense that we obtained higher phase, but not power, coherence in the low frequencies. Furthermore, while these results spanned different ranges of frequencies (delta, theta, alpha), comparisons to baseline indicate that the

theta range was somewhat preferred for phase entrainment compared to the other two.

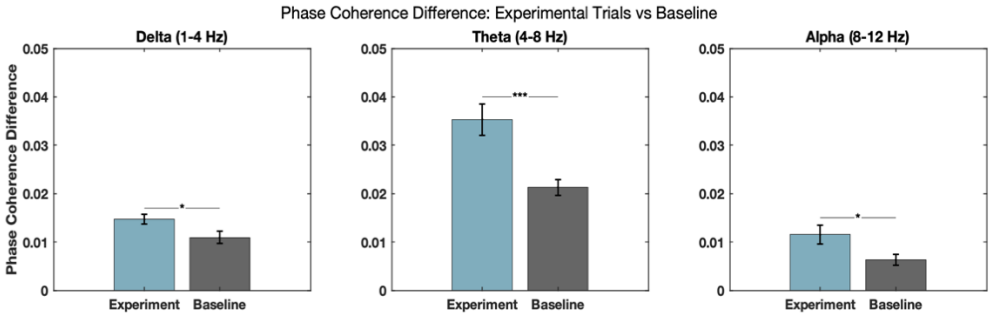


Figure 2. 6. Differences in phase coherence between baseline activity and experimental trials, at different frequency ranges. From left to right: differences in delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz). Error bars indicate ± standard error of the mean. Stars represent level of significance: *, p<.05. **, p<.001.

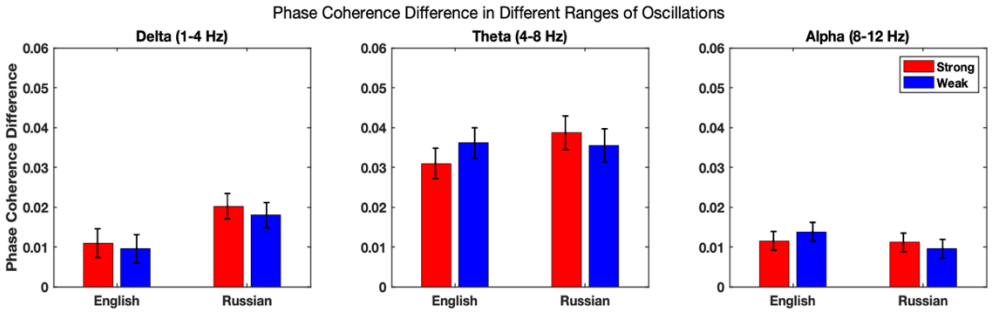


Figure 2. 7. Phase coherence difference to experimental trials, plotted per condition, at different frequency ranges. From left to right: delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz). Error bars indicate ± standard error of the mean. There are no differences between conditions at any of the three frequency ranges, but overall phase coherence difference in theta is greater than in the other two intervals.

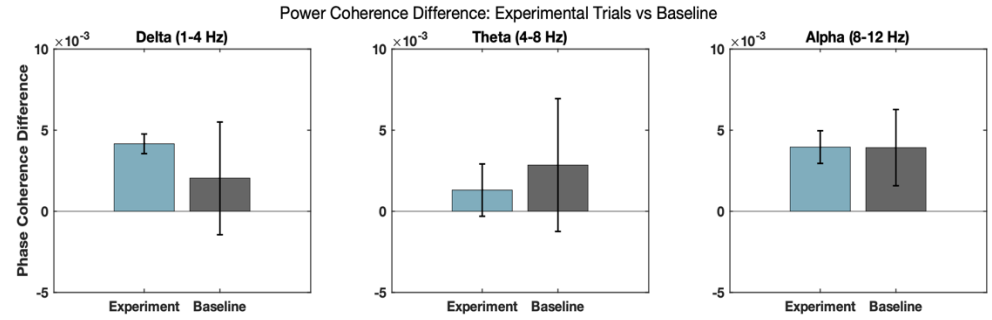


Figure 2. 8**.** The power coherence difference to experimental trials is plotted versus the power coherence difference in baseline activity, at different frequency ranges. From left to right: differences in delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz). Error bars indicate ± standard error of the mean. No significant differences were found at any frequency range.

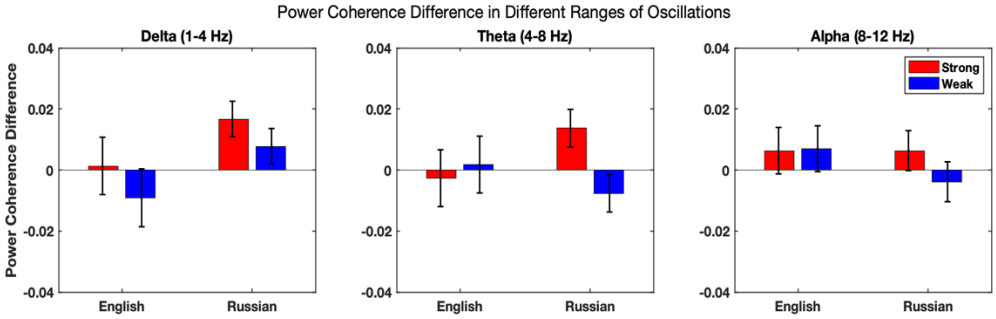


Figure 2. 9**.** Power coherence difference to experimental trials, plotted per condition, at different frequency ranges. From left to right: delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz). Error bars indicate ± standard error of the mean. There are no differences between conditions at any of the three frequency ranges, and no differences in the power coherence across the frequency ranges.

Discussion

An important finding of the present research is that the 'strong edge' stimuli indeed had greater values of normalised sharpness than the 'weak edge' stimuli, indicating that the syllable-initial consonants may affect sharpness and, specifically, that syllables starting with stop consonants were sharper than syllables starting with other consonants. However, we did not seem to find the expected effects of sharpness on entrainment, especially in the light of previous findings such as the ones from Doelling et al. (2014).

In line with other research, we found that the phase coherence between EEG trials was greater in the slow oscillations, with peaks showing between 2-10 Hz. Comprehension relying on a frequency range outside of theta (4-8 Hz) is not unusual (Drullman et al., 1994 a,b), so it possible that a larger interval of envelope fluctuations may have impacted phase locking in our study. This may be, in part, due to syllabic duration being both larger than 250 ms (4 Hz) and shorter than 125 ms (8 Hz). Indeed, in the English condition, the shortest syllable was 54 ms, and the longest syllable was 508 ms. However, most of our syllables were in the theta range: the median syllable duration was 220 ms, corresponding to 4.54 Hz. The

highest entrainment in our study as illustrated by the EEG phase coherence was between 4 and 8 Hz. These findings support previous research, which has shown that both amplitude fluctuations (1-10 Hz) in speech, as well as the phase coherence in response to sound envelopes, show peaks in this frequency range (for a review see Edwards & Chang, 2013).

Phase coherence in the delta, theta and gamma ranges, were all greater than baseline, with phase coherence in the theta range being higher than at other intervals. This is similar to findings from Luo and Poeppel (2007), who used the same calculations to show that entrainment to speech was strongest between 4 and 7 Hz, as well as a range of other studies emphasising the importance of theta oscillations in speech tracking (Ghitza, 2013; Ding & Simon, 2014). Furthermore, a slight effect of edge was noted in the phase coherence in the 8-12 Hz, with 'strong edge', but not 'weak edge', stimuli showing higher coherence than baseline. While it has not been suggested that speech entrainment is accompanied by an increase in the alpha rhythm (in fact, the opposite has been shown, see Weisz & Obleser, 2014), it is possible that peaks in this range may reflect entrainment to fluctuations corresponding to phonemes. Indeed, Drullman et al., (1994a,b) showed that removing frequencies between 8 and 16 Hz impacted the recognition of phonemes, especially the one of stop consonants. Therefore, it is possible that the high amount of stop consonants present in the 'strong' conditions led to more entrainment in the frequency range corresponding to their durations.

However, the peaks in the alpha range could also be explained by technical aspects, such as including all EEG channels in the analyses: it is known that alpha oscillations originate predominantly in the occipital region (e.g., Valera, Toro, John & Schwartz, 1981), which is not crucial for speech tracking. This could be avoided by

excluding occipital electrodes, limiting the analyses to a subset of channels corresponding to areas responsible for auditory processing or selecting the channels based on the highest auditory ERPs corresponding to each participant, using linear regression. Similar research reports using all of the channels in the analyses (e.g. EEG: Ding et al., 2017; MEG: Luo and Poeppel, 2007), but the studies showing significant findings from stimuli with variable syllable lengths generally use MEG and up to 128 sensors (e.g., Luo and Poeppel, 2007; Doelling et al., 2014). However, channel selection could be employed in future analyses.

An effect of language was noted in the delta range of the phase coherence, with Russian stimuli triggering more entrainment than the English ones. There was also a marginal effect of language in the power coherence, with Russian stimuli showing a positive phase coherence difference, and English stimuli, a negative one. This finding is counterintuitive with respect to intelligibility, as phase locking the delta range seems to often rely on comprehension (Ding et al., 2015; Molinaro & Lizarazu, 2018). However, our results may be explained by prosodic aspects, such as stress patterns. It has been shown that stress and prosody affect phase locking (Bourguignon et al., 2013). Even if English and Russian are both stress-timed languages, they differ in their stress patterns, or the properties which lead to stressed syllables. For example, when looking for stress cues in a native for foreign language, Russian speakers tend to look for intensity and duration patterns, whereas English and Mandarin speakers rely on pitch and vowel quality (Chrabaszcz, ,Winn, Lin, & Idsardi, 2014).

Despite a possible effect of language in the delta range, we did not find that the power coherence was stronger at any frequency interval, or at any point different from baseline activity. This is in line with previous findings (e.g., Luo and Poeppel,

2007) and suggests that speech tracking in the brain is not reflected by the amplitude of the neural oscillations, but only by the phase difference of oscillations in response to the same stimulus. This could also be considered as evidence for entrainment of endogenous oscillations (Zoefel et al., 2018).

However, the cerebro-acoustic coherence did not show any significant peaks. The reason for this finding remains unclear, especially given that both the stimuli and EEG trials showed strong representation of low frequencies when their spectrograms were inspected, but could be due to technical aspects such as the downsampling of the envelope when computing the cerebro-acoustic coherence. It is possible, however, that the phase differences between the speech waveform and the brainwaves were less consistent than those between brain oscillations recorded to the same stimuli. In fact, Doelling et al. (2014) report using a slightly different equation when calculating the cerebro-acoustic coherence. However, the reason that would give them higher speech-to-MEG phase coherence remains unclear.

Furthermore, unlike Doelling et al (2014), we used natural stimuli which did not differ between each other in the overall positive derivative values of the envelope, but only when these were normalised to the overall envelope. This may have affected the outcome of the experiment. On the other hand, the artificial stimuli used by Doelling et al. (2014) may have had sharpness values far beyond natural stimuli. The sharpness values were expressed by very distinct syllabic edges, therefore leading to better coherence between the MEG and the sound.

Another reason for the weak coupling in our findings may be due to the short durations of our stimuli and epochs: we only used 1.5 seconds of each trial in our analyses. Longer stimuli may have led to stronger entrainment and indeed, the majority of studies report using stimuli at least 4 seconds long (e.g., Luo and

Poeppel, 2007), with the exception of Doelling et al. (2014), who used 100 stimuli between two and three-seconds long, and repeated them four times, which is comparable to our study. However, their stimuli comprised of spoken digits, the majority of whom were monosyllabic words, whereas our stimuli were sentences with complex semantics as well as prosodic patterns.

The length of stimulation may be of particular importance given that it is possible that sometimes entrainment may occur later than the onset of the stimulus, with longer stimuli being more likely to lead to entrainment. This is suggested by research in animals, which has shown that sometimes, macaques or different species of parrots, respond to rhythmic stimuli after these have ceded (parrots: Hasegawa, Okanoya, Hasegawa, & Seki, 2011; macaques: Zarco, Merchant, Prado, & Mendez, 2009). This may be due to what some scientists have described as a 'build-up' of entrainment (Bee et al). According to one theory, the endogenous theta oscillations start shifting their phases to match attended speech stimuli, and consequently streaming out irrelevant stimuli with mismatching phases, but importantly, it has been shown that this process does not happen instantaneously, with stronger entrainment occurring after longer periods of attending auditory stimuli (Riecke et al., 2015).

Investigating the evoked potentials present in the data may also have been helpful in clarifying the underlying neural processes which arose as a results of this experiment. In the future, the analysis of studies using a similar approach could involve computing the STRFs of the neural responses by detrending ERP data using a linear model, similar to Di Liberto and Lalor (2017) and Di Liberto et al. (2015). These could be applied to the EEG response to the start and end of each word, then averaged across conditions, checking for features which may be specific to each

condition (strong, weak, native or foreign language). Unfortunately, this is a laborious process which was difficult to implement now due to the lack of event markers within EEG epochs, but which could be easily solved in a new experiment.

Another potential issue which can explain the lack of difference between conditions is our sentences were either somewhat implausible, or in a foreign language. Both of these factors may have affected intelligibility and in consequence, a lower level of intelligibility may have led to less stimulus tracking. This would be consistent with previous research, which has shown reduced entrainment to stimuli which were not easily comprehended (e.g., due to high background noise: Ding, Chatterjee, & Simon, 2014; due to degraded acoustic content: Zoefel & VanRullen, 2015). The unusual English stimuli could also explain why there was no difference between the two language conditions, despite evidence that listening to native speech leads to more neural entrainment than foreign speech (Pérez et. al, 2015). Unfortunately, we were restricted in our choice of words by the fact that each syllable could only start with a limited number of consonants. However, a smaller number of sentences, which were also longer, and repeated more times, may have been able to elicit stronger coherence, especially between the EEG and the stimuli.

While it has been suggested that onsets may possibly be the preferred landmarks for entrainment, as mentioned by Doelling et al. (2014), and because restoring acoustic information at the beginning of the syllables may improve entrainment (Zoefel & VanRullen, 2015), it has not been established that this is the case. Furthermore, our stimuli not only comprised of consonant-vowel (CV), but, mainly, of consonant-vowel-consonant (CVC) syllables, and sometimes consonant-consonant-vowel (CCV), or consonant-consonant-vowel-

consonant (CCVC). However, it was always only the first consonant of any given syllable that varied across conditions. While it seems that this was enough to affect sharpness, it may have not been enough to elicit differences in entrainment across conditions.

Entrainment to speech can also be investigated using periodic stimuli. For example, Ding et al. (2015) used sentences formed of monosyllabic words of identical durations to show entrainment at phrase and sentence levels, whose frequency was not explicitly present in the stimuli. Furthermore, isochronous stimuli have also been found to show effects in the power or magnitude of entrainment (Will and Berg, 2007), compared to variable speech, because this leads to additional evoked activity, similar to the case of ASSR. By using periodic stimuli and investigating 'entrainment in the broad sense' (Obleser & Kayser, 2019), we may be able to see differences in entrainment due to sharpness or due to differences in syllable-initial consonants.

In summary, we showed that stop consonants placed at the onsets of syllables led to greater amounts of sharpness as measured by the normalised positive derivative of the envelopes of the stimuli, but not by the Doelling sharpness. The effects of sharpness were possibly reflected by the EEG phase coherence in the 8-12 Hz range, which was higher for 'strong edge' than for 'weak edge' stimuli, while showing no effect of language. However, this may reflect entrainment to phonemes and not the effect of different phonemes on phase locking to the syllabic rhythm. Nonetheless, the results from our experiment showed the characteristic entrainment to speech in the theta range, as indicated by the phase coherence between EEG trials. The lack of power coherence is in line with previous findings, possibly due to the mechanisms of entrainment such as build-up, which assumes that responses to

the same speech stimuli are not necessarily the same in magnitude. Build-up may also lead to an increase of entrainment over time, which could also explain the absence of phase locking in the cerebro-acoustic coherence. Furthermore, short, implausible stimuli, insufficient repetitions of each stimulus, and not using equipment powerful enough for the demands of the task may also explain the minimal effects of our manipulation. Reinvestigating the cerebro-acoustic coherence as well as auditory ERPs may also be able to clarify the present results. The next study addresses some these issues in order to be able to investigate more thoroughly the effects of sharpness and phonemic properties on entrainment.

3. Do phonemes affect neural entrainment to the syllabic rhythm?

Introduction

Previous research suggests that the brain is able to track speech through oscillatory mechanisms which act at discrete temporal windows of activity, in order to parse the incoming information (Arnal & Giraud, 2012). The duration of one such window, or oscillatory cycle, corresponds to that of speech syllables, which, at an acoustic level, are best represented by the slow temporal fluctuations in the envelope (2-5 Hz) (Edwards & Chang, 2013). Importantly, alterations to the syllabic rhythm of speech have been consistently found to lead to impairments in both phase locking and comprehension, more so than reductions in the frequency content outside the syllabic range (Ding & Simon, 2014).

Researchers have proposed that the syllabic rhythm is not only conveyed by the slow temporal fluctuations of the envelope, but perhaps by specific landmarks found within the syllables (Ghitza, 2013). For example, Zoefel and VanRullen (2015) found that when participants listened to stimuli without any apparent amplitude modulations in the envelope, they still showed robust phase locking in the theta range if the acoustic information located at the onsets of the syllables was preserved, but not the one at any other locations. Doelling et. al (2014) found similar results by playing modified speech waveforms where frequencies below 10 Hz were removed, and introducing clicks at the previous locations of the syllables. Moreover, they found that the amount of entrainment was correlated with the 'sharpness' of the envelope. They measured sharpness as the sum of the positive derivative of the envelope, which corresponds to the sum of all slopes pertaining to the rises in the envelope. Because envelope rises are mostly located at the onsets of syllables, they

suggested that the ‘acoustic edges’ provided by the onsets may be possible landmarks for entrainment. Nevertheless, Doelling et al. (2014) used artificial stimuli which had sharpness levels far beyond those of natural speech.

In Experiment 1, we investigated whether natural speech sentences with different levels of sharpness would lead to differences in neural entrainment, by manipulating sharpness through the nature of the syllable-initial consonant. We created sentences with stop consonants placed at the onset of syllables with stronger edges, and other consonants, specifically a selection of fricatives and liquids, at the onset of those with weaker edges. Subsequent measurements confirmed that indeed, sentences whose syllables started with stop consonants had higher sharpness than those using other consonants as syllabic onsets, but only when sharpness was normalised to the overall envelope. The EEG responses to the syllabic rhythm did not vary significantly between conditions, although the phase coherence was higher than baseline at frequencies between 1 and 12 Hz. An effect of sharpness showed between 8-12 Hz, where the phase coherence to sentences with ‘strong edges’ was higher than that to stimuli with ‘weak edges’. Nonetheless, the frequency range between 8 and 12 Hz is outside the syllabic rhythm, so these results cannot confirm that sharpness affected entrainment to syllables in our experiment.

However, our stimuli did not differ in their ‘Doelling sharpness’, but only in the amount of ‘normalised sharpness’. It is unclear whether this is because natural stimuli do not present any obvious fluctuations in sharpness. If this is the case, even if our results cannot discredit the influence of acoustic edges in neural speech tracking, we cannot be sure that the ‘Doelling sharpness’ is alone able to describe such edges. This is equally true for the ‘normalised sharpness’. One reason for this

may be because the positive derivative of the envelope which is used to calculate sharpness takes into account all rises in the envelope. Even if most of the rises are present between the onset and the peak of the envelope, the fact that sharpness takes into account all rising fluctuations of a syllable may less clearly indicate the presence of a particular landmark. For example, if onsets provide edge information more than other syllabic locations, it may be necessary to measure envelope properties which relate to the onset of the syllable alone. Furthermore, the onset of the syllable has not been established as the location containing the only necessary information for entrainment. In fact, some researchers emphasise the vowel locations as more important, because these contain envelope peaks (Ghitza, 2013), and almost no literature exists concerning the role of syllabic codas on entrainment.

It is possible that our selection of phonemes did not result in differences not only in the 'Doelling sharpness', but also in the quality of acoustic landmarks across conditions. In Experiment 1, we only broadly differentiated between our syllable-initial consonants based on the manner of articulation, whilst ignoring other acoustic features related to phonemes. Some of these features have distinct spectral characteristics which may play a role in identifying landmarks during neural speech tracking.

Stevens (2002) argues that speech is a continuous signal which can be split into different segments, which are stored in memory as discrete units. The segments can be distinguished based on a number of articulatory features with different acoustic correlates, and it is the acoustic difference in features across or within the same segments which leads to word identification. For example, he gives a more complete interpretation of phonemes by explaining that the first segment in 'bat' is different from the first one in 'pat' based on a single feature, which allows us to

distinguish not only between words, but also between the phonemes /b/ and /p/. We are asking whether the brain is able to distinguish between these features during neural entrainment to the syllabic rhythm of speech.

Acoustic features can be seen as peaks, valleys or discontinuities at certain frequencies in the sound spectrum (Stevens, 2002). For example, vowels contain higher intensities than consonants, which can be identified as peaks at the frequency of the first formant. On the other hand, consonantal segments can be identified as discontinuities in the acoustic signal, in the sense that the amplitude of low and mid-frequencies is lower in the spectrum of consonants than that of vowels. This happens either due to a constriction (complete or partial) or due to a narrowing in the oral tract.

The type of constriction is given by the manner of articulation (Hannah and Davenport, 1998). If the constriction of the or closure is complete, then the consonants are plosives or stop consonants. Nasals (/m/,/n/) are also stop consonants, because the air does not pass through the vocal tract, but through the nasal cavity. Nasal consonants are sonorants, while plosives are obstruents. The difference between obstruents and sonorants is that in the first case, the spectral discontinuity is caused by an abrupt change in the pressure of the air flow through the oral cavity, while sonorant-related discontinuities are due to a sudden change in the path of the airflow (through the oral or nasal cavities), and not to differences in air pressure. Formants can be seen in the spectrum of sonorants because the air passes unrestrained during their production, allowing for resonance to occur. Other sonorants include liquid consonants (e.g., /l/, /r/), where the air passes through narrow openings in the mouth, and vowels, where the oral tract is completely open. Consequently, the sonorant feature is not just restricted to consonants, which

suggests that there is a certain similarity between vowels and sonorant consonants. In Experiment 1, liquids were used at the onset of syllables in the 'weak edge' condition, but their similarity to vowels could indicate the presence of powerful edges which are necessary for neural speech tracking.

Fricatives are a group of consonants which are produced with partial closure, which causes a continuous turbulence in the air stream (Stevens, 2002). These consonants are also known as continuants, and include the phonemes /f/, /v/, /s/, /z/. The last two consonants are sibilants, or stridents. When paired with a vowel, strident consonants show a greater amplitude in the high frequencies of the spectrum than the neighbouring vowel. Even if they do not have the same overall intensity as the vowels, a larger amplitude in the higher frequencies of the spectrum could result in a peak in the speech envelope, at the latency at which a fricative is produced. While this may not lead to sharper envelopes, it could imply that fricatives also provide strong landmarks for neural entrainment.

Some acoustic features like voicing have direct effects on the adjacent vowel (House and Fairbanks,1952; Stevens, 2002). Voicing refers to the vibration of the vocal cords during articulation. If vibration occurs during production, the consonant is voiced, but if the vocal cords are stiff, then it is voiceless. The fundamental frequency at the beginning of the vowel is increased when this follows a voiced consonant, and is lowered when it follows a voiceless consonant. Furthermore, the voicing of the consonant also affects the neighbouring vowel's power and duration (House and Fairbanks, 1952). Both plosives and fricatives can be voiced or voiceless consonants.

The effects of consonants on adjacent vowels, as well as consonantal features which arise as a consequence of articulation, and have distinct acoustic

properties, can influence the speech envelope in different ways, or provide different, but nonetheless strong, edges. We did not fully explore the effects of different consonant and vowel features in Experiment 1. In the next experiment, we addressed these issues by making a series of fundamental changes to our manipulation. First, we used consonant-vowel or vowel-only syllables, which allowed us to directly test the effects of different consonants placed at the onset of different syllables. Second, we tested the effects of different consonants separately, by creating conditions which each contained only one consonant located in the beginning of the syllables. We investigated the same consonants as in the previous experiment (stops: /b/, /d/, /g/, /k/, /p/, /t/; fricatives: /f/, /v/; sibilants or strident fricatives: /s/, /z/; liquids: /l/, /r/), but also added nasal stops (/m/, /n/), in order to obtain a broader picture of how different phonemes impact entrainment. Lastly, the syllables in Experiment 2 were isochronous, i.e., they all had approximately the same durations. By using isochronous syllables, we would obtain peaks in entrainment similar to the ones reported by ASSR experiments, or the ones described in Ding et. al (2015). This also means that we did not test 'entrainment in the narrow sense' (Obleser & Kayser, 2019), because the apparition of steady-state potentials are always due to evoked activity, but this type of procedure also ensured higher observable phase coherence. Thus, this allowed us to more easily test whether any differences in entrainment were solely due to the syllable-initial consonants, because only these were different across conditions.

The existing research tells us little about how different phonemic features are processed by the brain, and, more specifically, how they affect entrainment. So far, we know that different consonants are processed by discrete regions of the auditory cortex, and that separate clusters can be noticed depending on the manner of

articulation and voicing (Mesgarani et al., 2014). However, an fMRI study also showed substantial overlap between regions responsible for different consonant groups (Arsenault and Buchsbaum, 2015). While not directly related to entrainment, these results may suggest that different consonants may be more similar than they are different in the quality of acoustic edge that they provide. The areas of the brain which show discrete processing of phonemes are also responsible for comprehension, in the sense that these respond stronger to speech stimuli than to non-words or noise (Binder, Frost, Hammeke, Bellgowan, Springer, Kaufman and Possing, 2000). This is significant if we consider that the syllabic rhythm, but not necessarily the phonetic information in the envelope, is largely considered responsible for comprehension (Edwards & Chang, 2013). Furthermore, Di Liberto, O'Sullivan and Lalor (2015) found that when phonemic information is added to the speech envelope, the correlation between the spectro-temporal response function of the auditory cortex and the enriched envelope is stronger than between these and the envelope alone. The results from these two studies may indicate that phonemes provide information which is important for tracking the syllabic rhythm.

In a recent experiment, Oganian and Chang (2018) found that the correlations between neural responses and the speech envelope were the highest at the time of the peak derivative of the envelope. This was thought to be due to the fact that the peak derivative conveys the maximum rate of change in the envelope, possibly corresponding to formant or CV transitions. We considered both the size and latency of the peak derivative in the analyses for Experiment 2, alongside 'Doelling sharpness' and normalised sharpness. We also extracted the latencies of CV transitions manually, which allowed us to see the correspondence between these and the latency of the peak derivative. Furthermore, we calculated the maximum

amplitude of the envelope and its latency, which are associated with vowel effects. Lastly, we considered the overall shape of the envelope, including ascending and descending slopes, by extracting the Gini Index (see Methods). All the different edge markers were then correlated with the EEG responses.

We calculated entrainment to speech by extracting the power and phase coherence of the Fourier transform of the EEG waveforms, at the frequency of stimulation as well as harmonic frequencies. Harmonic responses were considered because sometimes, for complex stimuli containing a range of frequencies, the power or coherence of the Fourier transform can be higher at multiples of the fundamental frequency than those at the rate of stimulation (see General Introduction or Zhou et. al (2016). Furthermore, the existence of harmonic responses may reflect the possibility of activation of multiple cortical cell populations which are mutually synchronised, such as discrete regions activated by different phonemes which are coupled to the syllabic rhythm. Harmonic patterns of responses have been observed in research using multi-unit recordings and researchers argue that the nonlinear coupling between cortical oscillations with different preferred resonance frequencies is responsible for this phenomenon (Langdon, Boonstra and Breakspear, 2011). Consequently, the integration of harmonic responses could help us observe differences in the processing of multiple phonemic features.

We expected that strong edge markers, such as high sharpness, earlier latencies of peak derivative and envelope peaks, or high peaks of the envelope or its derivative, to be correlated with entrainment. We also expected the edge markers to be correlated with one another. Nevertheless, we did not make any specific predictions about which edge marker would be the best landmark entrainment, or which consonants would lead to more phase locking. Because little research was

conducted in the past on the topic of phonemic edges and syllabic entrainment, we considered that this to be a largely exploratory study, and reserved interpretations for the discussion of this chapter.

Methods

Participants

Twenty-five right-handed native English speakers (17 females, mean age = 23.68 years old, standard deviation = 5.28 years), without any learning disabilities or hearing impairments, were recruited using University of Bristol's Experimental Hours System or through social media advertisements. They were rewarded for their time with either course credit or financial compensation (£10/hour). Participants were allowed to withdraw at any moment from the study, in conformity with the University of Bristol Human Participants Ethics Guidelines.

Design

The experimental design was within-subjects and there were 15 conditions, depending on the nature of the stimulus: stimuli in 14 of the conditions comprised of consonant-vowel syllables, and one contained vowel-only stimuli.

Stimuli

The syllables were obtained by recording a female native English speaker in a soundproof room, using Cool Edit Pro software (Adobe Systems Inc.). For each syllable needed for the experiment, she uttered the same syllable ten times, and we kept the clearest recording of that syllable. The syllables were then normalised to 70 dB SPL and shortened using a custom Python script which applied the Pitch

Synchronous Overlap and Add (PSOLA) algorithm for duration modification. While the target duration was 250 ms, the lengths of the syllables differed between each other by a maximum of 10 ms. This was a consequence of using gammatone filters in order to preserve the original pitch of the sound. We also considered that a slight difference in the duration of the syllables would avoid possible effects of adaptation and was therefore beneficial for the experiment.

The syllables were either vowel-only or consonant-vowel, containing one of the following consonants: /b/, /d/, /g/, /k/, /p/, /t/, /m/, /n/, /s/, /z/, /l/, /r/, /f/, /v/. The vowels used were /a/ (as in the “a” in “bar”), /e/ (“e” in “error”), /i/ (“ee” in “bee”), /o/ (“o” in “pot”) or /u/ (“oo” in “coo”). We built 3 separate streams for each condition, in which the order of the vowel was pseudo-randomised, such that the same vowel was not repeated in consecutive syllable. For example, in the vowel-only condition, the first five syllables of each stream were, in order: /a/, /u/, /e/, /a/, /i/; /e/, /a/, /i/, /u/, /e/; and /i/, /o/, /u/, /o/, /a/. The order of the vowels was then kept the same for each stream of the additional CV streams. For example, the first five syllables of each of the three streams in the /b/ condition were: /ba/, /bu/, /be/, /ba/, /bi/; /be/, /ba/, /bi/, /bu/, /be/; and /bi/, /bo/, /bu/, /bo/, /ba/.

The stimuli were five seconds long and contained 20 such syllables. Each stimulus was repeated 10 times. We also created filler stimuli, to assure that participants would remain awake throughout the duration of the experiment. Fillers were stimuli which contained a single syllable starting with a different consonant from the dominant one (e.g. a single “fa” syllable in a /b/ stream, such that “ba bo bee fa be” would be the last five syllables in the stream). Participants were required to detect the “different” syllable, which was always inserted after the second half of the stimulus, to ensure that as much attention as possible was given to every new

stream. Each filler stimulus was only presented once. In total, there were 450 target stimuli, including repetitions, and 50 filler stimuli. Examples are given in the Appendix 3.1. All stimuli are freely available on the Open Science Framework website (see link in Appendix 3.1).

Apparatus

The apparatus was identical to Experiment 1.

Procedure

The duration of the EEG setup, described previously in Experiment 1, was approximately 40 minutes, and the experiment lasted 1 hour and 5 minutes. Participants were told that on some trials, they will hear a syllable starting with a different consonant than the other syllables in the stimulus (as in “ba bo bee **fa** be”), and were instructed to remember it. After each filler stream, a question appeared on the screen asking them to type in the syllable. Participants typed “none” if they could not hear a different syllable. A sad or a smiling emoji was shown on the screen after each keyboard response, as feedback for their performance.

Participants were given examples of target syllables before the experiment, with the correct spelling for each vowel, such that their performance was not affected by spelling, but only by the degree of attention or intelligibility. A practice block was played in the beginning of the experiment. This contained four /b/ stimuli, in which the orders of the vowels were different than the ones used in the main experiment, and two fillers based on the same consonant, which were also not present in the main tasks. For the main tasks, we recorded the performance of participants and

those with more than 50% incorrect responses to filler stimuli were eliminated from the analysis.

There were five experimental blocks, which lasted 20 minutes each and containing 100 individual streams of syllables. The stimuli were pseudo-randomised so that each block comprised of 10 fillers and 90 experimental streams, i.e., 10 repetitions of streams beginning with three different consonants, for which three separate streams existed. These stimuli were not played in any other blocks.

The inter-stimulus interval was two seconds and thirty-second breaks were provided between each 20-minute block. Participants had the choice of taking a longer break as they could only start the next block after pressing the “Enter” key, ensuring that they obtained the necessary amount of rest.

Data analysis

EEG

All EEG data were processed in Matlab, using the EEGLAB toolbox for pre-processing. The data were low-pass filtered at 50 Hz, re-referenced to the average of all channels and split into five second-long epochs. We used custom scripts for time-frequency analyses. These are available online in the link provided in Appendix 3.2.

We conducted ICA in order remove to eye-movement related components. A component was removed if the frontal channels in its topography contained more than 12% of total EEG power. In an EEG study using isochronous stimuli, Ding et al. (2017) reported removing a component if this power exceeded 10%. However, because removing an ICA component has the potential of interfering with data at the lowest frequencies, or affecting the size of the EEG, and because we used less EEG

channels than Ding et al. (2017), we set the higher threshold of 12%. Very few components were removed as a result.

The first 500 ms of each epoch were removed from subsequent processing, in order to prevent potential interference with auditory ERPs (Ding et al., 2015; 2017). A fast Fourier transform was obtained for each epoch using a Hanning taper, and subsequently the Evoked and Induced Power, as well as the Inter-trial Phase coherence (ITC), were calculated using the following formulas:

$$Power_{Evoked} = \left| \frac{\Sigma_K X_k(f)}{K} \right|^2 \quad (3.1),$$

$$Power_{Induced} = \frac{\Sigma_K |X_k(f)|^2}{K} \quad (3.2),$$

$$ITC = \left(\frac{\sum_K \cos(\theta_{ij})}{K} \right)^2 + \left(\frac{\sum_K \sin(\theta_{ij})}{K} \right)^2 \quad (3.3),$$

Where $X_k(f)$ is the value of the Fourier transform at frequency f and K = 10 is the number of repetitions per stream.

The Evoked Power reflects responses which are phase locked to the stimulus, while induced power reflects the power corresponding to endogenous activity, which is different from baseline activity (David, Kilner and Friston, 2006). If peaks are observed in the Evoked, but not Induced Power, one could confirm the existence of entirely stimulus-dependent activity. Therefore, subsequent analyses on the Evoked Power and ITC would be sure to reflect the entrainment to our stimuli.

To test the effects of periodic stimuli on brainwaves and separate these from background activity, the ASSR literature reports using the F-ratio to compare the neural signal at the frequency of interest to the same signal averaged over neighbouring frequencies. Typically, for a periodic stimulus of 80 Hz, the ITC at 80 Hz would be compared to the average ITC taken over five Hz in each direction:

between 75 and 80 Hz, and 80 and 85 Hz, but excluding 80 Hz exactly (John, Lins, Boucher and Picton, 1998). The choice of comparing the peak at the stimulation rate to the average of the bins spanning five Hz in each direction seems arbitrary, but also relates to the higher frequency of stimulation. Due to the low-frequency of our stimuli (4 Hz), we did not choose a range of 5 Hz for these comparisons. Instead, we ran paired T-tests to compare the average EEG measure (ITC or evoked power) at 4-Hz syllable rate with each of the 9 neighbouring bins in either direction. For example, the ITC at 4 Hz was compared to each bin from 2 Hz to 6 Hz, spanning a total of 4 Hz, or 18 bins. This was repeated in the exact same fashion for harmonic responses. The p-values of multiple comparisons were corrected using false discovery rate (FDR).

To test the differences in entrainment between the groups of consonants to which the different stimuli belonged to, we conducted univariate repeated measures analyses of variance (ANOVA), together with post-hocs, as well as paired, two-tailed T-tests. All statistical analyses between responses to different phonetic groups were done in RStudio version 1.2.335.

Stimuli

We extracted different stimulus properties to be able to study the relationship between sound sharpness and the EEG response. As in Doelling et al. (2014), we obtained the narrowband envelope of each syllable by applying a cochlear filter of 32 log-spaced frequency bands, spanning between 80 and 8000 Hz. Using the Hilbert transform, we calculated the envelope of each band separately, and summed them together to obtain the final signal. Further analyses were conducted on the summed

envelope of each syllable, which were averaged over the number of syllables in a stream, and where applicable, over the number of streams in a single condition.

We measured the original 'Doelling sharpness', by taking the sum of the positive derivative of the summed envelope. A normalised version of sharpness was also obtained by averaging the total Doelling sharpness by the sum of the envelope (see Methods section of Experiment 1). However, the 'Doelling sharpness' and its normalised version only give the total value of the ascending slopes in the speech envelope, without focusing on any specific syllabic landmarks. While this may be especially problematic for syllables containing multiple sounds (e.g., consonant-vowel-consonant, consonant-consonant-vowel), the lack of sharpness effects on entrainment in Experiment 1 indicated that we needed to consider other edge markers.

Doelling et. al (2014) were especially concerned with the ascending slopes in the envelope measured from the point of syllabic onset and to its peak, which describe acoustic edge. If these two points are connected, it appears that the edge which they form can be accounted for by the maximum amplitude of the envelope and its latency (see Figure 3.1). We therefore extracted the values of the maximum amplitude and its latency separately, which allowed us to test whether either or both of these edge markers would be correlated with entrainment. Furthermore, the peak of the envelope reflects the maximum intensity of the vowel (Ghitza, 2011). Therefore, if vowel-related measures such as the peak envelope and its latency were more correlated with entrainment than other considered edge markers, this could reflect the advantage of vowel landmarks over other possible ones.

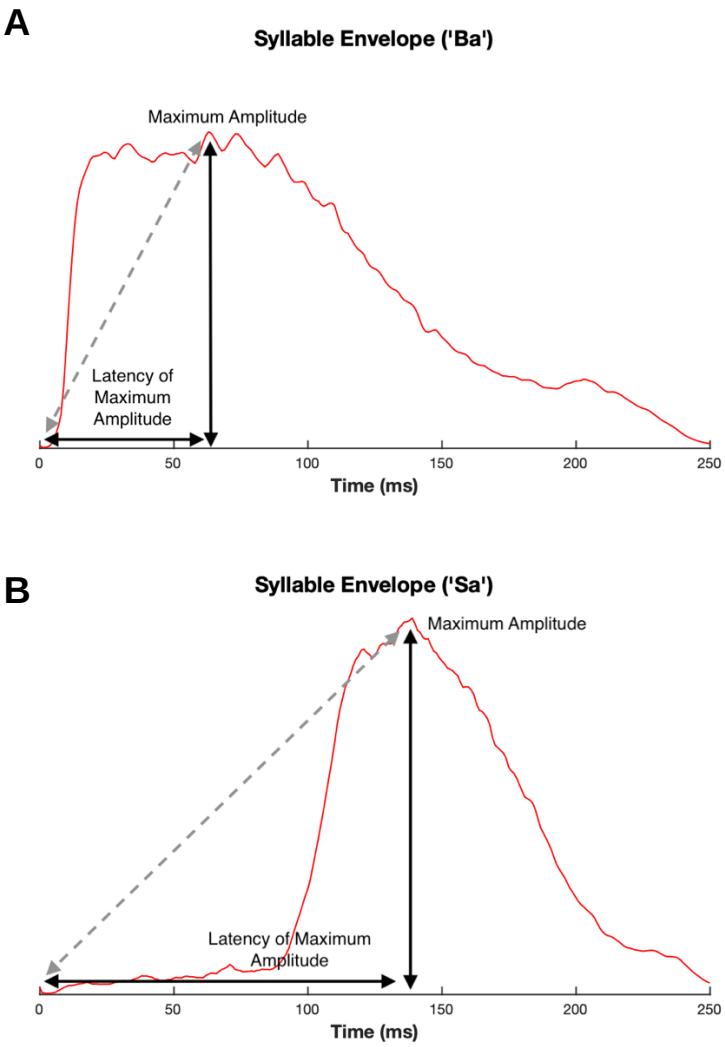


Figure 3. 1. Syllables Envelopes with black arrows showing the time of the maximum amplitude of the envelope, as well as its latency. Notice how they define different slopes (interrupted arrows), with the slope for 'ba' being steeper than the one for 'sa'. These slopes, however, are not the same as sharpness, which describes all of the slopes of the rises in the envelope, but the two may be related. We hypothesised that syllables containing some phonemes (such as stops) at the onset of syllables would have more sharpness than others (such as sibilants) based on the steepness of the slopes contained in their envelopes. One way of testing this, as well as investigating other factors which may be important for entrainment, is to look for the separate effects of the maximum amplitude of the envelope, and its latency, on phase locking to speech. A. Envelope of syllable 'ba'. B. Envelope of syllable 'sa'.

The peak derivative is thought to be especially reflective of onset properties (Oganian and Chang, 2018). Peak derivative values and their latencies were obtained from the broadband envelope of each syllable, and averaged over condition. The broadband envelope was calculated by low-pass filtering the absolute value of the sound waveform at 10 Hz, and by then extracting the positive values of the filtered sound. Figure 3.2 shows the difference between a summed narrowband envelope and a broadband envelope. Note that the broadband envelope is

smoother, so it may convey the maximum rate of change more accurately than the summed narrowband envelope which is noisier. The latencies of the peak derivative were compared with CV-transitions which were manually extracted using Praat software (Ding et al., 2017).

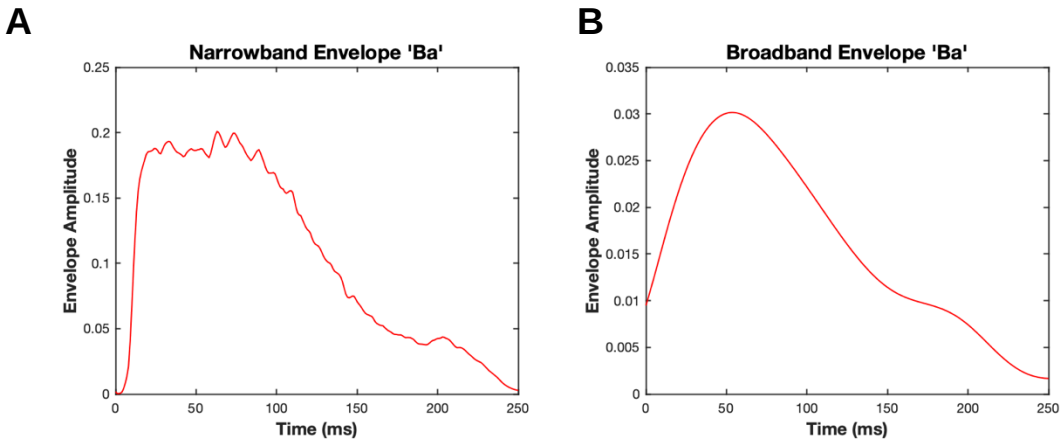


Figure 3. 2. A. Envelope of syllable ‘ba’ was calculated by summing the narrowband envelopes obtained for waveforms filtered between frequencies 80 to 8000 Hz from the original sound. B. Broadband envelope of syllable ‘ba’, obtained by filtering frequencies below 10 Hz. Notice how B resembles an average shape of A.

The CV transitions of the narrowband envelopes were obtained by visually inspecting each syllable, and taking the time at which the vowel started, or where its periodicity became apparent. The syllable was thus decomposed into a “consonant part” and a “vowel part”. The method of manually extracting CV transitions is explained in Figure 3.3. By playing either parts in isolation, one would not be able to hear the other part, e.g., by playing the /s/ in Figure 1, one would not hear the /a/.

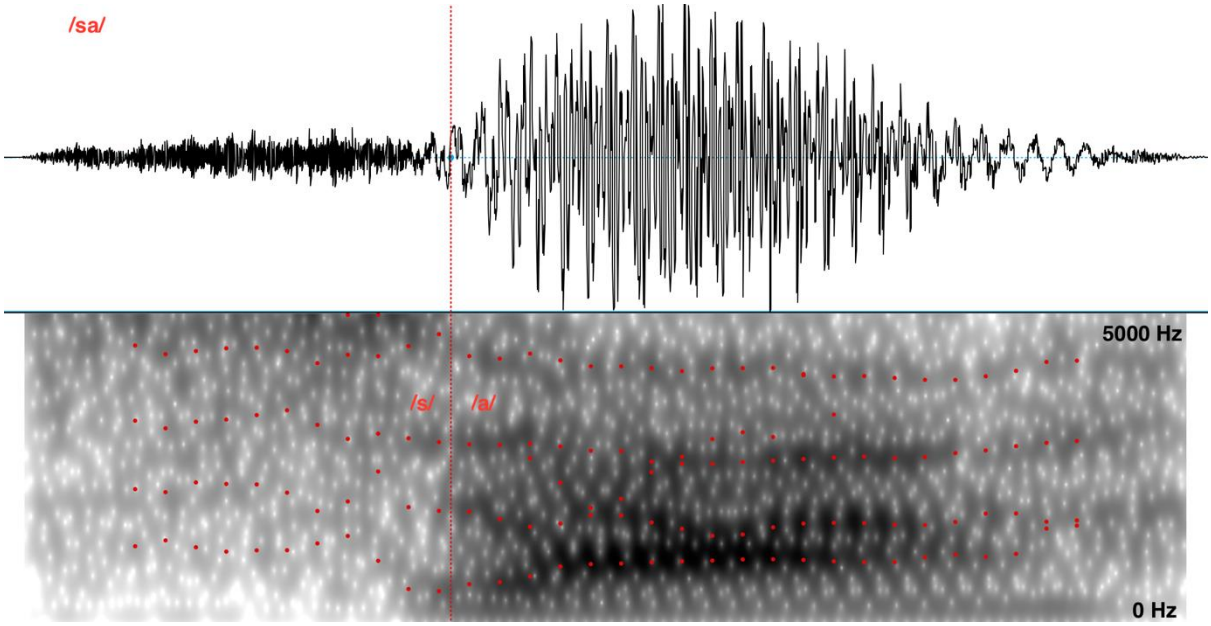


Figure 3. 3**.** Waveform and spectrogram of syllable "sa", demonstrating the location of the CV transition. The top half represents the waveform, and the bottom graph is the spectrogram, calculated in Praat, between 0 and 5000 Hz. The dotted red line splits the syllable into a 'consonant part', as indicated in the spectrogram of the syllable, to the left of the line, and a 'vowel part', to the right of the line.

Lastly, we also calculated the Gini index of the each syllable. This a measure of inequality in a given distribution, commonly the distribution of income in a population (Gini, 1921). The Gini index is expressed in equation (4):

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{2n\sum_{i=1}^{n}x_i} \quad (4)$$

Where x_i is the income corresponding to person i. For example, if we arrange the percentages of people from the lowest to the highest income against the cumulative share of the total population income, we will obtain an income distribution whose shape is defined by the Lorenz curve (see Figure 3.4). If the distribution of income is equal, the Lorenz curve would be identical to the 45 degree line (also known as the line of equality). The Gini index can also be expressed as the ratio of the area defined between the Lorenz curve and the line of equality (A), over the total area

beneath the 45 degree line. This coefficient is a number which takes values between 0 and 1, with 0 representing complete equality, and 1 being complete inequality.

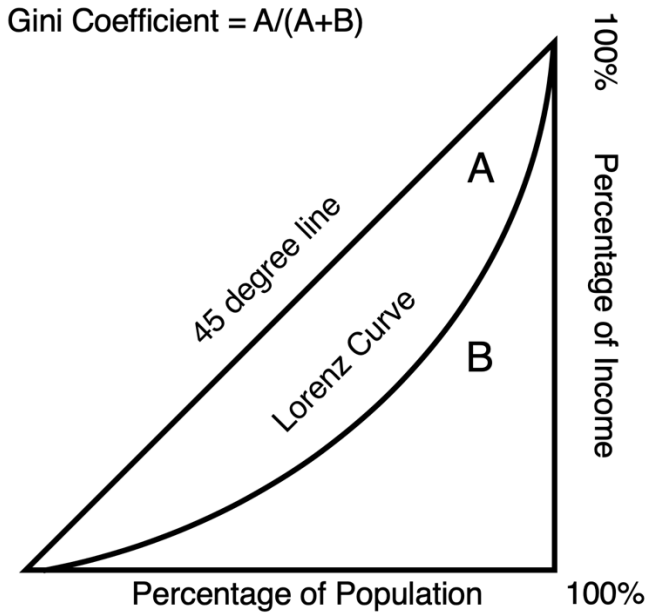


Figure 3. 4. Expression of Gini Index in terms of the Lorenz curve. The cumulative share of population from the smallest to the highest income is plotted against the cumulative percentage of income. The shape of this distribution is defined by the Lorenz curve. 45 degree line represents total equality. A = the area defined by the Lorenz curve and the 45 degree line. B = the area under the Lorenz curve. Gini = A/(A+B).

We applied the Gini index to the syllabic envelope, as a means of testing landmark distribution. For example, we considered that if the envelope had a prominent peak surrounded by steep ascending and descending slopes, this would lead to a high Gini index, and therefore would contain a clearly defined landmark. On the other hand, we expected a more uniform envelope to have a low Gini index, and less obvious landmarks. The Gini index would thus be equivalent to the Doelling sharpness, the difference being that, unlike sharpness, the Gini index accounted for both rises and falls in the envelope amplitude.

Finally, each of the edge markers was averaged across all five syllables used in an individual condition, such that each of the 15 conditions were attributed with

their own Doelling sharpness, normalised sharpness, maximum amplitude of the envelope and its latency, Gini index, as well as the peak derivative and its latency .

Results

EEG

Peaks were seen in the frequency response of the evoked power and ITC at 4 Hz and harmonics, at 8, 12 and 16 Hz (Figure 3.5). In the ITC, these peaks were significantly higher when compared to responses in neighbouring bins ($p<.001$ at 4 and 8 Hz, $p<.01$ at 12 Hz, $p<.05$ at 16 Hz; FDR-corrected). In the evoked power, the characteristic alpha response was seen between 8-12 Hz, and only the peak at syllabic rate was significantly higher than activity at nearby frequencies (Maximum $p<.05$, FDR-corrected). No peaks other than alpha were seen in the induced power, as expected.

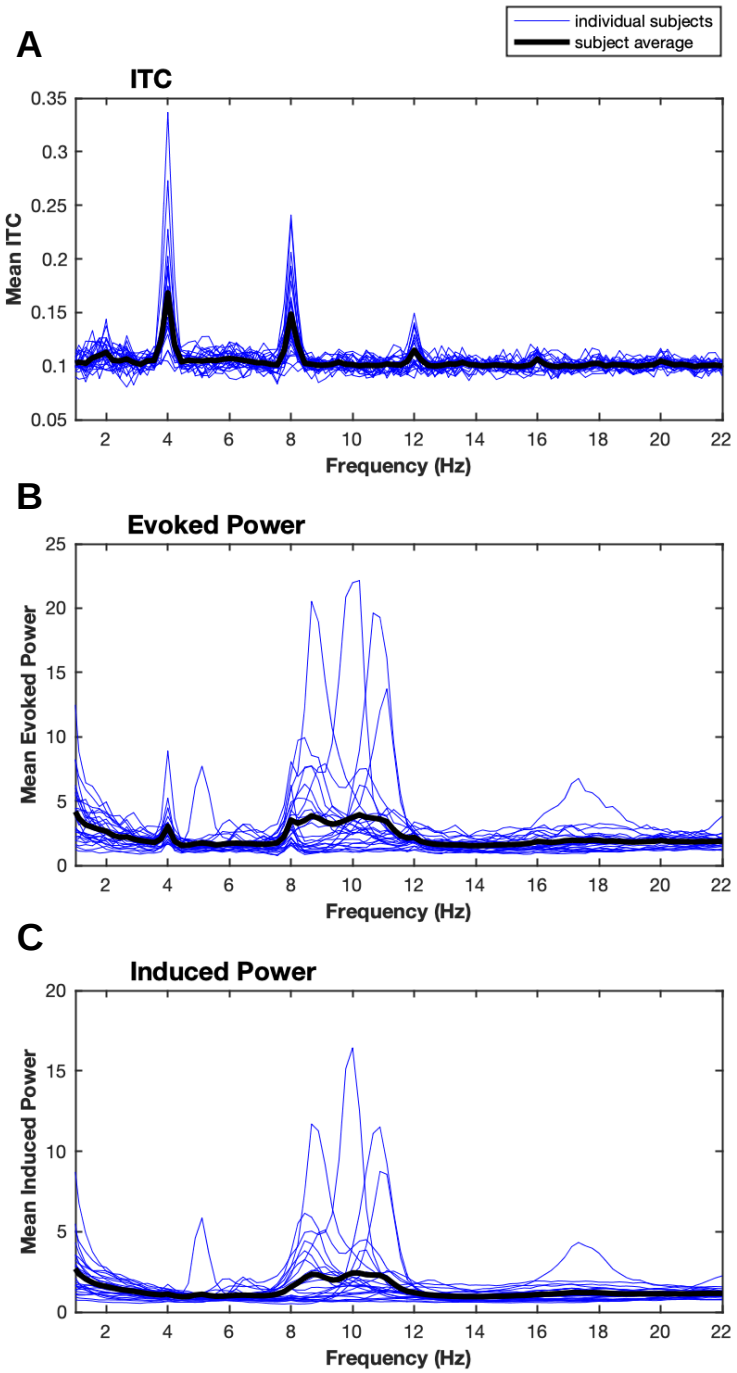


Figure 3. 5. ITC, Evoked Power and Induced power, averaged over channels and conditions, are plotted as a function of frequency, between 1 and 22 Hz. Bold black lines represents averages over all subjects. Each of the blue lines represents an individual subject. Peaks can be noticed at 4, 8, 12 and 16 Hz in ITC and Evoked Power, but not in induced power. A. ITC B. Evoked Power. C. Induced Power.

We performed Pearson's correlations on the ITC averaged over channels and conditions, at the syllable rate and its harmonics, in order to see if the significant peaks in phase locking reflect the same responses to speech. We found moderate positive relationships between the ITC at 4 and 8 Hz (Pearson's r = 0.54, p<.01), 4 and 12 Hz (Pearson's r = 0.59, p<.01), 8 and 12 Hz (Pearson's r = 0.56, p<.01) and

8 and 16 Hz (Pearson's r = 0.58, p<.01). No significant correlations were obtained between the ITC at 4 and 16 Hz (Pearson's r = 0.37, p = n.s.), and the ITC at 12 and 16 Hz (Pearson's r = 0.25 , p = n.s.). When correlations were conducted between ITC values per stream, a significant relationship was found only between the ITC at 4 and 8 Hz (Pearson's r = 0.55, p<.001). Figure 3.6 depicts the relationship between the ITC at 4 and 8 Hz, for individual streams, as well as when these were averaged over their corresponding conditions.

The existence of moderate correlations between most, but not all of the peaks in the ITC may be due to the fact that cortical subpopulations of neurons with different resonant frequencies, coupled in a nonlinear fashion, were responsible for the existence of harmonic responses. Because the nonlinearity implies that sometimes harmonic responses can be stronger than the ones to the driving frequency (Langdon et al., 2011), and because activations of different cell populations can reflect differences in phonetic feature processing, we considered all the significant peaks in the ITC for statistical tests.

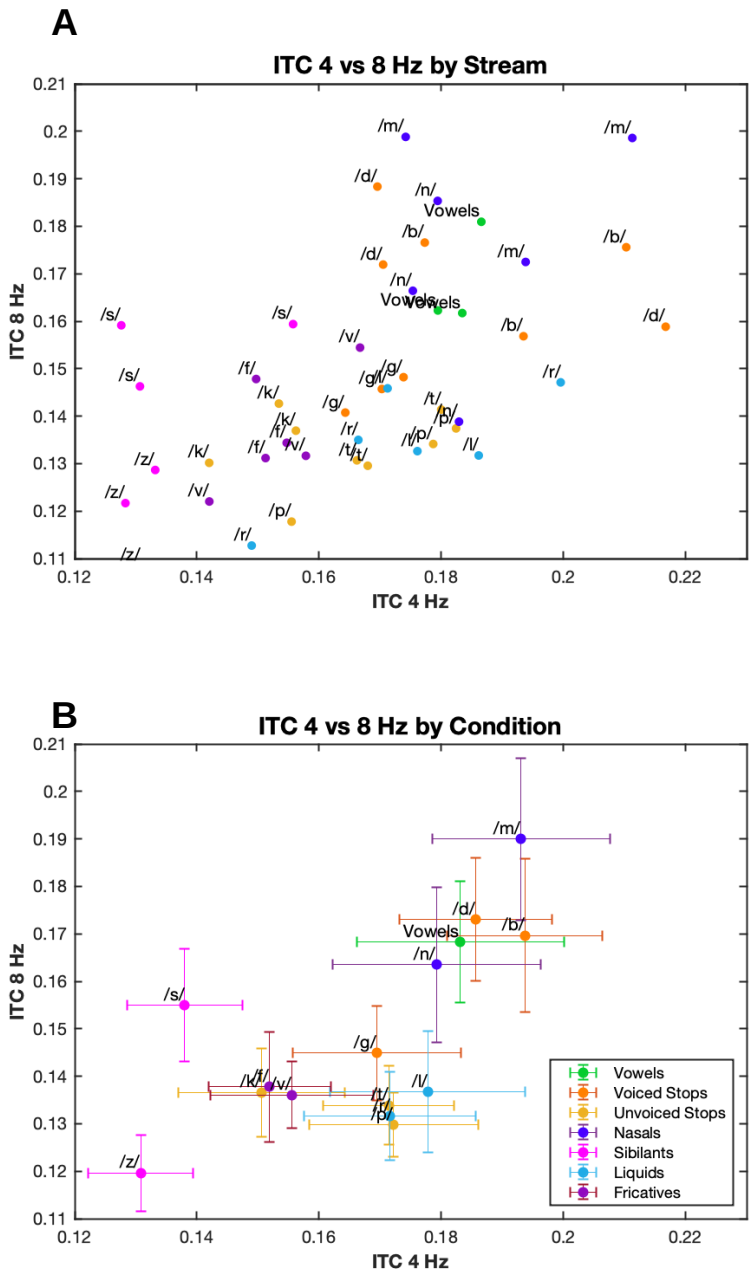


Figure 3. 6. The ITC values at 4 Hz are plotted against those at 8 Hz. Different colours represent the phonetic category corresponding to the manner of articulation of the consonant, or vowels, respectively (see legend in B). A. ITC values are plotted for individual streams. The dominant consonant of each stream is indicated above each scatter point. Vowel-only streams are also plotted. B. ITC values are averaged for streams belonging to the same consonant (or to the vowel) condition. Errors bars represent ± standard error of the mean for the ITC at each of the two frequencies.

Preliminary statistical analyses were conducted for different phonemic groups in the ITC at 4 and 8 Hz. Initially, we split these groups into vowels, stops, nasals, rhotics, fricatives and sibilants. Using a Bonferroni correction for six groups, we found that in the 4 Hz ITC, sibilants were significantly different from nasals ($p<.01$), rhotics ($p<.05$) and stops ($p<.001$). On the other hand, the 8 Hz ITC showed a significant difference between rhotics and nasals only ($p<.01$, Bonferroni). Because

results were so different between the 4 and 8 Hz ITC, but because these were also positively correlated with each other, we decided it was useful to combine our significant ITC peaks into individual measures using PCA.

Principal Component Analysis of ITC

Principal Component Analysis (PCA) was conducted on the ITC at 4, 8, 12 and 16 Hz. PCA is an orthogonal transformation applied to correlated variables in order to produce another set of perfectly uncorrelated variables, called the principal components (Pearson, 1901). Principal components are linear combinations of the input variables. The first component always explains the most variability in the data, with each successive component having the highest possible variance whilst being orthogonal to the previous components. Through PCA one can obtain a new orthogonal basis for the data, where the data in each direction are uncorrelated.

In statistical analyses, only the components which are able to account for a significant amount of variance are kept. The most common rules of thumb for PCA component choice are to keep those components which together explain between 70% and 90% of the total variance across variables (Joliffe, 2002), or those which individually account for over 5% of the variance (Xue et al., 2011). Because we had a low number of variables – the four ITC peaks – we avoided more computationally intensive methods and followed the above rules. We considered the first two components of the PCA for further analysis, which explained 76.2% and 20.08% of the variance, respectively.

The factor loadings describe the relative contribution of each of the original variables to each of the components, and the scores represent linear combinations of the original variables for each sample (Xue et al., 2011). The positive values of the

loadings describe the positive correlations amongst the variables, and the negative values indicate the presence of negative correlations. The PCA loadings corresponding to the ITC at each of the four measured frequencies can be found in Table 3.1. The ITC values at 4 Hz affected the first principal component the most, followed by moderate weights from the ITC at 8 Hz, and marginally by the ITC at 12 and 16 Hz. All these influences are positive, implying that the first component only considered the positive correlations between the ITC peaks, and is best described by the positive correlation between the ITC at 4 and 8 Hz (Figure 3.6). On the other hand, the second principal component was influenced the most by the ITC values at 8 Hz, followed by a moderate, negative effect of the ITC at 4 Hz, and then again marginally by the ITC at the remaining harmonics. Therefore, the second component, orthogonal to the first one, considered both positive and negative contributions of the ITC peaks.

The values of the ITC at 4 Hz and harmonics, averaged over channels, were multiplied by their corresponding loadings in each of the two components, leading to two different linear combinations of ITC results. Further statistical tests were conducted on these two combinations and, for parsimony reasons, we will refer to these as Compound ITC1 and Compound ITC2. The relationship between the PC1 and PC2 scores, for each stream or condition, is illustrated in Figure 3.7. In this plot, the separation of different consonants, even if not based on a particular feature, becomes more readily noticeable than in Figure 3.6, and also one can observe a reduction in the size of the error bars. Thus, we showed that the PCA analysis helped to show results in a more helpful way.

Table 3. 1. Factor loadings for the ITC values. They are given for each of the four principal components, after PCA. “PC” = principal component. The amount of variance explained by each component is noted below the loadings.

ITC at frequency:	PC1	PC2	PC3	PC4
4 Hz	0.8251	-0.5510	-0.1227	-0.0248
8 Hz	0.5318	0.8295	-0.1259	-0.1152
12 Hz	0.1807	0.0481	0.9758	0.1136
16 Hz	0.0620	0.0775	-0.1302	0.9865
Variance Eplained	76.2%	20.08%	2.98%	0.73%

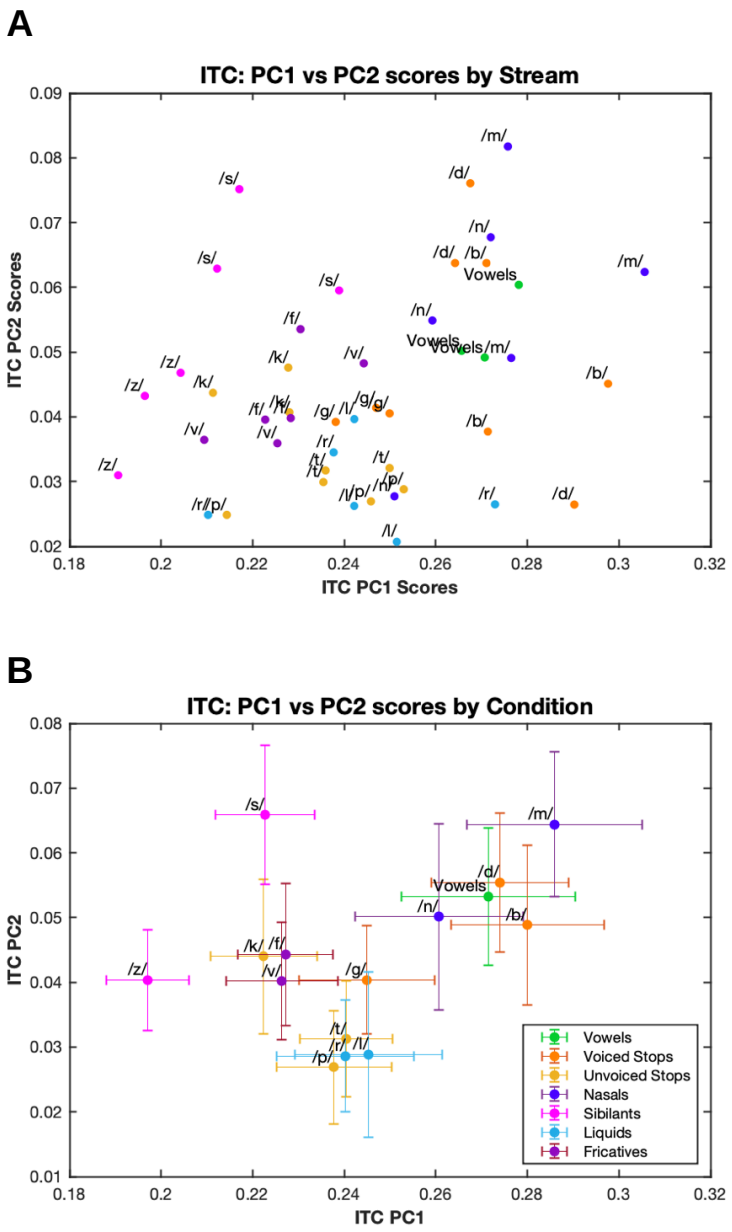


Figure 3. 7. The ITC at 4, 8, 12 and 16 Hz is multiplied by their respective PCA loadings for the first two principal components, and then plotted for each stream and each condition. Different colours in both graphs represent the phonetic category corresponding to the manner of articulation of the consonant, or vowels, respectively (see legend in B). A. ITC scores are plotted per stream. The dominant consonant (or vowel) in each stream is indicated above each scatter point. B. ITC scores are averaged over each consonant (or the vowel) in each condition. Error bars represent ± standard error of the mean for the ITC at each of the two frequencies.

Due to the alpha contamination at frequencies between 8-12 Hz and the lack of significant peaks in this frequency range, we did not conduct any correlations or PCA on the evoked power at 4 Hz and its harmonics.

Effects of phonemic group

The ITC and evoked power means from multiple conditions were compared by conducting one-way repeated measures analyses of variance (ANOVA), with Greenhouse-Geisser corrections where the assumption of sphericity was not met. First, the 4 Hz power as well as the two different compound ITC measures were averaged across the values for the three different streams from each condition. For the 15 conditions, the ANOVAs were significant for all three measures (Compound ITC 1: $F_{14,336}$ = 11.43, p<.001; Compound ITC 2: $F_{14,336}$ = 1.83, p<.05; 4 Hz Evoked Power: $F_{14,336}$ = 2.53, p<.01). These results indicate that the stimuli in different categories elicited results which were significantly different from each other, but because pairwise comparisons reveal little when the number of conditions is high, we conducted further analyses by collapsing the EEG measures into groups corresponding to the stimulus' relevant phonetic categories, depending on the manner of articulation. Initially, vowel conditions were omitted from the analysis.

First, we split the results into five phonetic groups, and then further combined some of these into three phonetic groups. The five phonetic groups were stops (/b/, /d/, /g/, /k/, /p/, /t/), nasals (/m/, /n/), sibilants (/s/, /z/), fricatives (/f/, /v/), and liquids (/l/, /r/). Then, we further averaged across nasals and liquids (sonorants), and sibilants and fricatives (sibilants are a subset of fricatives, see above), as such

consonants are often grouped together by linguists due to their similarity in the release of the air flow during articulation (Chomsky and Halle, 1968).

In the Compound ITC1, the ANOVA elicited a main effect of consonant group when conducted on five consonant categories ($F_{4,96}$ = 11.43, p<.001) and on three consonant groups ($F_{2,48}$ = 12.06, p<.001). Post-hoc T-tests using the Bonferroni method for multiple comparisons revealed that entrainment was the lowest in the sibilant category, this being significantly smaller than for nasals (p<.001), liquids (p<.05) and stops (p<.001). There was also a significant difference between nasals and fricatives, with the former eliciting higher Compound ITC1 values (p<.05). When paired with sibilants, fricatives showed significantly less phase locking than nasals/liquids (p<.01, Bonferroni) and stops (p<.001, Bonferroni), but there was no difference between nasals/liquids and stops. The average Compound ITC1 to both five and three groups of consonants can be seen in Figure 3.8.

The results of ANOVAs and post-hoc t-tests for five and three groups of consonants in the evoked power at 4 Hz and Compound ITC2 are given in Appendix 3.3. These findings were either similar or less significant than those in the Compound ITC1. However, post-hoc T-tests on Compound ITC2 indicate that nasals elicited higher entrainment than liquids (p<.05, Bonferroni), a difference which was not present in the other measures. These results also indicate that ITC1 mainly reflected the trends noticeable in ITC measured at 4 Hz, whereas ITC2 showed similar group differences as the ITC at 8 Hz.

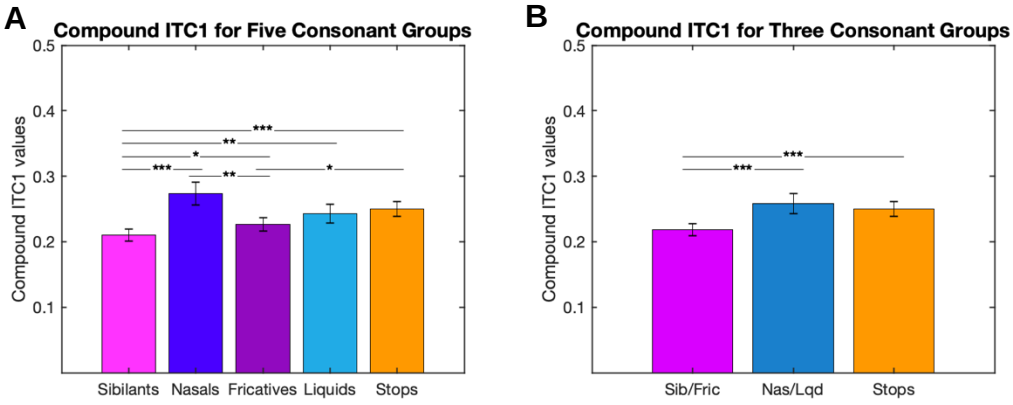


Figure 3. 8**.** A**.** Values of the Compound ITC1 are averaged over five consonant groups: sibilants, nasals, fricatives, liquids and stops. A one-way ANOVA conducted on these groups was significant, p<.001. Significance between groups is also shown above the bars, with * indicating p<.05, ** p<.01, *** p<.001. Error bars represent ± standard error of the mean. B. The compound ITC1 is averaged over three groups: sibilants/fricatives, nasals/liquids and stops. Again, ANOVA for these groups was significant, p<.001. Significance concerning individual group differences is represented above the bars, like in A. Error bars represent ± standard error of the mean.

We also conducted paired T-tests between responses to voiced and unvoiced consonants, by first averaging across EEG measures to both stops and fricatives, and then running separate analyses on each of these respective groups. The T-tests revealed an effect of voicing only in the Compound ITC1. Voiced consonants produced higher ITC when responses were averaged across both stops and fricatives ($t(24) = 2.39$, $p<.05$), as well as when the stop consonant category was studied separately ($t(24) = 3.78$, $p<.01$). However, unvoiced fricatives led to more tracking than voiced ones, albeit this difference was almost marginally significant ($t(24) = -2.06$, $p=.05$). Because voicing had opposite effects on stops and fricatives, we argue that another acoustic feature related to voicing may be responsible for the differences found at each of these consonant groups. This is explored in the next section about correlations between entrainment and stimulus edge markers.

Lastly, we conducted a separate ANOVA on the Compound ITC1, which included responses to vowels, separate responses to voiced and unvoiced stop consonants, as well as those to sibilants, fricatives, nasals and liquids. Here, we

aimed to explore the differential effects of vowels, as well as voiced and unvoiced stops, amongst the investigated phonemic groups. We did not separate voiced and unvoiced fricatives because the small difference between them obtained using an uncorrected T-test was unlikely to reflect in the multiple comparisons conducted between multiple groups. The ANOVA was significant ($F_{6,144}$ = 10.75, p<.001), and the values of the seven groups are illustrated in Figure 3.9, with significance values emphasising vowel and stop consonant groups. Post-hoc comparisons revealed that the vowels triggered significantly more phase locking than sibilants only (p<.001, Bonferroni), while voiced stops elicited stronger tracking than sibilants (p<.001, Bonferroni), fricatives (p<.05) and unvoiced stops (p<.05, Bonferroni). Unvoiced stops only showed smaller phase locking compared to voiced ones, but were not different from any other phonemic groups.

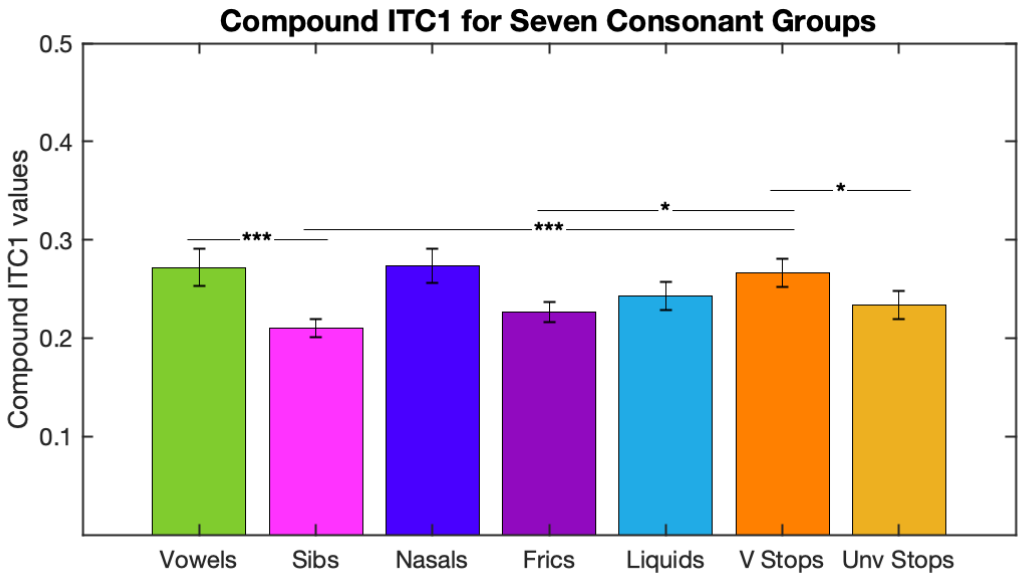


Figure 3. 9**.** Values of the Compound ITC1 are averaged over seven consonant groups: vowels, sibilants, nasals, fricatives, liquids, voiced stops and unvoiced stops. A one-way ANOVA conducted on these groups was significant, p<.001. Above the bars, significance values indicate: *p<.05, ** p<.01, *** p<.001. These are only plotted for differences between vowels and voiced/unvoiced stops, or between these and other groups. Error bars represent ± standard error of the mean.

The present results suggest that phonemic features may impact neural entrainment to syllables in different ways. Sibilants showed the least Compound ITC1, followed by fricatives, with the distinction between liquids, nasals, stops and vowels being less clear. Liquids showed less entrainment than vowels in the Compound ITC2 only. In terms of voicing, voiced stops showed greater Compound ITC1 than unvoiced stops, but the opposite effect was observed in fricatives. Therefore, the phonemic separation in syllabic entrainment based on manner of articulation and voicing remains ambiguous, but this could indicate that other phonemic features account for differences in phase locking. In the next section, we explore the correlations between the Compound ITC1 and a range of acoustic edge markers we extracted from our stimuli, which may give us a better explanation of our findings.

Correlations between stimulus edge markers and EEG

Edge markers were Doelling sharpness, normalised sharpness, maximum amplitude of the narrowband envelope and its latency, as well as the peak derivative of the broadband envelope and its latency. The Compound ITC1 showed significant correlations with most of the edge markers as seen in Table 3.2, apart from the maximum amplitude of the envelope. As expected, there was a negative relationship between the ITC and the latencies of both the maximum amplitude (Pearson's $r = -0.78$, $p<.001$) and that of the peak derivative (Pearson's $r = -0.84$, $p<.001$). The ITC was also positively correlated with the size of the peak derivative (Pearson's $r = 0.74$, $p<.01$). These results suggest that faster rises in the envelope and earlier, more abrupt changes in the rate of speech lead to better phase locking. However, the ITC was negatively correlated with Doelling sharpness (Pearson's $r = -0.57$,

p<.05), normalised sharpness (Pearson's r = -0.81, p<.001) and the Gini index (r = -0.59, p<.05). This is somewhat counterintuitive, as we expected more sharpness to lead to more, not less, entrainment.

Table 3. 2**.** Pearson's correlations between Compound ITC1 and edge markers.

Edge markers	Pearson's r Score	Significance Value (p)
Doelling Sharpness	-0.57	<.05
Normalised sharpness	-0.81	<.001
Maximum amplitude (MA)	0.32	*n.s.*
MA Latency	-0.78	<.001
Gini Index	-0.59	<.05
Peak derivative	0.74	<.01
Latency of peak derivative	-0.84	<.001

As shown in Table 3.3, the latencies of both the peak derivative and that of the maximum envelope were positively correlated with Doelling sharpness, normalised sharpness and the Gini index. This may suggest that a syllable was sharper the later its envelope reached its maximum amplitude. However, this contradicts our previous assumptions – for example, we claimed that the sudden bursts of energy corresponding to the air release in the production of stop consonants would lead to quicker rises in envelope (and consequently, earlier peaks) as well as more sharpness. A negative correlation between sharpness and the latency of the maximum amplitude of the envelope could be explained by the fact that later peaks of the syllables are also higher, and therefore have greater slopes, but this does not seem to be the case.

There was no significant relationship between the peak of the envelope and its latency (Pearson's r = - 0.03, p = n.s.). Moreover, the maximum amplitude of the envelope was negatively correlated with the normalised sharpness (Pearson's r = -0.31, p<.01), but there was a positive correlation between this and the Gini index (Pearson's r = 0.42, p<.001). Note that the Gini Index was positively correlated with

the normalised sharpness (Pearson's r = 0.48, p<.001). Consequently, the role of the size of the envelope peak as a landmark for entrainment, as well as its relationship with sharpness and envelope slopes remains unclear.

The maximum amplitude of the envelope was nonetheless positively correlated with the size of the peak derivative (Pearson's r = 0.40, p<.001). The peak derivative showed more consistent relationships with the other edge markers: there were negative correlations between this and the two measures of sharpness, as well as between it and the latencies of both the peak derivative and the maximum amplitude of the envelope. The last two were expected, in the sense that earlier changes in the rise of the envelope would also be more abrupt. The negative correlation between the peak derivative and sharpness could potentially explain the positive relationship between the latter measures and peak latencies. Therefore, for the future analyses we kept the peak derivative, but removed the maximum amplitude of the envelope.

Table 3. 3**.** Correlation matrix between seven different edge markers. Scores indicate Pearson's r values, stars indicate significance levels. MA = maximum amplitude.

	Peak Derivative	*Peak Derivative Latency*	*Sharpness*	*Normalised sharpness*	*Gini*	*Latency of MA*	*MA*
Peak Derivative	1.00	-0.44***	-0.62***	-0.65***	-0.03	-0.60***	0.40***
Peak Derivative Latency	-0.44***	1.00	0.39***	0.75***	0.66***	0.73***	-0.06
Sharpness	-0.62***	0.39***	1.00	0.68***	0.21	0.76***	0.09
Normalised sharpness	-0.65***	0.75***	0.68***	1.00	0.48***	0.74***	-0.31**
Gini	-0.03	0.66***	0.21	0.48***	1.00	0.42***	0.42***
Latency of MA	-0.60***	0.73***	0.76***	0.74***	0.42***	1.00	-0.03
MA	0.40***	-0.06	0.09	-0.31**	0.42***	-0.03	1.00

**p<.05*
***p<.01*
****p<.001*

Because these variables were correlated, we conducted PCA on the Doelling sharpness, normalised sharpness, the peak derivative, the latency of the peak derivative, the latency of the maximum amplitude of the envelope, and the Gini Index. Decorrelating these variables gave us a new orthogonal set of values, on the basis of which we could determine a best predictor for entrainment.

Edge markers PCA and Entrainment

Each of the edge marker measures were z-scored before the PCA was conducted, i.e., the mean was subtracted from each value of each property, and then the results were divided by the mean of that measure. The decreasing eigenvalues of each component in the PCA are illustrated in Figure 3.10. We kept the first four components which together explained 95.1% of the total variance (63.33%, 19.59%, 7.58%, and 4.65%, respectively). The coefficients of each principal component were multiplied with each of the original values of its corresponding properties, resulting in new vectors of principal component scores corresponding to each syllable. The factor loadings for the components are given in Table 3.1, indicating the contribution of each of the edge markers to each component.

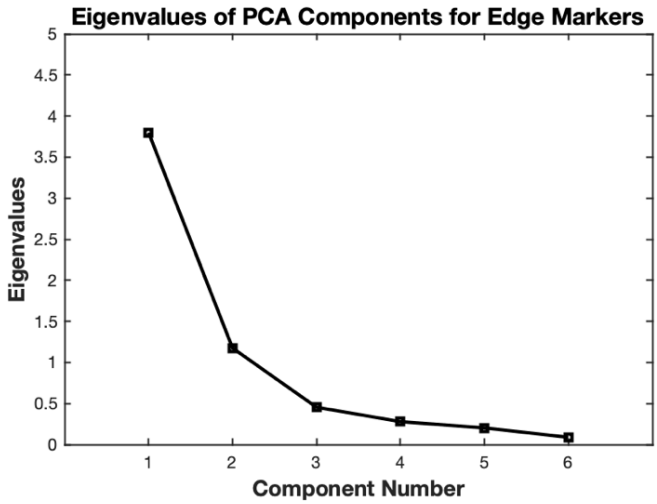


Figure 3. 10. Eigenvalues for each of the PCA components calculated across all edge markers showing significant correlations with each other. These decrease from the first component to the sixth, or the last one.

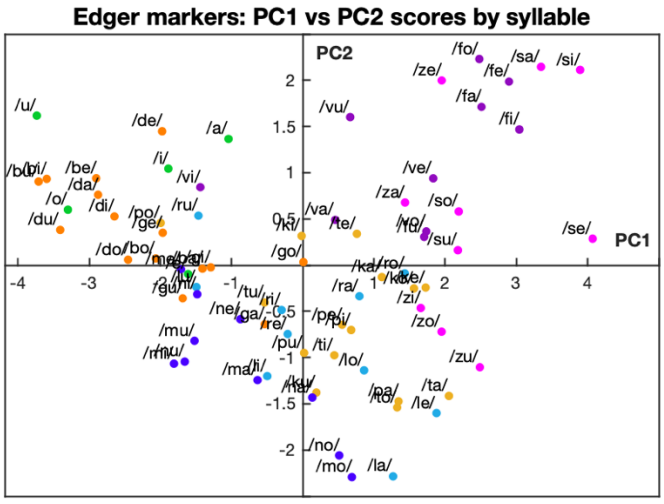
Table 3. 4. Factor loadings of edge markers, for each component of the PC. MA = maximum amplitude.

	PC1	PC2	PC3	PC4
Peak Derivative	-0.37	0.49	0.54	0.30
Latency of Peak Derivative	0.43	0.36	-0.38	0.42
Doelling Sharpness	0.40	-0.36	0.67	-0.19
Normalised Sharpness	0.47	-0.01	-0.17	-0.32
Gini Index	0.28	0.71	0.18	-0.47
Latency of MA	0.47	-0.06	0.25	0.61

The scores at PC1 and PC2 are plotted in Figure 3.11, for individual syllables, as well as the ones for syllables averaged across phonetic condition. These plots show some clustering of the syllables based on voicing and manner of articulation, with fricatives showing positive scores for both components, vowels and voiced stops generally having negative PC1 but positive PC2 scores, nasals showing negative scores for both components, while liquids and unvoiced stops have positive PC1 scores, but negative PC2 scores. Because of their relatively low PC2 scores, it seems that stimuli containing voiced stops (/b/, /d/, /g/), as well as fricatives (/z/,/f/, /s/, followed by /v/), contributed the most to PC1. Unvoiced stops, as well as nasals, liquids, and [v] had moderate scores for PC1, but higher scores for PC2. However, further correlations between the ITC and the scores of each component only revealed a relationship between the ITC and the scores of the first principal component PC1: Pearson's r = -0.85, p<.001. The ITC was not correlated with any of the other four principal components we investigated. The relationships between the Compound ITC1 and the first two components of the edge marker PCA are each

illustrated in Figure 3.12. The syllables with the highest PC1 scores led to the least entrainment, while the ones with the lowest PC1 scores led to the most phase locking (see Figure 3.12A).

A



B

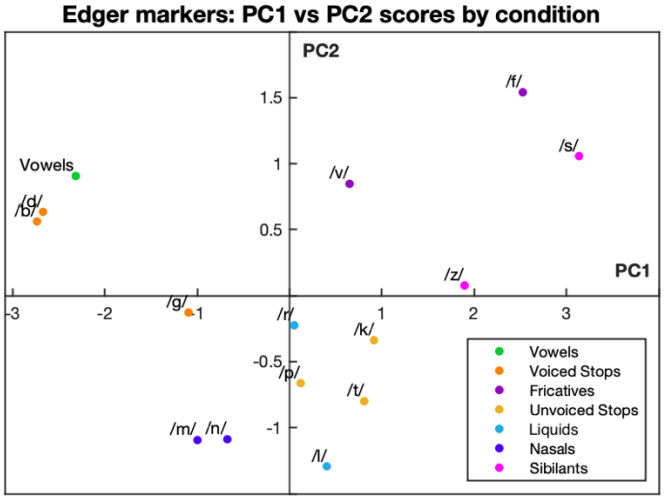


Figure 3. 11. Conducting PCA on the six edge markers (Doelling sharpness, normalised Doelling sharpness, latency of maximum amplitude, peak derivative, latency of peak derivative and the Gini index) of each syllable resulted in different component scores for each syllables A**.** Scores of the first two principal components (PC1 and PC2) are plotted for each syllable. Different colours represent different phonetic groups, identified by the manner of articulation (see legend in B). PC1 values correspond to the x-axis, and PC2, on the y-axis B. The scores of PC1 and PC2 are averaged over syllables pertaining to their corresponding conditions (i.e., syllables which are vowel only, syllables which start with [b], etc.). There were 15 such conditions. Colours represent the same phonetic groups as in A.

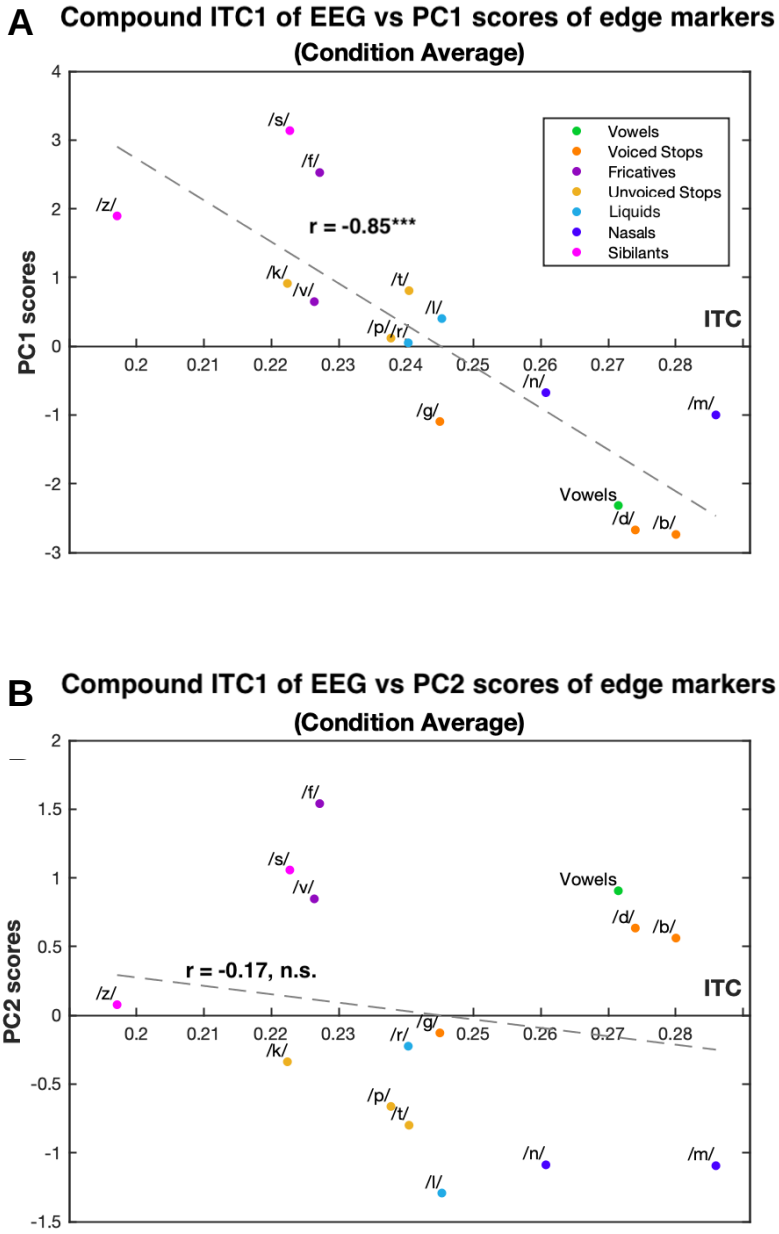


Figure 3. 12. A. The values of EEG Compound ITC1 are plotted against the PC1 scores corresponding to different edge markers of the syllables. Both Compound ITC1 values and PC2 edge markers are averaged over phonemic condition. B. The same Compound ITC1 values are plotted against the averaged PC2 scores of the syllabic edge markers. Each colour corresponds to a different condition (see legend in A). In both figures, interrupted lines represent lines of best fit. Pearson's correlation scores are given above these lines. *** represents p<.001, n.s., not significant.

The PC1 loadings of each edge marker given in Table 3.4 revealed that, while all edge markers had moderate effects on this component, the latencies of the peak derivative and the maximum amplitude and, most of all, normalised sharpness, stand out as the top predictors. On the other hand, the size of the peak derivative and the Gini Index had a slightly smaller influence on PC1. Combined with the results from

the correlation between the PC1 of edge markers and compound ITC1, these results seem to suggest that greater sharpness (both Doelling and normalised), as well as later peaks of the envelope and peak derivative, led to less entrainment.

While this could imply that sharpness may not be the best predictor for entrainment, or that this measure may, in fact, represent something else, it is also possible that a different factor may account for the differences, or, in fact, similarities, between all the considered measurements of edge, one which has not been yet explored. Loadings of each measurement for PC2, where the Gini Index and the size of the peak derivative have the highest influence, tell a similar story, in the sense that syllables starting with nasal consonants have negative scores, so possibly low Gini and height of peak rate. Nevertheless, they show relatively high phase locking as indicated by the ITC, again suggesting that the ascending slopes of the envelope do not need to be steep for entrainment to happen successfully, but actually, they might need to be low. The lack of correlation between the ITC and the maximum amplitude suggest, though, that only the latency of potential landmarks (peak rate, peak envelope) may predict successful entrainment. Furthermore, the correlation between the ITC and PC1 scores is only slightly higher than that between the ITC and the time of the peak derivative, suggesting that the latter may be indeed a potential clue.

Relationship between Peak Derivative and CV transition

Oganian and Chang (2018) suggested that the peak derivative may reflect the information-rich properties of the CV or formant transitions present in a syllable. In order to determine whether the CV transitions of syllables could be potential landmarks for entrainment, we correlated the latency of the peak derivative with the times of the CV transitions which we extracted manually for each CV syllable. We

obtained a significant, but moderate positive correlation between the latencies of the peak derivatives of CV syllables and the manually extracted CVs (Pearson's r=0.67, p<.001), with the mean difference between measures being -0.44 ms, the median -10.43 ms and the standard deviation 22.64 ms. The differences between peak derivatives and manually extracted CVs were considerable at times, and Figure 3.13 indicates that this was mostly true for syllables containing consonants such as "l", "m", "n", as well as "r" and "z", to a lesser extent. However, the boundary between a nasal or liquid consonant consonants and a subsequent vowel may be less clear: the resonant frequencies which are present in a CV syllable containing such consonants may indicate that a CV transition cannot be identifiable in a single point. This factor may explain the large differences (up to 60 ms) between the latency of the peak derivative and the manually extracted CV transitions.

The inconsistency between these two measures could indicate that manual extraction of CV transitions as a single point in time may not be recommended or informative for nasal or liquid consonants. Furthermore, the correlation between the Compound ITC1 and CV transitions did not reach significance (Pearson's r = - 0.50, p=n.s.). However, the values of the latencies of the peak derivatives were rather low for syllables containing nasal consonants at their onsets (as low as 11 ms), and it is unlikely that these latencies correspond to the start of the vowels in such syllables. Therefore, while we cannot argue that the peak derivative indeed measures the highest rate of change in a syllable, and that it may be a crucial factor for entrainment, it remains unclear whether this always corresponds to formant transitions.

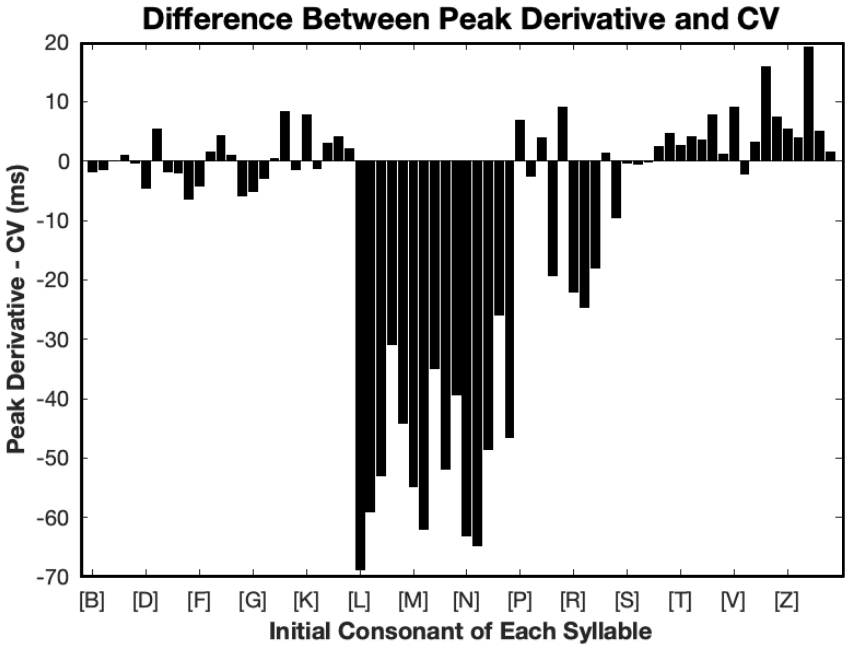


Figure 3. 13**.** Differences between the latency of the peak derivative and that of the manually extracted consonant-vowel (CV) transition. Each bar corresponds to the difference calculated for an individual syllable. The X-axis indicates the initial consonant of every five consecutive syllables, i.e., the first five bars correspond to syllables starting with "[b]" (e.g. /ba/, /be/, /bi/, /bo/, /bu/), the next five syllables start with "[d]", etc.

Discussion

In the present experiment, we showed, for the first time, that phonemic differences in the syllable-initial consonants led to differences in phase locking to CV syllables. Importantly, our syllables were almost isochronous, but peaks in entrainment were obtained at exactly 4 Hz (or harmonic frequencies) nonetheless. The fact that micro-variations in syllabic durations did not seem to matter indicate that the brain may indeed respond to an average frequency of stimulation, which is crucial in tracking a quasi-periodic stimulus such as speech. However, we note that our results refer to evoked rather than induced power, so, despite the positive findings in phase coherence measured by the ITC (normally considered as evidence for entrainment) , there is a possibility that our findings may not necessarily be an instance of synchronicity of endogenous oscillations. Despite this cautionary tale, we emphasise that the aim of the present experiment was to investigate how the brain responds to

different syllabic landmarks as given by their phonemic content, so evidence for ‘entrainment in the narrow sense’ was not a priority of the present study.

Both the phase coherence and evoked power showed peaks at 4 Hz, but harmonic responses were also seen in the ITC, at 8, 12 and 16 Hz. Because harmonic peaks are a natural consequence of applying the Fourier transform and reflect the responses to the frequency of stimulation (Zhou et al., 2016), they were investigated in further analyses. Indeed, we found that the ITC at 4 Hz and the values at harmonic peaks were correlated between each other. PCA was conducted on these ITC responses in order to obtain new uncorrelated variables, which were linear combinations of the first ones. Decorrelating the variables allowed us to better identify the best possible combination of ITC responses which was able to explain the data. We found that the first PCA component (Compound ITC1) best reflected the positive correlation between the ITC at 4 and 8 Hz, while Compound ITC2 was able to account for variation in a direction orthogonal to this correlation.

Subsequently, we grouped responses based on different phonemic groups, including vowels and consonant groups split into manner of articulation and voicing. Significant group differences were mainly found in the Compound ITC1, and showed effects of both manner of articulation and voicing. Thus, the least phase locking as given by the Compound ITC1 was found in sibilants, where responses were significantly smaller than in most other consonant groups. These were closely followed by fricatives, which showed less entrainment than nasal consonants. Conversely, stop consonants, nasals and vowels showed the most entrainment, but there was no differences between these groups. Voicing showed opposite effects between stop consonants and fricatives, in the sense that voiced stops triggered more Compound ITC1 than unvoiced stops, but voiced fricatives led to less

entrainment than unvoiced stops. Furthermore, responses to unvoiced stops were significantly different than voiced stops only, but not than other phonemic groups. At the level of Compound ITC2, we only found a significant difference between liquids and nasals, the latter triggering more phase locking. These results suggest that different phonemic features led to differences in syllable tracking, but because group differences were sometimes ambiguous, we cannot make any certain claims about the nature of features which are preferred by the brain for syllabic entrainment. The large number of phonemic conditions used in the comparisons almost certainly also prevented us from finding more significance between conditions.

We further investigated the different features of our syllables based on different edge markers. Doelling et al. (2014) proposed that one such marker was sharpness, or the positive derivative of the summed narrowband envelope of the stimulus. In Experiment 1, we also introduced a normalised version of sharpness, arguing that compared to the stimuli used by Doelling et al. (2014), natural stimuli could not show such dramatic differences in envelope rises. Indeed, the stimuli in Experiment 1 only differed in their amount of normalised sharpness. Nonetheless, because the positive derivative measures all rises in the syllabic envelope, we argued that sharpness is therefore less informative about the existence of discrete landmarks. Consequently, we added the maximum amplitude of the envelope and its latency to our range of edge markers. Furthermore, we measured the peak derivative of the broadband envelope and its latency, as suggested by Oganian and Chang (2018), who found entrainment to speech to be the highest for this specific landmark. We also added the Gini Index of the envelope, through which we aimed to describe the distribution of landmarks within the syllable. All these were calculated

for the envelope of each syllable, then averaged together for all five syllables within a given condition.

The values of the edge markers all showed to be correlated with each other. One surprising finding was that both the Doelling and the normalised sharpness were negatively correlated with the latencies of the peak envelope and that of the peak derivative. This implied that syllables with slower envelope rises were sharper, which was contrary to our predictions. Subsequently, we used PCA in order to decorrelate these and obtain new orthogonal variables which we could then individually explore as best predictors of entrainment. The maximum amplitude of the envelope was excluded from the PCA because the correlations between this and the other variables were inconclusive.

PC1 and PC2 of this analyses showed a separation of syllables based on phonemic features. For example, vowels, voiced stops and unvoiced fricatives show large PC1 scores, with the first two groups showing positive scores, while unvoiced fricatives only show positive PC1 scores. The latter also seem to have high PC2 scores. On the other hand, unvoiced stops, nasals, liquids and voiced fricatives show intermediate values of both PC1 and PC2 scores, with voiced fricatives again showing only positive scores on both components.

The Compound ITC1 was correlated with both edge markers and edge marker PCA components. As expected, there was a negative correlation between the ITC and the latencies of the peak derivative and the peak envelope, indicating that shorter, more abrupt envelope rises led to more entrainment. However, there was negative correlation between both measures of sharpness and Compound ITC1. Similarly, a negative correlation was found between the latter and the edge marker PC1. These findings support entrainment results, in the sense that, for example, stop

consonants led to more entrainment than sibilants. Nonetheless, it remains unclear why, in this experiment, we found that syllables with lower sharpness or PC1 edge marker scores led to more entrainment. It is possible, however, that both Doelling and normalised sharpness may not be clear indicators of syllabic landmarks, especially not for natural stimuli, which may contain positive slopes (sharpness) in the coda of the syllables as well. These other positive slopes may add noise to the overall measurement of the sharpness of natural syllables. Furthermore, due to their slightly lower PC1 scores, the Gini Index and the size of the peak derivative may also not be considered crucial for neural entrainment to the speech envelope. However, other measures, such as the peak derivative of the broadband envelope could explain edge tracking in a more accurate manner.

The peak derivative describes the highest rate of change in the envelope, and is thought to correspond to the acoustic changes between consonant and vowel parts of the syllable (Oganian and Chang, 2018). In order to test this, we also extracted the latencies of CV transitions for each of our syllables, and compared them to the latencies of the peak derivatives. There was a positive correlation between the two, and the differences between the two latencies were generally small for most syllables, apart from the ones starting with nasal and liquid consonants. As we discussed, this could be because the presence of formants in such consonants makes the border between them and adjacent vowels more difficult to identify. Therefore, while we cannot be sure that the peak derivative indeed corresponds to formant transitions, the latter remain to be investigated as potential landmarks which may be crucial for the neural tracking of speech.

One possible reason for which we did not find more distinctions between consonant groups, in both entrainment and edge markers, was because we used a

range of different vowels in our stimuli. This could have led to a spread in the neural responses based on the differences in responses to such vowels. Furthermore, vowels are thought to be some of the main landmarks which drive neural entrainment (Ghitza, 2011). Some support for this theory comes from research investigating speech comprehension: Fogerty & Kewley-Port (2009) showed that vowels, but not consonants, led to major impairments in comprehension. It is thus possible that by not studying the different vowels in isolation, we minimised the chances of finding more significant group differences in neural entrainment to the sound envelope.

However, research focusing on speech synchronisation between individuals indicates that speakers align their utterances with the onsets of syllables, but not with other syllabic locations (Włodarczak, Sˇimko, & Wagner, 2012). While our study seemed to find more evidence in favour of syllabic onsets as primary edge markers, because of the correlations we found between entrainment and the peak derivative of the envelope, it remains to be determined what the roles of envelope peaks and peak derivatives are, or how consonants and vowels affect syllabic entrainment.

Another possible limitation which may have led to ambiguity between consonant group differentiation was averaging over all EEG channels when calculating entrainment measures, so over the entire scalp. This was done because we used a small number of electrodes, which made source localisation difficult. Nonetheless, it would be useful to investigate whether different parts of the brain show distinct entrainment to CV syllables starting with different phonemes. Arsenault & Buchsbaum (2015) showed that regions of phonemic processing are not identical between the left and the right auditory cortex, but whether this would reflect in syllabic processing has not yet been researched.

Previous research showed that neural responses to different phonemic features cluster in discrete regions (Arsenault & Buchsbaum, 2015; Khalighinejad et al., 2017; Mesgarani et al., 2014). Furthermore, it seems adding phonemic feature information to the broadband envelope of speech improves the correlation between these and neural responses to speech (Di Liberto & Lalor, 2017). Future research could investigate how the mechanisms responsible for phonemic processing are involved in syllable tracking.

In the past, the degradation of phonemic content was found to lead to reductions in speech comprehension (Apoux & Bacon, 2008; Fogerty & Kewley-Port, 2009). Such results were obtained either by filtering our specific frequency bands, for example, between 8 and 16 Hz (Apoux & Bacon, 2008). Frequencies between 8 and 16 Hz are attributed with the intelligibility of stop consonants (Drullman et al., 1994 a,b). However, phonemes are sometimes associated with the fine structure of speech, which specifically refers to frequencies above 100 Hz; alterations in the fine structure of speech can also lead to reductions in comprehension, as well as entrainment (Ding et al., 2014; Zoefel & VanRullen, 2015). The role of phonemic information in speech comprehension, as well as syllabic or envelope entrainment, is therefore unclear, based on previous findings. However, the present study suggests that phonemes impact not only specific frequency ranges within the envelope, but its entire shape. This includes an effect on the broadband speech envelope, which only contains frequencies below 10 Hz. This is reflected especially at the levels of the peak envelope and the peak derivative. Future research could investigate the specific role of such edge markers not just in syllabic entrainment, but also in comprehension.

The findings of Experiment 2 showed that isochronous syllables comprising of natural sounds lead to differences in neural tracking, possibly because they provide different landmarks or edge intensities, as a result of differences in phonetic features, such as voicing or the manner of articulation. Amongst these edge markers, the latencies of the peak derivative and that of the peak envelope distinguished themselves as the most important ones. This can be attributed to them showing the expected relationship with entrainment as given by Compound ITC1, as well as being able to best explain the variability in stimulus conditions. Because of its relationship to the peak derivative, we suggested the CV transition of syllables to be a crucial landmark for speech tracking, but this cannot be determined on the present evidence alone. On the other hand, both the Doelling and normalised sharpness seemed to have an unexpected role on entrainment, with “sharper” stimuli leading to less entrainment. Consequently, we argued that sharpness may be a less reliable edge marker. We also excluded the Gini Index, and the amplitudes of the peak derivative and the peak envelope as primary landmark indicators. In Experiment 3, we investigate the role of specific landmarks in phase locking to speech sounds, with a particular reference to CV transitions and the maximum amplitude of the envelope.

4. The role of syllabic landmarks in neural entrainment to speech

Experiment 3A

Introduction

In Experiment 2, we investigated the roles of phonemic features and their related acoustic edge markers on neural entrainment to the syllabic rhythm in speech. We found that different syllable-initial phonemes led to differences in phase locking to isochronous syllabic streams and that the separation of neural responses depended on the phonemic group, as well as the manner of articulation and voicing of the consonants. A variety of different measures were used to describe the features of the sound envelope, these included measures which quantified the sharpness of the edge (Doelling, normalized) along with the peak latency and the maximum amplitude. These measures all showed correlation with the degree of neural entrainment. Most of the edge markers were also correlated with each other and we conducted PCA in order to obtain a new orthogonal basis which explained the stimulus landmarks. The syllabic conditions showed some clustering based on the first two components of the PCA and the distinctions between the PCA clusters seemed to depend on both voicing and manner of articulation. This suggests that phonemic categories differ in the quality of acoustic landmark which they provide.

Correlations between entrainment and the peak derivative or the maximum amplitude of the envelope indicate that these two discrete locations are potential landmarks for the neural tracking of the syllabic rhythm. Acoustically, the maximum amplitude of the envelope corresponds to vowel peaks (Stuart et. al, 1992) and the

peak derivative, which measures the highest rate of change in the envelope, is associated with formant transitions between a consonant sequence and the adjacent vowel (Oganian and Chang, 2018). We also found a moderate correlation between the latencies of the peak derivatives of each syllable and those of their manually extracted CV transitions, which indicated that there may be indeed a connection between the two.

In the present experiment, we explored different syllabic locations as potential landmarks for neural phase locking to the low frequencies in speech. Suggestions for such landmarks come from research investigating the perceptual correlates of syllabic parsing, or P-centres (Morton et al., 1976). Specifically, P-centres refer to the perceived onset of words or syllables and are thought to evoke the understanding of rhythm or regularity in speech and music (Marcus, 1981). Initial studies of P-centres were concerned with the subjective experience of syllabic timing (Rapp, 1971) or the 'location of rhythmic stress beats' (Allen, 1972), and the term itself, an abbreviation for 'perceptual centre', was not coined until later on, when Morton et al. (1976) reported that perceiving a succession of events to occur at regular intervals did not depend on the absolute regularity of these events.

Common experiments on P-centres involve modifying the timing of syllables so that they occur in a perceptually isochronous fashion (the 'rhythm adjustment method', e.g., Morton et al., 1976; Marcus, 1981), repeating words or syllables in line with a metronome beat ('metronome synchronisation': Šturm & Volín, 2016), or tapping to a sequence of isochronous items (the 'tap asynchrony' method; Vos, Mates, & van Kruysbergen, 1995). In this way, synchrony (or entrainment) lies at the basis of P-centre research. Psychological synchrony to an external event occurring at regular time intervals is thought to be achieved after the successful detection of

the beginning of an event (Villing, 2004). Thus, P-centres may be reflective of the mechanisms involved in neural entrainment to regular events and they may represent essential landmarks in the context of syllabic tracking.

However, the methodologies employed by the P-centre research impose certain limitations on this claim. The 'rhythm adjustment method' is considered tiring and difficult for participants, as the reliability of their judgements seems to decrease after prolonged testing (Villing, Repp, Ward, & Timoney, 2011). Furthermore, this method relies drastically on the internal timing of isochrony, which may not be related directly to syllabic parsing. On the other hand, synchrony experiments, and predominantly those involving tapping, measure the participants' motor actions. This implies that the observed behaviour is a consequence of entrainment in the motor system and not directly in the auditory one, although clear links between auditory and speech areas during both the perception and production of speech exist (Keitel et al., 2017). Tapping experiments also have the disadvantage of showing what is known as 'negative asynchrony', in the sense that taps tend to precede the onset of the sound (Vos et al., 1995), and have also shown larger variability between participants (Villing et al., 2011). Nonetheless, despite these limitations, the P-centre literature reports surprisingly consistent findings across methodologies, shows generally little individual variability (Villing et al., 2011) and P-centre identification seems to be independent of the acoustic or semantic context (Morton et al., 1976).

In general, P-centres seem to depend mostly on the syllabic features present at the onset of the syllable, and this seems to be consistent across languages (English: Harsin, 1997; Brazilian Portuguese: Barbosa, Arantes, Meireles and Vieira, 2005; Czech: Šturm and Volín, 2016). The location of P-centres has been found to be closer to the acoustic onset of the vowel (Morton et al., 1976), but this also

depends on factors such as the duration of the initial consonant (Marcus, 1981), its manner of articulation and voicing (Harsin, 1997), but also envelope features (Howell, 1988a). For example, P-centres were found at earlier latencies for voiced than for unvoiced plosives, or for fricatives and liquids compared to nasals (Harsin, 1997). Similarly, they were located more towards the onset if the vocalic energy occurred early in the syllable and more towards the coda if this energy occurred later (Howell, 1988a). Moreover, the variation in P-centre localisation was reduced if the rise times of the syllabic amplitude were short and abrupt rather than long or complex (Villing et al., 2011).

Some studies also found that P-centres were located later than the beginning of the vowel for syllables with longer initial consonants, as well as longer vowels or final consonants (English: Cooper, Whalen and Fowler, 1986; Brazilian Portuguese: Barbosa et al., 2005). The role of the coda in P-centre identification also seemed to be dependent on the language: for example, adding a sonorant at the end of a syllable delayed the localisation of the P-centre in Czech (Šturm & Volín, 2016), whereas this effect of the final consonant was not reported by studies conducted in other languages. Nonetheless, the P-centre seems to be associated most with the onset of the vowel or the consonant-vowel transition, but there is never a perfect overlap. Moreover, one tapping study in Cantonese found that, depending on the initial consonant of the syllable, participants' responses were closer to either the start of the vowel, the middle of the vowel, or even the start of the consonant (Chow, Belik, Tran and Brown, 2014). Researchers claim that the exact location of the P-centre is difficult to determine due to acoustic and phonemic influences, as well as methodological limitations of P-centre research.

However, the relationship between perceived syllabic onsets and neural tracking of speech has not been explored, even though it is possible that P-centre identification may arise as a consequence of neural entrainment to isochronous syllables. In this case, the unknown issue would be why P-centres are localised as closer to some syllabic landmarks more than other, and whether this is informative either for neural entrainment to speech, for perception mechanisms, or both. Furthermore, the role of precise syllabic landmarks in phase locking to speech has not been investigated, with the exception of the peak derivative, as found in Oganian and Chang (2018). In the present study, we considered the absolute onset of syllables, as well as the points of the CV transitions and vowel peaks, and their roles in neural entrainment.

We created stimuli comprising of two syllables, which had minimally audible noise added at one of three locations: onset, CV transition or the maximum amplitude of the envelope. The two syllables were “da” and “ta”, which show a pair of voiced/unvoiced plosives at their respective onsets. The unvoiced stop /t/ is longer than the voiced /d/, thus allowing us to explore both effects of voicing and duration of the syllable-initial consonant on entrainment.

Two types of noise were added: a click, or a single point in the waveform that was higher than the maximum amplitude of the syllable, and a 5 ms snippet of white noise. Both types of noise were barely detectable, as determined by the experimenters. We hypothesised that the click could enhance entrainment, because its maximum amplitude was higher than that of the syllable, which could be seen in the envelope of thus-modified syllables, and also because it left the original acoustic waveform intact, acting only as an additional landmark. Thus, there would be no differences between conditions containing the click, unless the latency at which this

was placed was itself of significance. On the other hand, the white noise was noticeable in the acoustic waveform, but not in the envelope of the modified syllables. We predicted that this may lead to a reduction in entrainment, especially if the noise masked a particularly important landmark. The white noise was also less detectable than the click, as concluded between the experimenters, but we did not exclude the possibility that this may also lead to a rise in entrainment in cases where participants did notice it as an additional landmark. Nonetheless, in line with findings from Oganian and Chang (2018) and the P-centre research, we expected effects to be stronger at the CV transitions if this is indeed the preferred landmark for neural entrainment, as opposed to other syllabic locations.

Methods

Participants

Sixteen participants (six males, mean age = 25.1 years, SD = 2.1 years) were recruited for the experiment via University of Bristol student groups on social media and compensated for their time (£10/hour). They were all right-handed native English speakers, without any neurological, language-related or sensory impairments.

Stimuli

Stimuli were repetitions of two CV syllables, “da” and “ta”, which we recorded for Experiment 2. We shortened the duration of each syllable using the gammatone filter procedure described in the previous experiment. Because all the stimuli in Experiment 2 contained syllables with different vowels, and thus, each stimulus contained syllables of different lengths, we also expected that the differences in duration between the syllables across the stimuli would not impact the results.

However, because in this experiment we only used two syllables, it was important that these were of identical length, in order to control for possible effects of duration. Subsequently, the total time of each syllable was adjusted manually to 250 ms, either by inserting a few ms of silence, or by removing the last few ms of the vowel. The average intensity of each syllable was changed to 70 dB.

The latencies of CV transitions were obtained manually for each of the processed syllables, using the procedure outlined in Experiment 2. The latencies of the absolute maximum amplitude of the vowel and that of the syllabic onset (located just before the start of the waveform) were also extracted for each syllable, as temporal landmarks where the noise was inserted.

We used two types of noise for this experiment: a click, a single positive change in amplitude at the zero-crossing nearest to the landmark, or a 5 ms snippet of white noise, which contained a multitude of audible frequencies. The amplitude and intensity of the noise were minimal, yet audible, as determined between three experimenters. The minimum and maximum amplitudes of “da” and “ta” were -0.27 and 0.29 arbitrary units (a.u.), respectively. The amplitude of the click was 0.4 and the intensity of the white noise was 45 dB, with its amplitude being negligible in comparison with those of the syllables. The waveforms of the white noise and click were then added to the waveforms of the syllables, at each location of interest (onset, CV, maximum amplitude). Adding noise to the syllables slightly changed the pitch from 70 dB (maximum 70.017 dB, for “Da Click Maximum Amplitude”). Control stimuli (“da” and “ta” without any modifications) were also used, in order to reliably test the effects of the noise on each syllable.

Fourteen stimuli were created for each condition: “da click”, “ta click”, “da white noise”, “ta white noise”, each with three locations of landmark alterations

(onset, CV, maximum amplitude), and the two control conditions, “da” and “ta”. The stimuli were 10 s long and were constructed as 40 repetitions of the same syllable. Each stimulus was repeated 10 times. To maintain the attention of participants, we also created ten filler stimuli, by inserting a syllable starting with a different consonant (e.g., “fa”) in a “da” or “ta” stream. The mismatch syllables were always added randomly in the second half of the stimulus, to ensure participants paid attention until the end of each sound presentation. Stimulus examples (including fillers) are provided in Appendix 4.A.1.

Apparatus

The apparatus was identical to Experiments 1 and 2.

Design

The experiment was within-subjects, with a 2 (consonant) x 2 (noise) x 3 (location) design, and two control conditions. The independent variables were consonant (two levels, “da” and “ta”) , noise type (two levels, “click” and “white noise”) and the location of the noise (three levels, “onset”, “CV” and “maximum amplitude”). The dependent variables were the inter-trial phase coherence and the evoked power.

Procedure

The experiment was split into seven blocks and lasted approximately 40 minutes. One block contained one or two fillers and ten repetitions of two of the main experimental stimuli. There were four blocks with 21 and three blocks with 22 stimuli in total. Each block showed stimuli from a “da” and a “ta” condition. The presentation of the stimuli was quasi-randomised both within and across blocks, so that no two

participants would listen to the same stimulus order within a given block, or the same order of conditions across blocks. Each block lasted approximately 5 minutes, and 20 second breaks were inserted between blocks.

As in Experiment 2, participants saw a message on the screen in front of them after each filler, asking them to type in the different syllable. Because we wanted to maximise the amount of time that participants saw this message, and to make the number of messages equal across most blocks, the question appeared on ten other occasions, after the main stimuli and not the fillers. Participants were instructed to type “none” if they did not hear a different syllable. In blocks containing 22 stimuli, the question followed a main stimulus only once, and in three of the blocks with 21 stimuli, the message was shown once after the main stimulus. There was one block which contained a single filler and where the question only appeared once after the main stimulus.

The main stimuli followed by a message were always control stimuli, because we considered that the noise might alter the perception of the syllable-initial consonant, especially after prolonged exposure to the same stimulus. Also, because the noise was barely detectable, we believed the altered perception of some consonants may happen for some but not all syllables in a given stream. Thus, participants would be sure that a control stream did not contain a different syllable.

Data Analysis

Edge markers

In this experiment we investigated the role of different syllabic landmarks on neural speech tracking, but we also aimed to observe how our different manipulations affected different edge markers. For each syllable (control or experimental), we

calculated Doelling sharpness, normalised sharpness, the Gini index, as well as the peak derivative and its latency, in the same ways as described in the summary of the Experiment 2. We expected that the two sharpness measures and the Gini Index to be fairly similar across conditions with the same syllable-initial consonant. Because the peak derivative of the envelope is thought to correspond to the CV transition (Oganian & Chang, 2018), we also expected the latency of peak derivative to be close to that of the manually extracted CV transition, but we predicted that the value of the peak derivative will differ between control and CV-altered conditions.

EEG

The ITC and Evoked power were calculated in the same way as outlined in the Methods section of Experiment 2. Like in Experiment 2, we tested the effects of entrainment by verifying the significance of peaks in the dependent variables with respect to their neighbouring values. Thus, we compared the average EEG (ITC and Evoked Power) at syllable rate and its harmonics with each of the 19 neighbouring bins in either direction (1.9 Hz, total of 3.8 Hz). Due to channels being correlated with each other, we used the non-parametric Wilcoxon signed rank test instead of T-tests, and the significance of probabilities from multiple comparisons were corrected for false discovery rate (FDR). Statistical analyses were conducted using SPSS (IBM Corp. 2017), in order to test differences between conditions.

Results

Stimuli

Figure 4.A.1 illustrates envelopes of syllables used in six of the 14 conditions: unaltered “da” and “ta”, “da click onset”, “da white onset”, “ta click CV” and “ta white

CV". The click is easily noticeable for both consonants and at either location, whereas the white noise is undistinguishable from the rest of the envelope.

Table 4.A.1 provides values of the different edge markers measured for each syllable (peak derivative and its latency, peak envelope and its latency, Doelling sharpness, normalised sharpness, Gini Index). Noise-altered syllables show similar values to their corresponding controls in all the edge markers. However, some modifications can be noticed. The peak derivative was sometimes earlier for altered than for unaltered syllables, including when noise was placed maximum amplitude locations, the two types of sharpness also show small increases or decreases compared to controls, and the latency of the peak envelope is slightly different for syllables containing a click at the maximum amplitude of the acoustic waveform, compared to all other syllables. However, the differences within conditions which contain the same syllable-initial consonant are not as large as between those whose syllables start with different consonants. Consequently, we did not expect differences in envelope properties to affect entrainment.

The time of the peak derivative extracted from control syllables was also remarkably close to the latencies of CV transitions manually determined for both "da" and "ta" (latency of peak derivative for "da": 7.41 ms, manually extracted CV: 8.3 ms; latency of peak derivative for "ta": 45.69 ms, manually extracted CV: 43 ms).

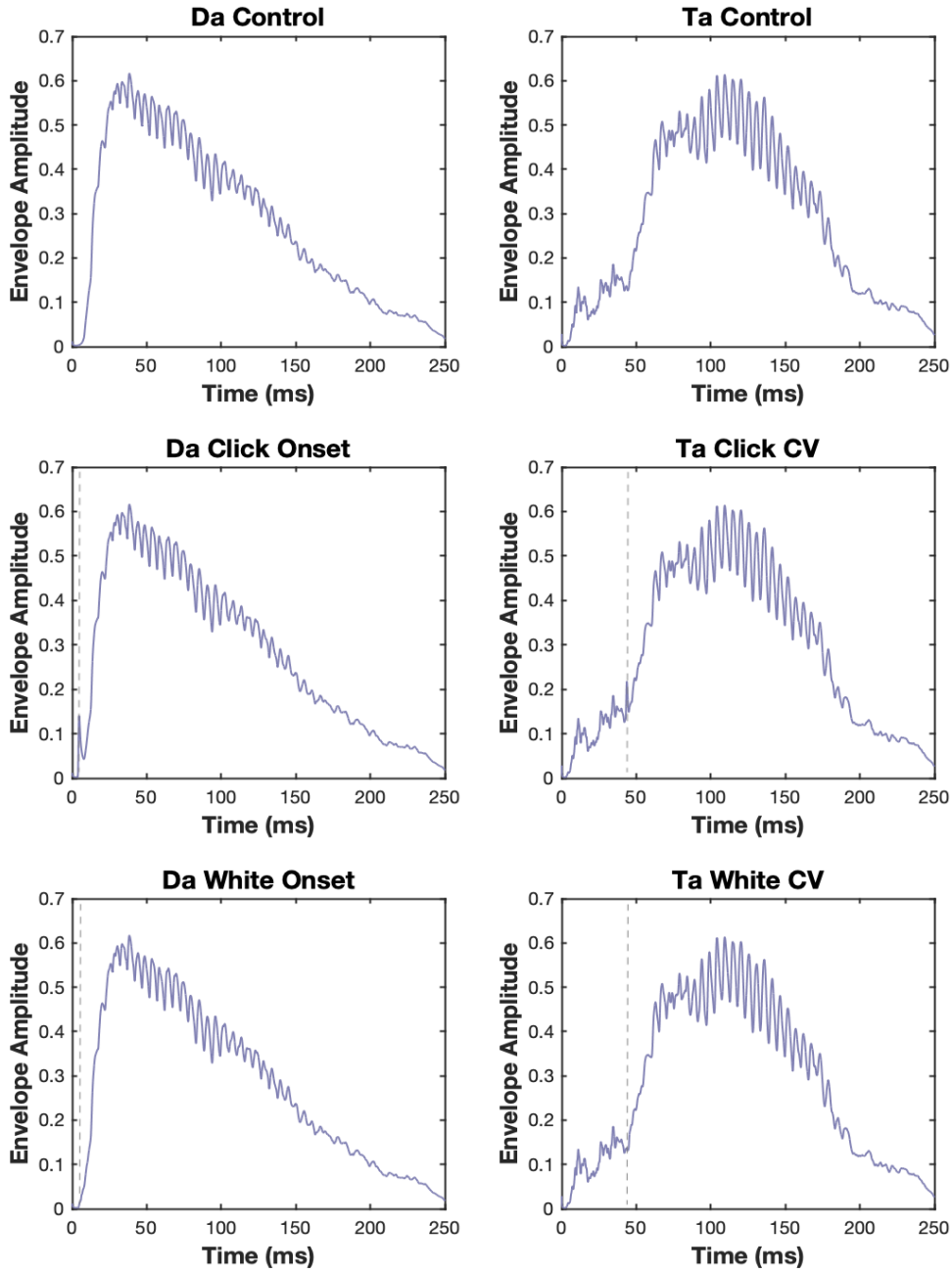


Figure 4.A.1**.**Envelopes of syllables in several different conditions , obtained via the narrowband method described in Experiment 2. Dotted lines represent the location at which noise was added. Note how the click can be seen for both consonants and how the white noise did not affect the shape of the envelope.

Table 4.A.1. Edge markers for each condition. MA = Maximum amplitude of the acoustic waveform.

	Latency of Peak Derivative	*Peak Derivative*	*Latency of Peak Envelope*	*Peak Envelope*	*Doelling Sharpness*	*Normalised sharpness*	*Gini Index*
Da	7.4150	0.00004	38.3447	0.6160	257	0.8305	0.9371
Da Click CV	7.3243	0.00004	38.3447	0.6167	260	0.8370	0.9368
Da Click MA	7.4150	0.00004	27.6417	0.6424	249	0.8032	0.9372
Da Click Onset	7.3243	0.00004	38.3447	0.6155	257	0.8265	0.9365
Da White CV	7.3243	0.00004	38.3447	0.6152	257	0.8298	0.9370
Da White MA	7.4150	0.00004	38.3447	0.6154	253	0.8173	0.9372
Da White MA	7.2562	0.00004	38.3447	0.6163	259	0.8357	0.9369
Ta	45.6916	0.00002	109.0930	0.6130	436	1.4142	0.9362
Ta Click CV	45.3741	0.00002	109.0930	0.6132	445	1.4393	0.9361
Ta Click MA	45.7370	0.00002	109.0930	0.6168	432	1.3988	0.9363
Ta Click Onset	45.7596	0.00002	109.0930	0.6131	431	1.3902	0.9355
Ta White CV	45.4422	0.00002	109.0930	0.6125	438	1.4198	0.9361
Ta White MA	45.7370	0.00002	109.0930	0.6162	428	1.3879	0.9362
Ta White Onset	45.8957	0.00002	109.0930	0.6129	436	1.4119	0.9360

EEG

The peaks of both ITC and evoked power were significantly higher than neighbouring bins at 4, 8, 12 and 16 Hz ($p<.001$, FDR-corrected). Means for both ITC and evoked power can be seen in Figure 4.A.2, with the evoked power graph showing the characteristic alpha between 8-12 Hz. Both entrainment measures show peaks at the syllabic rate and its harmonics. However, the evoked power also showed characteristic alpha activity between 8 and 12 Hz, and therefore these two harmonics and the later one at 16 Hz (which was smaller) were excluded from further analyses.

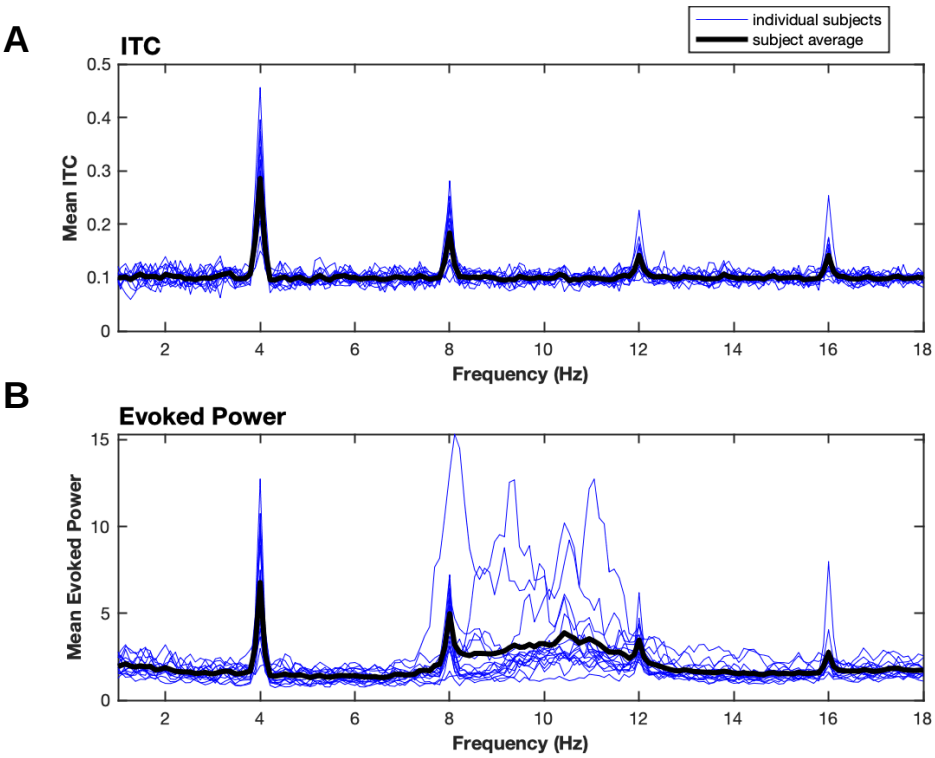


Figure 4.A.2**.** ITC and Evoked Power, averaged over channels and conditions, are plotted as a function of frequency, between 1 and 18 Hz. Bold black lines represents averages over all subjects. Each of the blue lines represents an individual subject. Peaks can be noticed at 4, 8, 12 and 16 Hz in ITC and Evoked Power. A. ITC B. Evoked Power.

There was a positive correlation between the ITC peaks at 4 and 8 Hz (Pearson's $r = 0.55$, $p<.05$), and a positive correlation between the peaks at 12 and 16 Hz (Pearson's $r = 0.66$, $p<.01$).

Like in Experiment 2, we decorrelated the ITC peaks by applying PCA and obtaining linear combinations of the ITC values as new independent bases for entrainment. The factor loadings for each of the four components as well as the amount of variance explained by the latter, are given in Table 4.A.2. For this section, we kept the first component explaining 64.4% of the variance for statistical testing. We called this component Compound ITC1. The factor loadings for Compound ITC1 were 0.88, 0.47, 0.02 and 0.005 for the ITC at 4, 8 , 12 and 16 Hz, respectively, suggesting that the ITC at 4 and 8 Hz had the highest impact.

We conducted a series of repeated measures factorial ANOVAs in order to test the significance of EEG results across conditions (excluding the two controls). We present here the ones for the compound ITC1 and the evoked power at 4 Hz. The tests were run over all 12 conditions, then separately for "da" and "ta", noise type and location. Statistical tests revealed similar results for the 4 Hz ITC, but, like in Experiment 2, we decided PCA components of the ITC would give a more complete picture of the underlying neural processes, as some effects may be carried, to a smaller degree, in the harmonic peaks. The analyses yielded similar results for the other PCA components of the ITC as well and these, along with those for the 4 Hz ITC, are provided in Appendix 4.A.2.

Table 4.A.2**.** Factor loadings for each of the four ITC values (4,8,12, and 16 Hz) are given for each of the four principal components, after PCA. “PC” = principal component. The amount of variance explained by each component is noted below the loadings.

ITC at frequency:	*PC1*	*PC2*	*PC3*	*PC4*
4 Hz	0.882	-0.312	-0.352	-0.021
8 Hz	0.470	0.559	0.683	0.012
12 Hz	0.020	0.441	-0.389	0.809
16 Hz	0.005	0.629	-0.509	-0.587
Variance Explained	64.47%	22.79%	9.48%	3.26%

The consonant x noise type x location ANOVA revealed a main effect of location ($F_{2,30}$= 6.108, p<.01) and a consonant-by-location interaction ($F_{2,30}$= 4.34, p<.05) in the Compound ITC1. Post-hoc tests further showed that the effect of location was driven by the Compound ITC1 in onset conditions being always higher than that in the maximum amplitude, when averaged over consonant condition and noise type (p<.05, Bonferroni). In the 4 Hz evoked power, there was a main effect of consonant ($F_{1,15}$ = 5.53, p<.05), with “da” stimuli generally eliciting higher power than “ta”. However, there was no difference between control “da” and “ta” syllables, as indicated by paired two-tailed t-tests conducted on either ITC or evoked power (p = n.s.).

Noise type x location ANOVAs conducted for “da” syllables revealed a main effect of location in the Compound ITC1 ($F_{2,30}$= 5.09, p<.05). Subsequently, post-hoc tests showed that both the Compound ITC1 and the 4 Hz Evoked Power were greater for syllables with altered onsets than for those with modified maximum amplitudes, when these responses were averaged over noise type (Compound ITC1, 4 Hz Evoked Power: p<.05, Bonferroni). A similar effect of location was found by the Compound ITC1 ANOVA conducted on “ta” syllables ($F_{2,30}$ = 5.04, p<.05), with entrainment being higher to onset than to CV-altered syllables (p<.05, Bonferroni). At

first, this may suggest that onset-altered syllables always led to more entrainment. Nonetheless, when Bonferroni-corrected T-tests were conducted separately for “da click”, “da white noise”, “ta click” and “ta white noise”, no further significant differences were found between the different locations. Onset-altered conditions also did not differ in the amount of entrainment compared to their corresponding controls.

At CV-altered conditions, two consonant x location ANOVAs showed main effects of consonant and noise type in both the 4 Hz evoked power and the Compound ITC1: “da” always led to more entrainment than “ta” (Compound ITC1, $F_{1,15}$ = 7.52, p<.05; 4 Hz Evoked Power, $F_{1,15}$ = 7.94, p<.05), and “click” always more than “white noise” (Compound ITC1, $F_{1,15}$ = 7.78, p<.05; 4 Hz Evoked Power, $F_{1,15}$ = 4.78, p<.05). Furthermore, a series of paired two-tailed T-tests revealed “Da Click CV” showed significantly more entrainment than both “Da Control” (Compound ITC1, t(15) = 2.47, p<.05) whilst the Compound ITC1 to “Ta White CV” was lower than that to “Ta Control” (t(15) = -2.79, p<.05). Therefore, the differences between all conditions were the largest at the latency of the CV transition, which suggests that the CV transition may be more important than the other landmarks in the neural tracking of syllables.

Figure 4.A.3 show how the effects of the experimental manipulation are the most noticeable in the CV conditions, with differences between groups being the highest at this location, smallest in the maximum amplitude condition, and the onset condition falling in between. This was shown for both the Compound ITC1 and the 4 Hz Evoked Power. Effects at CV locations were opposite for “da” and “ta” conditions and also seemed to depend on the noise type. White noise led to a reduction in entrainment for “ta CV” syllables, but the click did not seem to affect phase locking to “ta CV” stimuli with respect to control, or the other click conditions. On the contrary,

the click led to greater entrainment in “da CV” syllables, but white noise did not generate a significant difference between “da CV” streams and other conditions. This could imply that at the CV location, the white noise masked critical information for “ta”, while the click acted as an entrainment enhancer for “da”. A possible interaction of the noise type with the acoustic properties of both “da” and “ta” is evaluated in the discussion of this chapter.

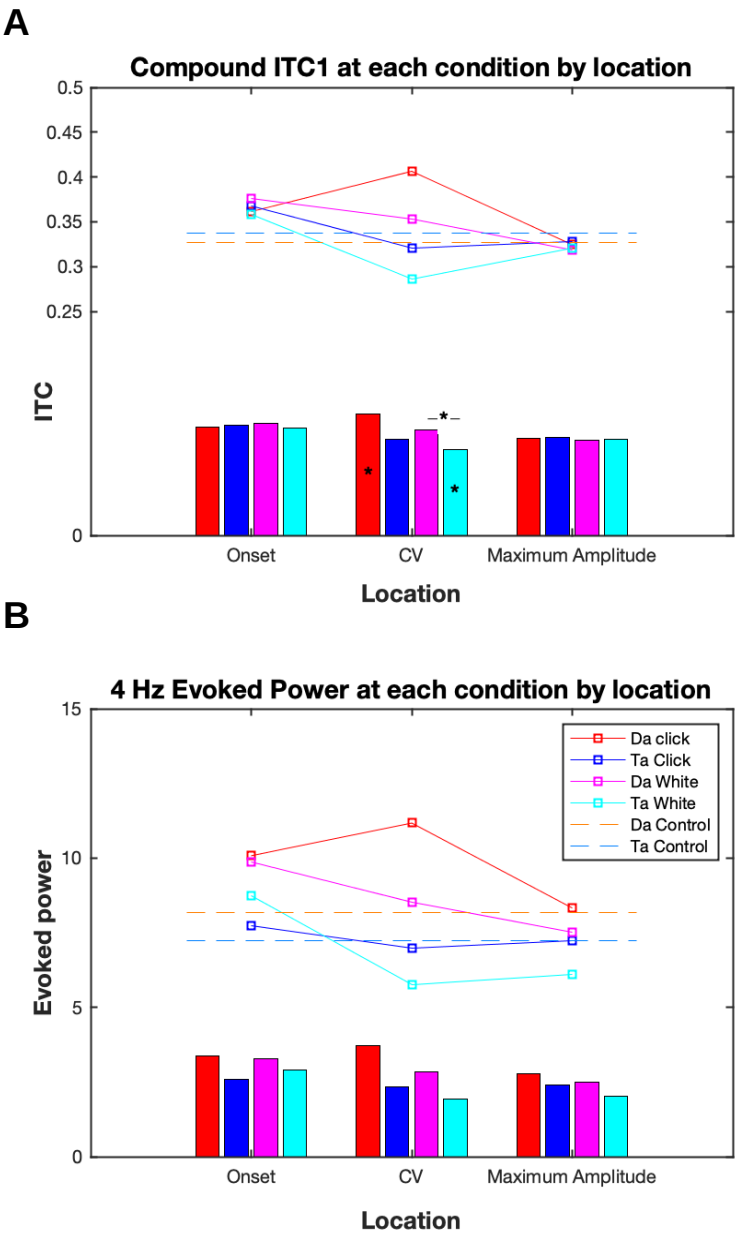


Figure 4.A.3. The bars represent the values of Compound ITC1 and 4 Hz Evoked Power at each syllable-altered condition, and are delimited depending on the “Location” factor. Lines above the bars represent each condition, with solid lines for altered syllables, and also dotted lines for controls. The colour codes for each condition are provided in the legend in B. Stars represent the significance of the comparison between two groups, with stars inside the bars suggesting difference between the bar condition and its respective control, and stars between the bars the significance between their corresponding conditions, The latter are Bonferroni-corrected. * is p<.05.

A. Compound ITC1.

B. Evoked Power.

Discussion

Previous research investigating neural phase locking to speech suggested that, when tracking a stimulus with irregular amplitude fluctuations, the brain tracks specific acoustic landmarks such syllabic onsets or vowel peaks (Doelling et al., 2014; Ghitza, 2013). On the other hand, behavioural entrainment experiments have found that the perceived onsets (P-centres) of isochronous syllables are located closer to the CV transition of the syllables (e.g., Barbosa et al., 2005; Šturm & Volín, 2016). In this experiment, we tested the role of different syllabic landmarks in neural entrainment to speech sounds, by adding noise at one of the three locations of isochronous syllables: onset, CV and maximum amplitude of the sound waveform.

Experiment 3A showed that on average, onset-altered syllables showed more entrainment in both the Compound ITC1 and the 4 Hz Evoked Power, compared to other syllable-altered conditions. When we averaged over noise type, time-frequency responses were higher at onset-altered conditions than at maximum amplitude ones, for “da”, and than CV ones, for “ta”. Moreover, individual onset- and maximum amplitude-altered conditions were not significantly different from their corresponding controls, but they were consistently higher, or lower than control, respectively. There were also no differences due to noise type in entrainment to maximum amplitude and onset-altered stimuli.

The present onset and maximum amplitude results may be explained by findings from the P-centre literature. The P-centre of syllables starting with stop consonants was located closer to the beginning of the vowel (Harsin, 1997), and, in tapping experiments, this was sometimes considerably ahead of the CV transition (Vos et al., 1995). Consequently, a landmark preceding the P-centre, such as noise at the onset of syllables, could have led to more entrainment because it acted as a

predictor of the P-centre. Conversely, a later landmark placed at the maximum amplitude of the syllable may have disrupted the perceived isochrony of the stimuli, as well as entrainment. Nonetheless, this interpretation is constrained by findings from CV conditions, which showed both higher and lower entrainment than onset and maximum amplitude conditions.

Entrainment in the Compound ITC1 and the 4 Hz Evoked Power was highest at “Da Click CV” and lowest at “Ta White CV”. These two conditions were also significantly different from their consonant-relevant controls, as shown by Compound ITC1. Unlike for onset and maximum amplitude conditions, the two types of noise had opposite effects on the two different consonant conditions, when syllables had altered CV transitions. It is possible that in “da” syllables, the click was more noticeable at the CV transition, but also that this may have marked more clearly the onset of the vowel, which is necessary for entrainment.

As seen in Table 4.A.1, the peak derivative is higher and occurs earlier for “da” syllables, which begin with a voiced stop consonant, versus “ta”, which starts with an unvoiced plosive. A higher and earlier peak derivative could imply that the acoustic boundary between the consonant and the vowel is also clearer for “da” than for “ta”. This may explain why the click caused the CV transition in “da”, but not in “ta”, to be more noticeable. While the CV transition is equally important for “ta”, the fact that this occurs later and presents a less abrupt rate of change in envelope information suggests that this could be a less stable landmark for entrainment.

This claim is partially supported by the P-centre literature. It is known that unvoiced plosives last longer than voiced ones, and syllables beginning with longer consonants have later and more variable P-centres than syllables starting with shorter consonants (Villing et al., 2011). However, this variability in P-centre

identification seems to be similar to both voiced and unvoiced plosives, when these are found in the beginning of the syllables (Šturm & Volín, 2016). In Experiment 2, we found that syllables starting with voiced stop consonants triggered more entrainment than those beginning with unvoiced stops. The P-centres of syllables starting with unvoiced stop consonants could therefore be less stable than that of those starting with voiced stops, and the CV transitions of such syllables may also be less reliable landmarks for entrainment.

Consequently, a small click added at the CV transition of “ta” may have not been sufficient in creating a more stable landmark, whereas white noise masking information at this location may have further destabilised it, leading to poorer neural tracking of such syllables. On the other hand, the small amount of noise present in the formant transition of “da” syllables may have not been enough to conceal the strong landmark present in the acoustic information at this location. Nonetheless, it remains unclear to what degree we affected the P-centres of our stimuli and further studies are needed to investigate these hypotheses.

Changes in edge markers due to the manipulation in the present experiment do not seem to account for the findings. For example, the latency of the peak derivative was slightly earlier than control for both “da click CV” and “da white maximum amplitude” syllable, but only the former showed greater phase locking compared to unaltered “da” syllables. Furthermore, “da click CV” and “ta white CV” both presented higher amounts of Doelling sharpness than their corresponding controls, but showed opposite trends in entrainment with respect to each other.

Nonetheless, the present findings could be explained by the noise being more noticeable at certain locations than others, without revealing anything about the acoustic properties of the underlying syllables. Even if the loudness of the click and

that of the white noise were the same within their respective conditions, the differences between their amplitude and the amplitude of the rest of the syllable varied depending with location. For example, the click may have been the most noticeable at “da CV”, which could explain why this condition triggered the most entrainment. On the other hand, the white noise may have been most disruptive at “ta CV”. The placement of white noise at this location may have interfered with the natural P-centre of the syllables, causing them to sound less uniform, and possibly less isochronous. In Experiment 3B, we explored the perceptual effects of noise added at different syllabic locations, by asking participants to rate how noticeable or how disruptive the noise was in each syllable. This experiment allowed us to eliminate potential confounds, such as the perceived loudness of the noise, and to explore more valid explanations for the present results.

Experiment 3B

Introduction

In Experiment 3A, we found that adding noise to isochronous syllables affected entrainment the most when this was placed at the CV transition of each syllable. However, the effects on entrainment depended on both the initial consonant of the syllable (i.e., whether this was /d/ or /t/), as well as the type of noise that was added (click or white noise). “Da” syllables with a click added at the CV transition showed more entrainment when compared to unaltered controls, whereas syllables in the “ta white noise CV” condition showed less phase locking than their matched controls. This may be due to the fact that the CV transition is a crucial landmark for entrainment, compared to the absolute onset and the maximum amplitude of syllables.

Furthermore, the acoustic quality of each landmark at the CV transition may depend on the initial consonant(s) of the syllable, which could explain why the different types of noise affected each syllable in different ways. For example, while the main landmarks for entrainment may be found at the CV transition for both “da” and “ta” syllables, they are different from each other in their acoustic and spectral properties. A clearer boundary between the voiced consonant and the vowel at “da” may have been enhanced by placing a click at exactly its location, whereas a longer, more diffuse boundary between the unvoiced /t/ and /a/ may have been perturbed by a longer segment of white noise.

Nonetheless, we did not exclude that the findings in Experiment 3A were due to perceptual effects of noise. The difference in intensity between the noise and the underlying syllabic waveform was not the same at all locations, implying that the noise may have either been more noticeable or more disruptive at some locations than others. For example, the click may have appeared louder at “da CV”, whereas the white noise could have been more disruptive at “ta CV”. The effects of loudness and disruption of the acoustic content by noise are explored in this experiment.

It is possible that in conditions where the noise was louder or more noticeable, this provided a stronger landmark for entrainment. Even though the effects of landmark intensity on syllabic entrainment have not been investigated, previous studies may support our claims. For example, louder tones were found to trigger more entrainment in the gamma band than tones with of lower intensities (Schadow et al., 2007). Furthermore, Zoefel and VanRullen (2015) added tone pips of identical frequencies and amplitudes on top of speech-noise constructs and found that the tones were easier to detect when the spectral energy of the carrier stimulus was higher rather than lower. Because the noise was uniform, higher spectral energy was

found at the peak envelopes of speech. Equally, we found that pitch intensity was highest when the click was added to the maximum amplitude of the syllables. While this does not perfectly explain why entrainment in Experiment 3A was highest for “da click CV” syllables, it is also possible that, if the noise in this condition was more noticeable, such syllables may have also appeared more stressed. Indeed, some experiments found that an increased amount of stressed syllables present in infant-directed speech leads to more neural entrainment in children, both in the delta and theta bands (Leong et al., 2014, 2017).

On the other hand, the white noise led to a reduction in entrainment at “Ta CV”, implying that the effect of the white noise was different than that of the click. One possibility is that this noise type disrupted the phonemic content of the syllable, in the sense that “Ta white noise CV” syllables sounded less like unaltered “ta” syllables. Cooper et al. (1986) showed that by eliminating portions of the initial consonant of a syllable, one would alter the perceived category of that phoneme as well as the location of its P-centre: e.g., if a certain amount of frication was eliminated from “sha”, this would sound like a “cha” and its P-centre would be identified earlier than for “sha”. It is possible that we altered the locations of our P-centres through the addition of noise and also, because the segment of white noise was longer than that of the click, it is possible that this may have led to different perceptions of the consonant. Altering the location of the P-centre may have also meant that some stimuli were perceived as less isochronous than others. Finally, such disruptions in perception could have resulted in a reduction in entrainment.

In the present experiment, we asked participants to rate the syllables used in Experiment 3A, either by judging how disrupted they were, or how noticeable the noise was, compared to each other, as well as controls. We expected CV-altered

syllables containing white noise to be classified as more disrupted than others, because the noise at these locations masked phonetic information which was crucial for the correct categorisation of speech sounds. In contrast, the information at both the onset and the maximum amplitude of the envelope was more uniform, so syllables containing white noise at this location would sound less disrupted than the other ones. On the other hand, we expected clicks added at the CV of stimuli to be more noticeable than at other locations, because these would emphasise the boundary between the consonant and the vowel.

Methods

Participants

Twelve students from the University of Bristol were recruited via social media groups (five males, mean age = 26.3 years old, SD = 3.5 years). They were reimbursed for their time (£10/hour). All were native English speakers and had no hearing impairments.

Stimuli

We used the same syllables as the ones in each condition described in Experiment 3A: “da” and “ta” with noise added at one of the three locations explored in Experiment 3A (onset, CV, maximum amplitude). The syllables were 250 ms long. Two more conditions were added for each type of noise, such that we used clicks of three levels of sound amplitude (0.4 – original, 0.7 and 1), and white noise with three levels of intensity (45 dB – original, 55 dB and 65 dB). Each of the six noise segments were then added at one of the three locations of either syllable. This resulted in 36 altered syllables, and two controls (the unaltered “da” and “ta”). Each

syllable was repeated three times in the creation of a stimulus. Syllabic repetition was maintained for the present stimuli, in order for the results of the two experiments to be comparable.

The stimuli were presented in pairs, as follows: for each consonant and noise condition, at each level of intensity, a stimulus with an alteration at one location was compared with a stimulus containing the same type of noise at each of the other two possible locations. Each of the 36 altered syllables were also paired with their respective controls. To these, we added comparisons at each location between click and white noise, but only for the levels of noise used in the previous experiment. For example, a “da” with a 0.4 click at CV was paired with a “da” with 45 dB white noise at CV. There were 78 pairs in total: all combinations are provided in Table 4.B.1. “Da” stimuli were never compared with “ta” syllables.

Each pair of stimuli was presented six times per experiment, the order of the stimuli being counterbalanced across all trials. All pairs of stimuli were uploaded online to the link provided in Appendix 4.B.

Apparatus

The experiment was run using PsychoPy 3.1 and Python 3 on a 15-inch retina display MacBook Pro. The script for the experiment can be found in the online Github repository listed in Appendix 4.B. Participants listened to the stimuli through Pioneer headphones, model SEC-MJ101-k. Analyses were conducted in Matlab 2018b.

Table 4.B.1. Table showing all comparisons performed in an experiment (78 in total). For example, for “da click” of 0.4 amplitude, we compared syllables with noise added at onset and CV, onset and MA, CV and MA, after which each syllable with noise was compared with an unaltered control. We also compared the stimuli with 0.4 clicks with those containing 45 dB noise, at each of the three locations: onset, CV and MA. CV = consonant-vowel transitions, MA = maximum amplitude.

Da						***Ta***					
Click			**White noise**			**Click**			**White Noise**		
0.4	0.7	1	45 dB	60 dB	70 dB	0.4	0.7	1	45 dB	60 dB	70 dB
O - CV	O - CV	O - CV	O - CV	O - CV	O - CV	O - CV	O - CV	O - CV	O - CV	O - CV	O - CV
O - MA	O - MA	O - MA	O - MA	O - MA	O - MA	O - MA	O - MA	O - MA	O - MA	O - MA	O - MA
O - Control	O - Control	O - Control	O - Control	O - Control	O - Control	O - Control	O - Control	O - Control	O - Control	O - Control	O - Control
CV - MA	CV - MA	CV - MA	CV - MA	CV - MA	CV - MA	CV - MA	CV - MA	CV - MA	CV - MA	CV - MA	CV - MA
CV - Control	CV - Control	CV - Control	CV - Control	CV - Control	CV - Control	CV - Control	CV - Control	CV - Control	CV - Control	CV - Control	CV - Control
MA - Control	MA - Control	MA - Control	MA - Control	MA - Control	MA - Control	MA - Control	MA - Control	MA - Control	MA - Control	MA - Control	MA - Control
Onset		**CV**		**MA**		**Onset**		**CV**		**MA**	
Click 0.4 – White Noise 45 dB		Click 0.4 – White Noise 45 dB		Click 0.4 – White Noise 45 dB		Click 0.4 – White Noise 45 dB		Click 0.4 – White Noise 45 dB		Click 0.4 – White Noise 45 dB	

Design/Procedure

We created two experimental groups based on the type of comparison that participants made between stimuli. After being presented with a pair of stimuli, the participants in either group answered one of the following questions: “which stimulus contained the more noticeable noise?” or “which stimulus was more disrupted?”. The same question was repeated after each pair of stimuli was presented. There were six participants in each experimental group.

We used a counterbalanced block design: participants were assigned to either group (“noticeable noise”/”disrupted syllable”) in accordance with the participant number, i.e., in an alternating fashion, starting with the “noticeable noise” group. Participants first listened to blocks of stimuli containing only clicks, followed by white noise-only stimuli, or vice versa. This order was counterbalanced across all twelve participants. A block that included mixed pairs with two types of noise was presented at the end of the experiment.

In the “noticeable noise” group, the loudest version of the click was played twice, in isolation, before the click-only blocks, and a 65 dB white noise was presented in a similar way before the “white noise” blocks. The white noise segment was described as “hissing”. The two types of noise were played in advance in order to familiarise the subjects with the sounds which they needed to identify during the trials. Conversely, the participants in the “disrupted syllable” group heard the unaltered versions of the “da” and “ta” syllables before each of the click-only and white noise-only blocks. The original, uncompressed syllables were presented twice each, with a two-second gap between each presentation. Participants in this group did not explicitly know that the experiment contained two different kinds of noise. The

unaltered syllables were presented before each noise type block as a reference point for the experimental stimuli.

Each of the click/white noise blocks were further separated into six subblocks. A subblock contained all possible pairs for both consonants containing a particular type of noise (e.g., “click” or “white noise”), as well as pairs including controls, and lasted approximately three minutes. The order of the pairs within a subblock was randomised. A subblock was repeated three times, after which the same counterbalanced subblock was also played three times, so that each pair of stimuli was played six times. For a given pair, a “1” would appear on the screen during the playback of the first stimulus, and a “2” when the second stimulus was presented. There was a 0.5 second inter-stimulus interval for stimuli within a pair. The relevant question was shown after the second stimulus was played. The participants were instructed to press either “1” or “2” for the stimulus which they thought was more disrupted, or contained the more noticeable noise. The next pair of stimuli was presented 0.5 seconds after an answer was typed.

Participants were given 20-second breaks between subblocks, during which they heard nature sounds and were shown animal pictures, in order to prevent adaptation and exhaustion. The pictures were found through Google searches and were not protected by Copyright. All nature sounds were free to download from the BBC Sound Effects Archive found at http://bbcsfx.acropolis.org.uk/.

The mixed block of trials containing comparisons between the two kinds of noise was presented after the click and white noise blocks. This lasted three minutes and, like the other blocks, contained six presentations of each pair (including counterbalancing). The total duration of the experiment was approximately 40 minutes. The structure of the experimental design is depicted in Figure 4.B.1.

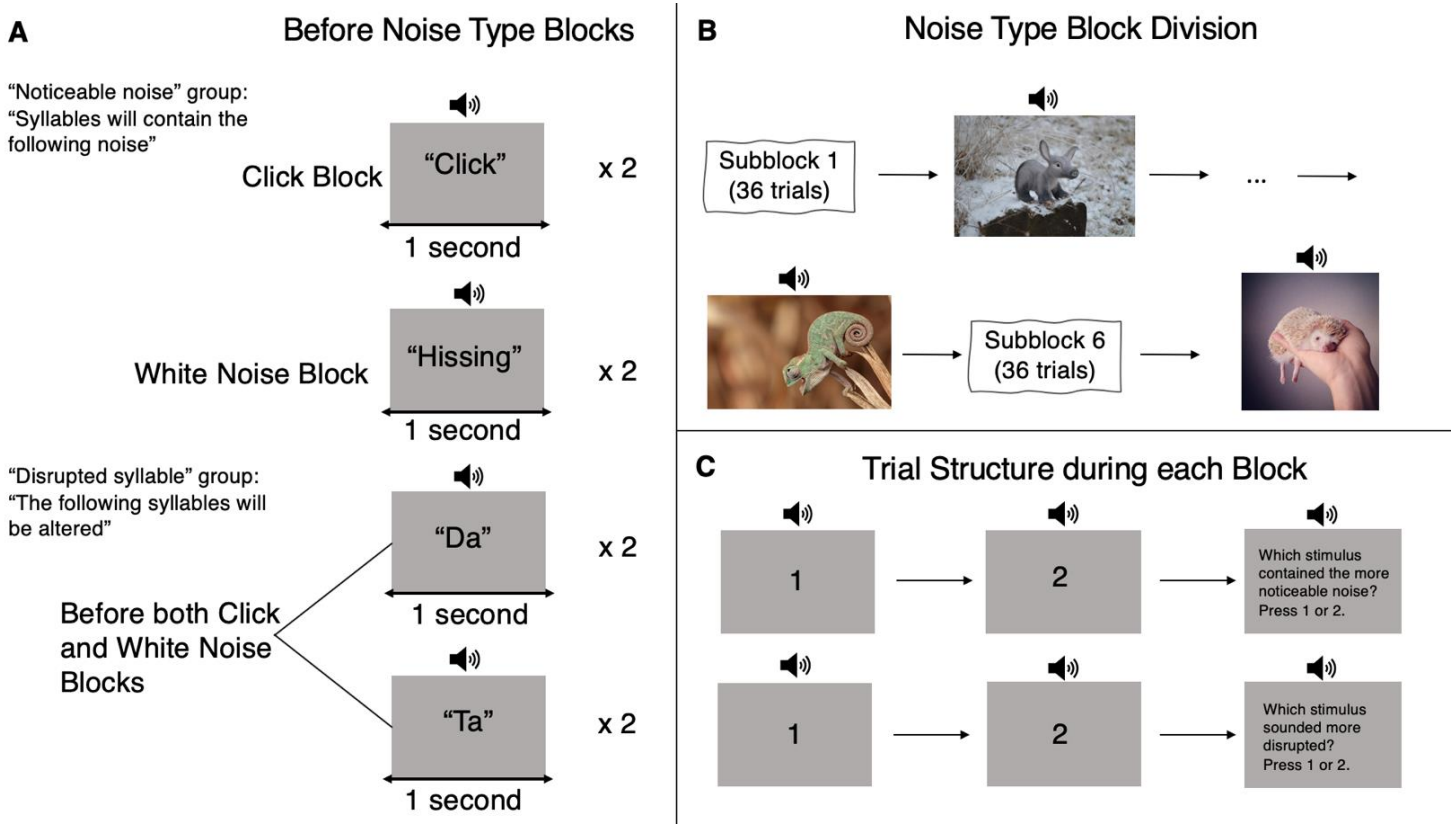


Figure 4.B.1**.** Experimental structure. Grey rectangles represent laptop screens. A. Sounds played before noise type blocks (click-only or white noise-only). In the "noticeable noise" group, the click was played twice before the click block and the white noise was played twice before the white noise block. There was one second of silence between each noise. In the "disrupted syllable" group, each unaltered syllable ("da" or "ta") was presented twice before either the click or white noise groups. There was one second of silence between the presentations of each syllable. B. Each noise type block (click-only or white noise-only) was divided into six sub-blocks containing all 36 trials for that particular noise type, which were presented in a random fashion. Twenty-second long breaks were provided between each sub-block, during which participants saw an animal picture on the screen and listened to a nature sound. C. Trial structure. A "1" was presented on the screen during the playback of the first stimulus in a pair. The second stimulus was played after an inter-trial interval of 0.5 seconds and a "2" appeared on the screen during its presentation. After a second break of 0.5 seconds, the question relevant to each group was shown: for the "noticeable noise" group, participants were asked which of the two stimuli contained the more noticeable noise; in the "disrupted syllable" group, they pressed the key corresponding to the stimulus which they perceived as more disrupted from the syllables they listened to before each noise type block.

Data analysis

To be able to order each stimulus, from the one containing the least noticeable noise to the one with the most noticeable noise, and, respectively, from the least disrupted to the most disrupted stimulus, we calculated the binomial probability of success for a stimulus in each of the pairs. First, we counted the number of times (k) that one of the stimuli was picked over the other one across all presentations of the same pair, including counterbalancing, for all participants in each experimental group. Subsequently, we tested if the probability that a stimulus in a pair was picked exactly k number of times was significant, using the following formula:

$$P = \binom{n}{k} p^k (1-p)^k \quad (4.\mathrm{B}.1),$$

Where P is the binomial probability, p is the probability of success for a stimulus in a single trial (0.5, for a single trial had only two possible outcomes) and n is the total number of trials that a pair was presented. We considered P was significant if this was below .05.

We ordered the triplets in each single syllable/noise condition as follows: for example, for “da click”, where the click was 0.4 in amplitude, the stimulus with noise at onset was considered to be less disrupted (“<”) than the one with noise at CV if the binomial probability that this was picked less than its pair was significant. An exact order was not determined if the binomial probability was not significant: for example, if the count for onset stimuli was less than for CV stimuli, but the probability was above .05, a “<=” was used as notation. The onset stimulus was then compared with the maximum amplitude one, and the remaining comparison between CV and maximum amplitude was also performed, in exactly the same fashion as described above. We finally ordered all three stimuli: for example, onset<CV, onset <MA and

CV<MA resulted in onset<CV<MA. We did not determine triplet rankings if the obtained order was intransitive (e.g., onset<CV, onset>MA, CV<MA). We never ordered stimuli containing different consonants, or across different noise levels. Furthermore, comparisons between altered syllables and non-altered controls, or single location comparisons across two different types of noise, were addressed individually. The same procedure was applied to both experimental groups ("disrupted syllable"/ "noticeable noise").

Results

For each syllable/noise type/noise level combination, we ordered the stimuli altered at onset, CV and maximum amplitude, with respect to each other. This was done based on the binomial probabilities obtained from comparisons between each of the two stimuli in that particular combination. A stimulus was excluded from the ranking if there were non-significant relationships with both of the other two stimuli.
The orders of the stimuli are provided in Table 4.B.2.

Table 4.B.2**.** Rankings of locations deducted from a triplet of comparisons (onset – CV, onset – maximum amplitude, CV – maximum amplitude), for each consonant, at each noise type and level. O = onset, C = CV, M = Maximum amplitude. In the "noticeable noise" group, the ranking C<O<M indicates that CV-altered syllables contained the least noticeable noise, followed by syllables containing noise at onset and maximum amplitude. In the "disrupted syllable group", the same order indicates that CV-altered syllables sounded the least disrupted, followed by onset and maximum amplitude-altered stimuli. A "<" indicates that binomial probabilities between stimuli in a given pair were significant, whereas a "<=" indicates that while the success counts for the stimulus on the left of the inequality sign was less than for the one on the right, their binomial probabilities were not significant.

	Da		*Ta*	
	"Noticeable Noise"	"Disrupted Syllable"	"Noticeable Noise"	"Disrupted Syllable"
				Click
0.4 (a.u.)	C<O<M	C<O<=M	O<C<M	O<M
0.7 (a.u.)	O<=C<M	n.s.	O<M<=C	O<C<=M
1 (a.u.)	O<=C<M	O<M	O<C<M	O<M
				White noise
45 dB	O<=C<M	n.s	O<C<M	O<M<=C
55 dB	O<=C<M	M<C	O<C=<M	O<C<=M
65 dB	O<M	n.s.	O<C<M	O<=C<M

In the "noticeable noise" group, for "da" syllables with 0.4 amplitude clicks, stimuli with noise at CV were chosen to contain the least noticeable noise, followed by those with noise at onset, and then those with those at the maximum amplitude of the sound waveform. These syllables produced the most entrainment in Experiment 3A compared to all other conditions. Furthermore, they contained the second highest pitch compared to all syllables used in Experiment 3A. Nonetheless, for "da" syllables containing clicks measuring 0.7 or 1 in amplitude, the noise at CV was not significantly more noticeable than at onset, but it was less noticeable than at the maximum amplitude. This finding was in line with our predictions.

However, "ta" syllables containing the same click amplitude showed onset-altered conditions to be the least disrupted or noisy. CV-altered syllables followed in ranking, having less noticeable noise than syllables altered at the maximum

amplitude. Moreover, CV-altered stimuli were not significantly more disrupted than either of the two other stimuli, in the “ta 0.4 click” grouping.

Most click stimuli containing noise levels higher than the ones used in Experiment 3A were significantly more disrupted, or contained significantly more noticeable noise, than unaltered controls. Exceptions in this case are “ta” syllables containing 55 dB white noise at onset, perceived as containing more noticeable noise than their respective controls, but not significantly ($p=.052$), and “da CV” syllables with 0.4 amplitude clicks. The latter were not different from control in terms of noticeable noise, in contrast with “da onset” and “da maximum amplitude” containing identical noise: these were always picked as noisier than controls. Furthermore, “da click CV” syllables were perceived as significantly less disrupted than unaltered control syllables, when the amplitude of the click was 0.4.

In most cases, syllables containing noise at the maximum amplitude were found to be either the most disrupted ones, or to contain the most noticeable noise, compared to the same syllables containing identical noise at one of the other two locations. Syllables with altered maximum amplitudes were found to be almost always more disrupted/noisier than onset-altered stimuli, which is in line with our predictions. The opposite was found in the EEG experiment, where entrainment was almost always higher at onset than at maximum amplitude conditions, and significantly so.

For “ta” stimuli, CV-altered stimuli always showed more disruption/more noticeable noise than the onset ones, but responses to these syllables were sometimes not different from those to stimuli containing noise at the maximum amplitude. This is of particular significance when considering results from the “disrupted syllable” group, at the level of 45 dB white noise. Here, both maximum

amplitude and CV-altered syllables sounded more disrupted than onset-altered stimuli. Importantly, CV-altered syllables showed the least entrainment for “ta white noise” conditions in Experiment 3A. On the other hand, CV-altered “ta” syllables were also found to have the same ranking as those containing noise at the maximum amplitude when the added noise was a 0.7 amplitude click, in both the “disrupted syllable” and “noticeable noise” group.

Overall, the results in the “disrupted syllable” groups are less clear compared to the “noticeable noise” group. This is especially the case for “da” syllables, which didn’t show any significant binomial probabilities for three of the six noise type/level combinations in the “disrupted syllable” group. Moreover, two other conditions in this group (“da click” of amplitude 1, “da 55 dB white noise”) only showed one significant probability each. On the other hand, rankings for “ta” stimuli in the “disrupted syllable” group were more consistent, and these showed more significant probabilities when white noise was added rather than a click.

Figures 4.B.2 and 4.B.3, constructed for the original levels of noise (click 0.4, white noise 45 dB), also reflect the preference for maximum amplitude over onset, and the slight differences between “da” and “ta” found at CV locations, for both the “noticeable noise” and “disrupted syllable” groups. The ascending height of the bars, from left to right, is also shown more in “ta” than in “da” conditions. “Ta” syllables generally showed a consistent trend in terms of the impact of noise added at different locations, irrespective of the experimental group (lowest at onset, followed by CV and maximum amplitude). This was especially the case when participants answered a question related to the noise loudness, and to a lesser extent, in the “disrupted syllable” group. Not as many significant probabilities were found for “ta” syllables in the latter group, especially when the added noise was a click. The

opposite was true for “da” syllables in the “disrupted group, which showed no significant probabilities when these were part of the “white noise” condition. Furthermore, unlike “ta” syllables, “da click CV” syllables were chosen as both the most disrupted syllables and the ones containing the most noticeable noise.

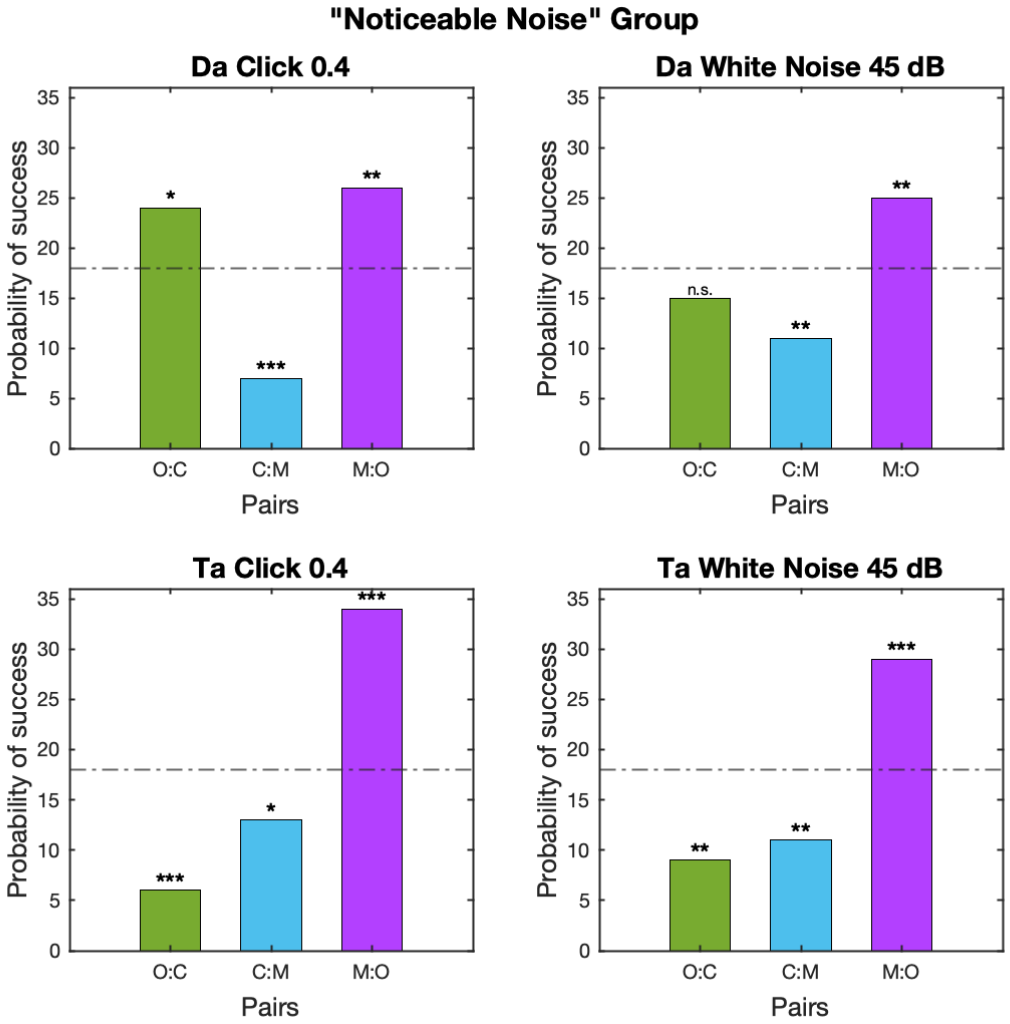


Figure 4.B.2**.** “Noticeable noise” group: number of successes for a stimulus in a given pair. O=onset, C=CV transition, M=maximum amplitude. Pairs are provided on the x axis of each graph. Number of successes for a stimulus (y axis) represents the binomial probability that a stimulus on the left side of the colon was picked over the one on the right (i.e., how often the “O” in the “O:C” is likely to be picked over “C”), as having more noticeable noise. Results are only provided for stimuli containing 0.4 clicks and 45 dB white noise. The dotted line represents chance level (50%). If a bar was under this level, then the item on the left of the colon was chosen more often, but if the bar was above the line, then the right item was chosen more often. Symbols above bars represent significance levels of the binomial probability associated with success: * = .05, ** = .01, ***=.001, n.s. = not significant.

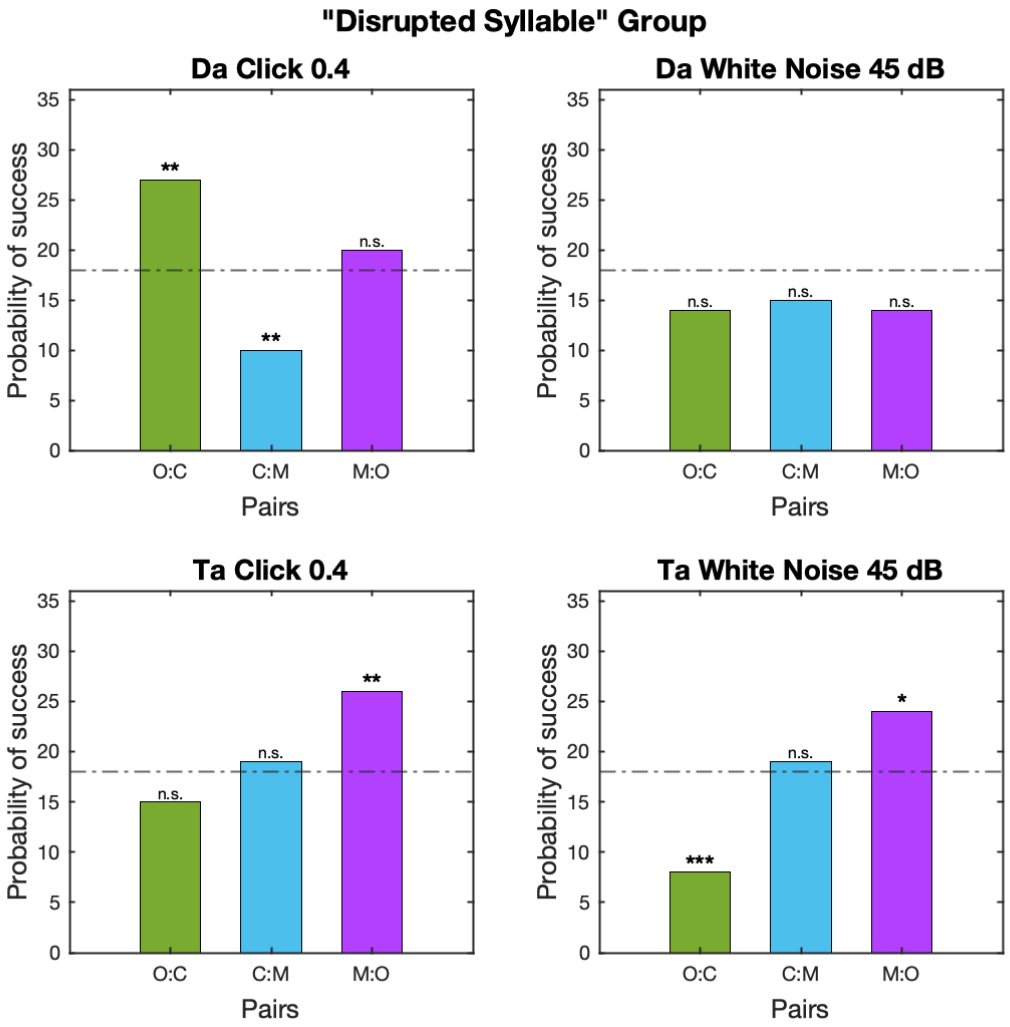


Figure 4.B.3**.** "Disrupted syllable" group: number of successes (y axis) for a stimulus in a given pair (x axis). Results are only provided for syllables containing 0.4 clicks and 45 dB white noise. Number of successes for a stimulus represents the binomial probability that a stimulus on the left side of the colon was picked over the one on the right as being more disrupted. Dotted lines represent chance level. Symbols above bars represent significance of binomial probability: * = .05, ** = .01, *** = .001, n.s. = not significant. See also Figure 4.B.2.

Results from the "disrupted syllable" group do not inform us whether the altered syllables sounded less like their corresponding controls, to the extent that the syllable-initial phoneme was perceived as different one (e.g. /k/ instead of /d/), or whether the perceived isochrony of the altered syllables was different from that of control stimuli. This distinction was not directly investigated in the current study. Overall results are different between the "disrupted syllable" and the "noticeable

noise" group, indicating that different perceptual factors were measured in the two groups. Conversely, comparisons between click and white noise-altered stimuli, at the levels of noise used in Experiment 3A (see Table 4.B.3), indicate that click syllables were perceived both as containing more noticeable noise, and were also more disrupted, than white noise stimuli.

Table 4.B.3. Rankings of click versus white noise comparisons, at each location, for original levels of noise (0.4 click, 45 dB white noise). WN = white noise. MA = maximum amplitude. Stars indicate significance levels of the binomial probability. No stars indicate probability was not significant.

	Da		*Ta*	
	"Noticeable Noise"	"Disrupted Syllable"	"Noticeable Noise"	"Disrupted syllable"
Onset	Click<WN**	Click<WN	WN<Click*	WN<Click**
CV	WN<Click	Click<WN*	WN<Click ***	WN<Click **
MA	WN<Click***	WN<Click *	WN<Click ***	WN<Click ***

*$p<.05$
**$p<.01$
***$p<.001$

Discussion

In the present experiment, syllables containing noise at the maximum amplitude of the sound waveform were identified either as most disrupted (sounded the least like their unaltered controls), or containing the most noticeable noise, whereas the opposite was found for the onset-altered stimuli. This applied for both types of noise (click and white noise), at most levels of intensity. It is possible that the perception of the noise increased with the sound amplitude of the carrier syllable. The amplitude of the acoustic waveform of the syllable was also the smallest at onset, followed by CV, until it reached its maximum value. Similar results were found in the "disrupted syllable" group, where syllables altered at the onset were also perceived as the least disrupted, followed by the other two conditions, in the same order as for the other experimental group. It is thus possible that a stimulus was perceived as disrupted for

the simple reason of containing noise, with louder noise leading to more disruption, or less similarity between the altered syllable and its control. This is confirmed by comparisons between stimuli containing clicks and white noise at the same location, for the levels of noise used in Experiment 3A, where click stimuli were almost always found as either more disrupted, or as containing more noticeable noise than syllables in the other condition.

Nonetheless, differences between the two experimental groups pose a problem for this interpretation. For “da” syllables in the “disrupted” group, the differences between stimuli were not found to be significant as often as they were in the other group. Furthermore, binomial probabilities for “ta” stimuli in the “disrupted” group were found to be more significant if the syllables contained white noise, but not clicks. The “noticeable noise” group did not show similar trends for either of the two syllables. These findings suggest that while similar perceptual factors were measured in the two experimental groups, these were not identical. Secondly, the two types of noise affected the perception of the two stimuli differently, presumably due to differences in the syllable-initial consonant. There is a slight indication that the white noise, but not the click, led to “ta” syllables sounding more disrupted, whereas both types of noise were noticeable in “da” syllables without necessarily disrupting their perception.

In the present experiment, participants identified the level of disruption of a stimulus with reference to a single presentation of its unaltered control. Thus, a “ta” syllable containing white noise at CV was selected as sounding less like “ta”, but the exact implications of that comparison are not clear. For example, the /d/ in an altered “da” syllable may have sounded less like the /t/ in a control syllable, or altered “da” syllables may have seemed less isochronous than unaltered ones. The fact that

control “da” syllables sounded more disrupted than “da” syllables with 0.4 amplitude clicks added at CV is a particularly mystifying finding. Results from the “disrupted syllable” group remain somewhat difficult to interpret. Individual differences in terms of how participants interpreted the word “disrupted” may have had a considerable impact on the findings. Thus, we cannot confirm that a syllable sounded more disrupted simply because it contained more noticeable noise.

There were also exceptions to the onset-CV-maximum amplitude order, in terms of both noticeable noise and syllable disruption levels. Importantly, these exceptions were generally found for stimuli which elicited either the most or the least entrainment in Experiment 3A. When the click was 0.4 in amplitude, “da click CV” syllables were chosen either as the ones being the least disrupted, or the ones containing the least noticeable noise, compared to similarly-altered onset and maximum amplitude syllables. Furthermore, such “da click CV” syllables were found to be less disrupted and to contain less noticeable noise than unaltered “da” syllables. On the other hand, “ta CV” stimuli containing 45 dB white noise were chosen as the most disrupted ones compared to stimuli with identical noise at the other two locations. However, they were not also found to contain the most noticeable noise. One possibility for these results is that spectral information is the least uniform at CV: a single click may have not been detectable enough for “da” syllables, which also show a more abrupt rise time than “ta” syllables, whereas white noise at “ta” syllables may have made this information appear more uniform. However, it is unclear why these findings were not replicated at higher levels of noise, or why the differences between the two syllables were so drastic.

Another interpretation for these findings is that the 0.4 click caused “da CV” syllables to sound more isochronous, whereas the 45 dB white noise had the

opposite effect on “ta” syllables. It is possible that adding noise at different syllabic locations altered the P-centre of the syllables, but this remains to be determined. Previous findings indicate that, while the P-centres for both voiced and unvoiced consonants seem to be located in the vicinity of the CV transition (Barbosa et al., 2005; Šturm & Volín, 2016), there is more variability in the identification of P-centres for syllables starting with longer consonants than for shorter ones (Villing et al., 2011). This could imply that the location of the P-centre is more fixed for “da” than for “ta” syllables. Thus, a click at “da” may have helped the identification of an already clear P-centre, whereas white noise at “ta” may have caused an unclear P-centre to become even less clear. This claim is somewhat contradicted by the fact that “ta” syllables containing a 0.7 click at CV were found to contain as much noise and be as disrupted as the ones altered at the maximum amplitude. It is however possible that a less clear P-centre at the CV of “ta” syllables may benefit from a louder click than “da” syllables, although not too loud (a click of amplitude 1 did not elicit the same results). Further investigations need to be conducted in order to test these various claims.

Overall, the present results confirm that the findings in entrainment from Experiment 3A were not just due to the perception of the noise itself, but because of the interaction between the acoustic properties of the noise and the underlying syllable, interactions which were most likely the strongest at the CV transition.

4.3. General Discussion Experiments 3A and 3B

Previous research has suggested that the brain entrains to the syllabic rhythm of speech by tracking specific landmarks present within the syllable. Neural phase locking was stronger for theta oscillations, which are associated with speech envelope tracking, for acute, short-lived tones more so than for sinusoidal ones (Prendergast et al., 2010), indicating a preference for discrete landmarks over continuous information, which is more difficult to parse. The location of such landmarks has nonetheless remained debatable. On the one hand, some studies indicated onsets as preferred locations for entrainment. For example, theta entrainment to speech-noise constructs which did not present any envelope fluctuations was stronger if the phonemic information at the onsets, but not the coda of syllables was present in the stimuli (Zoefel & VanRullen, 2015). Oganian and Chang (2018) found that phase locking to the speech envelope was strongest at the peak derivative of the speech envelope and suggested this may be landmark for neural tracking, whilst also emphasising the crucial importance of the rich acoustic information present in the onset of syllables. Moreover, Oganian and Chang (2018) associated the peak derivative of the envelope with syllabic formant transition. Behavioural entrainment studies found that when presented with sequences of isochronous syllables, participants identified the onsets of the syllables (known as P-centres) as closer to the onset of the vowel, or the CV transition of the syllables (e.g., Barbosa et al., 2005; Šturm & Volín, 2016). However, some researchers believe that the brain follows vowel peaks when entraining to speech, because these contain the highest sound intensity (Ghitza, 2013). Indeed, this claim is supported by comprehension studies in which concealing vowel peaks with noise severely damages speech comprehension, whereas noise replacing consonants at the onset

or coda of syllables does not seem to result in the same impact (Fogerty et al., 2010; Fogerty & Kewley-Port, 2009).

In Experiment 3A, we placed minimal noise at the onset, CV and vowel peaks of syllables “da” and “ta”, in order to test whether this would lead to more (or less) neural phase locking to syllables altered at one location, versus those with noise at other locations. We found that the effects of noise were strongest for syllables with modified CV transitions: these were the only conditions to which phase coherence was significantly different from unaltered syllables. Moreover, the direction of entrainment seemed to depend on both the syllable-initial consonant and the type of noise: “da click CV syllables” led to the most phase locking, whereas “ta white noise CV” stimuli showed the least entrainment. Because the click was identical in all click stimuli, and the white noise was identical in all syllables containing white noise, we claimed that differences in entrainment results for CV-altered stimuli were due to the importance of the CV transition over the other two locations for neural speech tracking. Nonetheless, because the sound intensities of the syllable were different across the locations at which the noise was placed, the noise may have also been perceived as more or less intense depending on where it was located. We considered the possibility that the findings in entrainment were due to the noise either being more noticeable at some locations than others, which may have triggered to more entrainment to the noise alone, or that the noise might have caused certain syllables to sound more disrupted than others, and resulting in a reduction of phase locking to the underlying syllabic rhythm.

In order to verify the perceptual impact of noise on entrainment, we conducted a behavioural experiment in which participants reported the prominence of the noise of syllables used in Experiment 3A, or the level of disruption of the stimuli.

Furthermore, in Experiment 3B we also constructed stimuli with clicks and white noise louder than in the previous experiment. In the event where the intensity of the background syllable affected the intensity of noise perception, an effect of location would be consistent across multiple levels of noise, unless, of course, the noise was more noticeable than the syllable itself. In general, stimuli containing either type of noise were chosen as the least noisy/disrupted if the noise was placed at the onset of the syllables, whereas the opposite was found for stimuli containing noise at the maximum amplitude, with CV-altered syllables falling in between the other two. Nonetheless, “da” syllables with 0.4 amplitude clicks at CV were found to contain the least noticeable noise or chosen to be the least disrupted, indicating that the perception of noise did not alter entrainment to such syllables in Experiment 3A. However, “ta” syllables where 45 dB noise was placed at CV locations were found to be as disrupted as those with the same noise placed at the maximum amplitude. This suggests that a reduction in entrainment from control to “ta white noise CV” syllables in Experiment 3A may have been due to altered syllables sounding less like unaltered controls. However, results from the “disrupted syllable” group did not clarify why altered syllables sounded less like unaltered ones, i.e., whether this was due to acoustic reasons such as the masking of phonemic information at CV, or because they were perceived as less isochronous. Differences in the perceived isochrony of the stimuli used in Experiment 3A is a potential explanation for the differences seen in entrainment to these stimuli. However, the fact that changes in either isochrony or neural processing of speech stimuli are most prominent when alterations occur at the CV transition of syllables indicates that this a crucial landmark for speech tracking.

The fact that the two types of noise affected the two syllables in opposite ways could be attributed to differences in voicing between the two syllable-initial

consonants. Whereas P-centre research shows that participants are equally consistent in identifying the perceived onsets of syllables at vowel onset locations for syllables starting with both voiced and unvoiced stop consonants (Barbosa et al., 2005; Šturm & Volín, 2016), some studies showed that there is more variability in P-centre identification when syllables start with longer than with shorter consonants (Villing et al., 2011). Furthermore, syllables starting with longer or multiple consonants seem to have P-centres located closer to vowel peaks. Therefore, we cannot suggest that the CV transition is an equally important landmark for all syllables. Further investigations could tackle the importance of landmarks in entrainment to different types of syllables, by measuring neural tracking to a variety of syllable-initial consonants, altered at more locations than in the present experiment, and using multiple level of noise, to account for all possible interactions between the acoustic properties of the syllable and the added noise.

5. General Discussion

The aim of the present research was to delve deeper into the mechanisms responsible for the neural tracking of speech, particularly concerning the entrainment of cortical theta oscillations to the syllabic rhythm. Neural phase locking to the slow (~2-5 Hz) envelope fluctuations of speech is a well-established phenomenon: without their presence in the speech signal, comprehension is gravely affected (Drullman et al., 1994 a,b), and the brain seems to entrain to amplitude modulations in this frequency range, even in the absence of intelligibility or comprehension (Ding & Simon, 2013; Howard & Poeppel, 2010); conversely, phase alignment to speech outside the theta range seems to rely on comprehension (i.e., in the delta range: Bourguignon et al., 2013; Molinaro & Lizarazu, 2018) or may depend on coupling to theta oscillations (i.e., gamma: Gross et al., 2013; Hyafil et al., 2015). Low-frequency amplitude fluctuations in the speech envelope mostly correspond to the syllabic rhythm and researchers claim that the 'beat' provided by syllables helps the brain parse the incoming speech information into chunks, which aids both acoustic intelligibility and semantic comprehension (Doelling et al., 2014; Ghitza, 2011).

However, little is known about the mechanisms by which brain oscillations temporally align to match the syllabic rhythm. It has been claimed that the purpose of oscillatory processes is to track and predict patterns in the surrounding world (Arnal & Giraud, 2012). Such patterns can be seen in the quasi-regularity of speech envelope fluctuations or the durations of syllables, which are always contained within certain duration intervals (the average being four syllables a second (Hyafil et al., 2015). Furthermore, syllables are marked by onsets and offsets in the envelope which could reset the phase of ongoing theta oscillations (Ghitza, 2011, 2013). As

such, speech tracking may rely on the detection of edges in order to successfully parse the signal into syllables.

The issue with this assumption is that it has not been fully established whether phase resetting of endogenous oscillations occurs in response to a rhythmic stimulus. There are clear advantages which this mechanism poses for neural processing: it utilizes less energy, especially when there is a close match between the eigenfrequency of the oscillation and the stimulus (Obleser & Kayser, 2019), it constitutes evidence that the brain is temporally predicting regular events (Zoefel et al., 2018). However, even in the case of perfectly periodic stimuli, researchers have found it difficult to show evidence in favour of this phenomenon, with some researchers arguing that 'entrainment' can best be explained as a superposition of evoked responses (e.g., Capilla et al., 2011). There is now increasing evidence for entrainment to perfectly regular stimuli (e.g. Notbohm et al., 2016), including for fixed-rate speech (Zoefel et al., 2018), but it remains to be determined whether this still happens in response to continuous, quasi-regular speech.

However, it has been suggested that one should consider 'entrainment in the broad sense' instead, which refers to the temporal alignment between the stimulus and the brain activity, without confirming that this necessarily refers to endogenous oscillations (Obleser & Kayser, 2019). In fact, phase patterns of evoked activity are always consistent in the presence of a rhythmic stimulus (Zoefel et al., 2018). Furthermore, it seems that even the amount of phase coherence of the evoked activity depends on the specific landmarks of the stimulus.

This was first suggested by Prendergast et al. (2010), who showed that tone sequences which were frequency-modulated at 4 Hz elicited stronger ASSRs when the tones were more pulsatile, or shorter-lived, than more continuous, sinusoidal

tones. Thus, a certain rhythm would be better tracked if it were marked by certain acoustic edges. Furthermore, an animal study investigating the LFP activity in the auditory cortex of rats during music and 1/f noise presentation, showed there was phase-resetting of LFPs at events happening with 2-4 Hz regularity (Szymanski et al., 2011). These events were spikes in activity which were found to occur spontaneously in the absence of exogenous sounds, or during noise stimulation, but happened regularly during music presentation, revealing entrainment to the beat of the music. Consequently, researchers believe that phase locking to such events helps process crucial information about the acoustic stimulus.

Nonetheless, it remains unclear what edges represent in the context of neural speech processing. Because envelope fluctuations are crucial to speech tracking, some researchers propose that some envelope landmarks may count as markers for entrainment. Doelling et al. (2014) introduced a method called sharpness, or the summed positive derivative of the narrowband speech envelope, to account for the total amount of rise time present in syllables, and found that stimuli with higher sharpness led to higher entrainment of theta oscillations. Subsequently, Oganian and Chang (2018) found that phase locking to the broadband envelope of continuous speech was highest at the latency of the peak derivative of the envelope, or when the speech rate was the highest. Furthermore, this point is also believed to represent the acoustically rich information which characterizes syllabic formant transitions, or vowel onsets.

Other researchers have claimed that mid-vowel locations may be the most important landmarks for entrainment, as these carry the highest acoustic energy and represent envelope peaks (Ghitza, 2011). Furthermore, vowels also seem to carry syllabic stress: stress impacts both the duration and intensity of syllabic peaks and,

to a certain degree, those of syllabic onsets, but never the codas (Greenberg et al., 2003). That vowels may be more important than other speech sounds comes from some comprehension studies: this seems to be more affected when vowels are replaced by noise, but not the surrounding consonants (Fogerty et al., 2010; Fogerty & Kewley-Port, 2009). However, in the case of stimuli presenting mainly consonant information, comprehension seems to be restored when some of the vowel information in the vicinity of the consonants is restored. These results indicate that, while vowel information seems to be crucial for speech processing, this may not be necessarily contained only by its peaks, but also at formant transitions.

Formant transitions also seem to be associated with syllabic onsets, as specifically claimed by Oganian and Chang (2018). Some researchers also emphasise the importance of fine-grained acoustic information placed at syllabic onsets in neural speech tracking, including entrainment of the theta oscillations. One study by Zoefel and Van Rullen (2015) showed that, when presented with speech-noise stimuli where all low-frequency envelope fluctuations were concealed, participants detected tone pips which were overlaid on top of the stimuli, if these were placed at syllabic onsets, but not at other syllabic locations. Moreover, a follow-up study by Zoefel and Van Rullen (2016) showed that the level of entrainment of theta oscillations to such speech-noise constructs was the same as for the intact speech stimuli (only the phase of entrainment was different), leading them to believe that the rhythmicity of phonemic information which occurs at syllabic intervals may be enough to trigger syllable tracking.

The latter findings are of particular importance in the light of current theories about the role of envelope fluctuations in speech processing. In fact, edges are always defined in the context of the envelope. For example, one study found that, in

the vicinity of edges, there was an increase in phase locking of theta oscillations, an increase in power of gamma oscillations, as well as a greater level of coupling between the two neural rhythms (Gross et al., 2013). However, the edges in this study were merely identified as sufficient increases in amplitude between successive onsets in the envelope, with the level of increase, as well was what defines an envelope onset, being arbitrary. The existing research does not offer enough clues as to what edges may be, and furthermore, whether differences in the acoustic quality of different edges may impact entrainment to natural speech.

In the present research, we investigated the effects of edges on entrainment to the syllabic rhythm, by altering the fine-grained information present at syllabic onsets, through the use of different phonemes. We aimed to test whether differences in syllable-initial consonants would lead to differences in phase locking of theta oscillations, as well as differences in speech edge markers, such as sharpness, or other envelope-related properties. Furthermore, we evaluated the relationship between theta entrainment and edge markers and lastly, between the phase locking of theta oscillations and acoustic modifications to specific syllabic locations.

In Experiment 1, we used different types of phonemes at the beginning of syllables to construct sentences with “strong” and “weak” edges: stop consonants were used as “strong” landmarks, whereas “weak” landmarks were represented by a selection of liquids, fricatives and sibilants. We constructed such sentences for two different languages, English, the native language of participants, and Russian, as the foreign language. The aim of this experiment was to investigate whether “strong” sentences would lead to more entrainment in the theta range than “weak” ones, and whether this effect was modulated by comprehension, as manipulated by the language condition.

We found that “strong” stimuli were significantly different from “weak” ones in terms of the normalised Doelling sharpness, but not in the original Doelling sharpness. This meant that envelope rises in the “strong” group were steeper than in the other group only when these were measured relative to the cumulative value of the amplitude envelope. We attributed these findings to the fact that, unlike Doelling et al. (2014), who constructed stimuli with significantly different sharpness values, we used natural stimuli, between which differences in the “edge” content were minimal, and would only show in a scaled version of their sharpness measure. We did not find any differences in sharpness, Doelling or normalised, between sentences corresponding to the two language groups.

Our results showed that the inter-trial phase coherence difference was higher at frequencies between 1 and 10 Hz than at frequencies above this range. Moreover, the inter-trial phase coherence difference was higher in the theta (4-8 Hz) than in the delta (1-4 Hz) or alpha (8-12 Hz) ranges. The phase coherence difference at frequencies between 1 and 12 Hz was higher during stimulation than at baseline, which was not found for the inter-trial power coherence difference. This is in line with findings from Luo and Poeppel (2007), who also did not find any peaks in the power coherence difference measured to continuous speech stimuli, indicating that this may be a result of entrainment to endogenous neural oscillations which were not accompanied by any additional evoked activity.

However, we found no effects of sharpness or language in the theta range of the inter-trial phase coherence. Nonetheless, Russian stimuli elicited more phase coherence than English ones for frequencies between 1-4 Hz, and “strong” stimuli also had greater phase coherence values than “weak” ones between frequencies spanning 8-12 Hz. The lack of differences between conditions in the theta range

remains unclear. Based on previous findings, we expected that at least an effect of comprehension on entrainment would be apparent at this frequency level.

For example, Perez et al. (2019) showed that neural phase locking to the syllabic rhythm of speech was greater for native than for foreign language stimuli. On the other hand, other research investigating the effects of intelligibility on entrainment showed mixed results. Some studies showed that there are no differences in theta phase locking between forward and reverse speech (Howard & Poeppel, 2010; Zoefel & VanRullen, 2016), while others showed that reverse speech triggers significantly less envelope tracking (Di Liberto et al., 2015). Furthermore, the lack of an edge effect may partly be explained by the fact that some syllables also contained consonants in the coda, although the role of coda information in entrainment to the syllabic envelope remains to be determined.

The results in the delta range of the inter-trial phase coherence were surprising: in general, entrainment at this level only shows in the presence of comprehension (Molinaro & Lizarazu, 2018), or if the stimuli contain strong prosodic fluctuations (Bourguignon et al., 2013). There is a possibility that Russian stimuli may have appeared more stressed than English ones, but this aspect was not investigated. The differences between sharpness conditions at the alpha level were also unexpected, especially given the lack of findings in the theta range. The alpha range is not considered a meaningful window for speech entrainment, because the frequency rates of syllables, phonemes or phrases do not fall in this range (Luo & Poeppel, 2012). However, some comprehension studies showed that eliminating envelope fluctuations between 8 and 16 Hz impaired recognition of certain phonemes, especially plosives (Drullman et al., 1994 a,b). It is possible that our results may reflect that stop consonants were indeed perceived as sharper than the

other consonants by the brain, which reflected in more entrainment to such consonants, but not that syllables containing stop consonants were also sharper than other types of syllables.

There were no effects present in terms the cerebro-acoustic coherence difference or the inter-trial power coherence difference: there were no frequency ranges which distinguished themselves amongst others, and these two measures were also not significantly different from baseline (the inter-trial phase coherence difference was also greater than baseline for frequencies between 1-10 Hz). However, we did not use the exact formula given in Doelling et al. (2014) to calculate this. We also discussed the possibility that, while our stimuli triggered similar neural responses to the same stimuli, which was illustrated by an increased amount of inter-trial phase coherence, the small number of repetitions for any given stimulus, the short duration of the stimuli, as well as their complexity in terms of both syntax and semantics may be responsible for the lack of cerebro-acoustic coherence.

The overall null results may also be explained by other methodological limitations pertaining to the EEG analyses. In this study, we averaged across all channels, including occipital ones, which may partially account for the peaks in the alpha range noticeable in the phase coherence differences. While similar research reports averaging across all electrodes (e.g., Ding et al., 2017), we could have extracted a subset of channels, either for the auditory region or using linear regression based on the highest auditory ERPs for each subject. Investigating ERPs may have also been helpful in computing a linear model of the waveform in response to speech, which could sometimes be a more accurate method of investigating phase locking to the stimulus. However, ERPs used to compute spectro-temporal

functions would have been difficult to extract due to the lack of necessary epochs. Future studies could account for these limitations.

Overall, Experiment 1 replicates findings from Luo and Poeppel (2007) in showing increased neural entrainment in the theta range, as given by the inter-trial phase coherence difference, which was not found in the power coherence difference, suggesting that only a mechanism involving the phase alignment of endogenous oscillations was required for the tracking of continuous speech. Furthermore, we found that stop consonants placed at the beginning of syllables may indeed lead to stimuli with stronger edges than the ones containing other consonants at syllabic onsets, as given by the normalised sharpness measure. However, the lack of differences between conditions in theta entrainment could not confirm an effect of edge on syllabic tracking. We discussed that, even if existent, such differences may have been difficult to notice due to the technical limitations involving both the construction of our stimuli as well as the analyses. Consequently, Experiment 2 controlled for such aspects.

In Experiment 2, we used stimuli comprising of nearly-isochronous CV syllables. The stimuli were also longer in duration, and were repeated a greater number of times than those used in Experiment 1. Longer, isochronous stimuli with a greater number of repetitions were likely to increase our chances of observing entrainment. Phase locking was found to “build up” over time, with preferred phases of entrainment of neural oscillations changing throughout stimulation (Riecke et al., 2015), which implied that longer stimuli could also lead to stronger phase locking. Even if by using periodic stimuli and generating evoked activity meant that we may have only measured ‘entrainment in the broad sense’ (Obleser & Kayser, 2019), we

were only interested in obtaining a greater neural response and not the exact nature of occurrence of oscillatory phase alignment,

We also used a different syllable-initial consonant for each of our conditions in Experiment 2, while the same set of vowels were used across conditions. By manipulating only the initial consonant, we were able to test differences in syllabic entrainment based on the acoustic properties of syllabic onsets, as well as investigate how these affected various envelope properties of the syllables in each condition. In this respect, we introduced other edge markers alongside the Doelling and normalised sharpness, which were initially used in Experiment 1. We calculated the value of the maximum amplitude of the narrowband envelope and its latency, the peak derivative of the maximum amplitude and its latency, as well as the Gini index, which measured the total level of inequality in the narrowband envelope of each syllable.

Stevens (2002) claimed that phonemes, such as vowels and consonants, provide different landmarks or features which are necessary for speech recognition and processing. These features depend on the acoustic and articulatory properties of the phonemes (Stevens, 2002), have been shown to be encoded separately by the brain (Mesgarani et al.,2014), and different acoustic features of consonants have also shown to impact neighbouring vowels in distinct ways (Stevens, 2002). Results from Experiment 2 showed that entrainment to the syllabic rhythm, measured either as the 4 Hz Power, or as a linear combination of the inter-trial phase coherence at the syllabic rate and its harmonics named the Compound ITC1, showed differences between conditions based largely on phonemic features, such as manner of articulation and voicing. Consonants such as stops, where the air passing through the vocal tract was blocked (plosives and nasals) led to the most entrainment, while

sibilants and fricatives, where the oral cavity is only partially blocked during articulation, led to the least entrainment. Moreover, voicing had opposite effects on the Compound ITC1 depending on whether the syllable-initial consonant was a plosive or a fricative, with voiced plosives eliciting more phase locking than unvoiced ones, while voiced fricatives showed less entrainment than unvoiced ones.

In order to test the differences in acoustic features between our conditions, as well as their effects on entrainment, we performed a PCA on all envelope edge markers. PCA allowed us to denoise the data and combine effects from ITC harmonics into fewer analyses. Similar clusters, based on the manner of articulation and voicing, were observed when plotting the first two components of the PCA, across the different conditions. Furthermore, the first component of the PCA was negatively correlated with Compound ITC1. The direction of the correlation was explained by the individual correlations of the edge markers and the Compound ITC1: the latter was negatively correlated with the latencies of the peak derivative of the broadband envelope, and that of the peak narrowband envelope, but negative correlations were also found between entrainment and sharpness (Doelling and normalised). A negative correlation between the latencies of the peak derivative and that of the peak envelope were expected: Doelling et al. (2014) also suggested that sudden, steep changes in the acoustic envelope provide more reliable edges for entrainment. However, the negative correlations between the Compound ITC1 and sharpness were unexpected, as in this study, more sharpness was associated with less phase locking.

Following from these results and those of Experiment 1, we argued that the Doelling and normalised sharpness may not be the best edge markers of natural stimuli, which show more envelope fluctuations than the stimuli used by Doelling et

al. (2014). Nonetheless, it is possible that sharpness may be a better suited option when applied to the broadband envelope rather than narrowband envelope of the syllable: as seen in Figure 3.2, the broadband envelope is a lot smoother than the narrowband envelope, and consequently, its positive derivative would be more likely to refer to amplitude rises located at syllabic onsets as opposed to the sharpness of the narrowband envelope, which also takes into account micro-variations in amplitude across the entire envelope. Future studies need to establish the validity of this claim.

In Experiment 2, we showed, for the first time, the effects of acoustic and articulatory properties of syllable-initial consonants on entrainment to the syllabic rhythm. This is particularly important considering how phoneme tracking has been interpreted by the current literature. In general, the brain is thought to rely on theta oscillations to parse the incoming speech signal into syllabic units, and by coupling with the theta rhythm, gamma oscillations are able to further divide syllables into their component phonemes (Ghitza, 2013; Giraud & Poeppel, 2012). Furthermore, if the coupling between theta and gamma is weakened, entrainment in the gamma range is also diminished, leading to the poorer identification of phonemes (Hyafil et al., 2015).

Equally, only the slow envelope fluctuations, or the syllabic rhythm, seem to be most important for comprehension: in their absence, but not in the absence of fine-grained acoustic information, comprehension is severely diminished (Doelling et al.,2014). Nonetheless, some studies have found that, in noisy environments, when envelope fluctuations are highly degraded, theta entrainment to speech remains unaffected (Ding & Simon, 2013; Zoefel & VanRullen, 2016). These researchers claim that this may be due to the rhythmicity present in at the level of the fine-

grained, or phonemic structure of speech, which provides landmarks for syllabic entrainment.

In Experiment 2, we showed that syllable-initial phonemes affect both syllabic entrainment, as well as the envelope properties of speech. Despite the high number of experimental conditions and edge measures employed in this experiment, the correlations between Compound ITC1 and the latencies of the peak envelope and that of the peak derivative indicated that both CV transitions and vowel peaks may have a role in speech entrainment, as previously suggested by a number of researchers (CV: Oganian & Chang, 2018; vowel peaks: Ghitza, 2011, 2013).

In Experiment 3, we aimed to alter the acoustic content of specific syllabic locations, without severely affecting envelope properties, i.e., some of the edge markers which we found to be significantly correlated with entrainment in Experiment 2. In this experiment, we used only two 250 ms syllables, “da” and “ta”, which were used in separate conditions. Phonemes /d/ and /t/ are usually regarded as a pair of stop consonants, because they are produced in similar ways (they are ‘alveolar’, because the sound is made when the tongue meets the alveolar ridge behind the teeth), with the main difference being that /d/ is voiced and /t/ is unvoiced. Furthermore, by using the vowel in both syllables, we were better able to investigate the differences between a voiced and unvoiced stop consonant in terms of their CV transitions, vowel peaks and edge markers.

The syllabic locations we investigated were absolute onset, the latency of the formant transition and that of the vowel peak, the latter of which also corresponded to the peak of the sound waveform, in each of the two syllables. We manipulated these locations by introducing two different types of noise at each of their latencies, a single-sample click and a 5 ms snippet of 45 dB white noise, which was less than the

average intensity of the syllables (70 dB). “Da” and “ta” were manipulated in the same way, by introducing either a click or white noise at any of the three syllabic locations. As expected, the CV and maximum amplitude of “da” was later than that of “ta”. The unaltered “da” and “ta” syllables were also used for separate control conditions.

In Experiment 3A, we recorded participants EEG activity while they listened to repetitions of the isochronous syllables in each condition. We measured entrainment in the same way as in Experiment 2, by taking either the 4 Hz evoked power, or the Compound ITC1 resulting from the PCA conducted on the inter-trial phase coherence taken at the frequency of stimulation and its harmonics. Results in Compound ITC1 showed that only two conditions elicited entrainment which was different from the control conditions: “da click CV” showed significantly more phase locking than unaltered “da” syllables, while “ta white noise CV” triggered significantly less phase coherence than unaltered “ta”. When altered conditions were compared to each other, onset-manipulated stimuli tended to trigger more entrainment than the ones with an altered maximum amplitude, but such conditions were never significantly different from control.

We claimed that these results were due to the CV being a more important landmark for entrainment than the other investigated two locations. Nonetheless, because the two different types of noise seemed to have different effects on the syllables depending on the syllable-initial consonant, we argued that the CV transition may have been a more stable landmark for “da” than it was for “ta”. This claim is supported to a degree by the P-centre literature. P-centres refer to the perceived onsets of syllables determined behavioural entrainment experiments, i.e., when participants listen to streams of isochronous syllables (Morton et al., 1976).

Such studies have found that P-centres tend to be located more closely to the onset of the vowel of CV syllables, or the CV transition (Cooper et al., 1986; Harsin, 1997; Marcus, 1981). Moreover, their location seems to be more variable when the syllable-initial consonants are long, rather than when they are short (Villing et al., 2011). Because /t/ is longer than /d/, it is possible that the P-centre of syllables starting with /t/ may also be more variable than for those beginning with /d/, although that specific result has not been reported in the existing literature. However, if that is the case, it may mean that the landmark provided by the CV transition of “ta” may be less reliable than for “da”.

Importantly, it seems that effects on entrainment were not due to envelope alterations. Even if the click was noticeable in the envelope of the syllables that contained it, this was not the case for white noise conditions. Furthermore, white noise syllables did not affect edge markers such as Doelling sharpness, normalised sharpness, the Gini index, or the values of the peak envelope, the peak derivative and their latencies. Amongst these, the most important finding concerned the peak derivative of broadband envelope: this was previously suggested as a potential landmark for entrainment and thought to correspond not only to the peak rate of change at the syllabic level, but also to formant transitions (Oganian and Chang, 2018). We also found that the latencies of the CV and that of the peak derivative were remarkably close. The fact that the value and latency of the latter remained unchanged across 45 dB white noise conditions suggests that the fine-grained spectral information at CV is more likely to represent a landmark for entrainment than its envelope properties. If so, these findings would support the theory that acoustic edges for neural tracking are found the rhythmicity found at the high frequency spectral level of speech (Ding & Simon, 2013; Zoefel & VanRullen, 2015).

Nonetheless, we argued that it was possible that the click at the CV of “da” syllables was more noticeable than in other locations, or that the white noise masked crucial acoustic information at the CV of “ta” syllables, perhaps causing them to sound less like “ta”. These assumptions were tested in Experiment 3B, in which we investigated the perceived effects of each type of noise placed at the different syllabic locations of “da” and “ta”. We used the same conditions as in Experiment 3A, but also added two more intensity levels for the noise and the click. If the audibility of the noise/the disruption of the syllable depended on the sound intensity of the syllable at any given location, we expected results to be consistent across multiple noise intensities. Indeed, this seemed to be the case for most conditions, where onset-altered stimuli sounded the least disrupted or were chosen to have the least noticeable noise, followed by stimuli with alterations at the CV transition or maximum amplitude. This result immediately implies that entrainment results were not affected by perceptual factors: CV-altered conditions seemed to be in the middle between the other two conditions in terms of perceptual effects, but for two of the CV conditions, entrainment results were found at extremes in Experiment 3A.

However, we found that “da click CV”, which elicited the most entrainment in Experiment 3A, was found to contain the least noticeable noise with respect to syllables containing the same noise type and level, whereas “ta white noise CV” syllables, which triggered the least neural entrainment, were chosen as the most disrupted ones amongst other syllables containing white noise of 45 dB intensity. Furthermore, original “da click CV” syllables were found to have less noticeable noise and were also less disrupted than control “da” syllables. There are two different ways in which these results can be interpreted. On one hand, the 0.4 amplitude click at the CV of “da” syllables seems to have not been audible at all,

implying that neural entrainment to such syllables was not affected by perceptual factors; at the same time, the 45 dB white noise placed at the CV of “ta” syllables caused these to sound less like the unaltered “ta” syllables, suggesting that this type of noise masked crucial phonemic information which did affect neural phase locking to these syllables.

In sum, there was a perceptual effect on neural entrainment for “da CV” syllables with 0.4 clicks, but not for “ta CV” stimuli containing 45 dB white noise. On the other hand, it is possible that the first syllables sounded more isochronous, whereas the latter sounded less isochronous than all other conditions, including controls. In any situation, it appears that the noise only interacted with the acoustic properties of the syllables at the CV location, making this a more important landmark for entrainment compared to onset and maximum amplitude. Nonetheless, the exact acoustic and perceptual reasons for this remain unclear and require further investigation.

It is important to note that the results of Experiments 2 and 3 may not apply to continuous speech, in the sense that the syllable-initial consonant may not trigger differences in theta neural entrainment across syllables. This is partly confirmed by a lack of significant effects of sharpness in Experiment 1, although limitations of this study were evaluated. However, by making syllables isochronous or nearly isochronous, we completely neglected stress patterns and their effects on entrainment. Some researchers believe that stress can also mark edges for the neural entrainment of syllables, as suggested both by behavioural (Quené & Port, 2005) and neural entrainment research (Leong et al., 2014, 2017). For example, one behavioural study asked participants to track a certain plosive placed in words or syllables which followed an unambiguous stress pattern, i.e. the stress intervals were

either exact or fell within a certain durational range (Quené & Port, 2005). Researchers found the reaction times of participants were much faster for regular than irregular stress patterns. Furthermore, prominent, regular stress patterns seem to improve neural entrainment in the theta range in young children (Leong et al., 2014, 2017). One possibility for the lack of differences between 'strong' and 'weak' edge conditions in the theta range in Experiment 1 was because these were matched in terms of stress patterns.

Stressed syllables have also been found to be longer and louder than other syllables (Greenberg et al., 2003). All syllables used in Experiment 2 were of the same average intensity and only contained micro-variations in duration. However, in natural, continuous speech, syllables starting with sibilants would be longer than syllables starting with plosives. Furthermore, when syllables were processed for Experiment 2, syllables starting with /b/, for example, suffered from a lesser degree of duration alteration than syllables starting with /s/. Perhaps, in natural speech, landmarks of syllables starting with 'weak' edges are provided instead by stress patterns, longer durations and greater sound intensity.

Another aspect which we have not accounted for in the present research is the preferred phase of syllabic entrainment to our stimuli. This could be useful especially when if differences in entrainment between syllable with different initial consonants are not apparent, e.g., between syllables starting with nasals stops and those starting with plosives in Experiment 2, or when looking for an effect of edge in continuous speech, such as for stimuli in Experiment 1. Exploring the phase of entrainment has been reported in other studies. For example, Zoefel and Van Rullen (2016) have shown that forward and reverse speech do not differ in the level of theta phase locking, but in the precise phases of entrainment. Another study by Power,

Mead, Barnes, & Goswami (2013) found that dyslexic children, who are thought to have difficulty in the processing of individual phonemes, which impairs phoneme-grapheme conversion during reading and writing, show a different preferred phase of entrainment to isochronous CV syllables when compared to control participants. Furthermore, the phase of entrainment seems to change during buildup, with prolonged entrainment showing phases which are closer either to 0° (in phase) or 180° (anti-phase) (Riecke et al., 2015). Thus, the phase of entrainment seems to be relevant for speech tracking, and may be informative about the role of different acoustic edges, or their differential impact on phase resetting neural oscillations to speech sounds.

In terms of the role of the CV transition for syllabic entrainment, it is possible that this might represent an attractor in the dynamics of a network of coupled oscillators which is responsible for the neural tracking of speech. In a mathematical model of a network of delayed pulse coupling, Ashwin and Timme (2018) showed that the phase of one of the oscillators resets after the firing of its cells, which occurs at discrete events in time. Furthermore, Ashwin and Timme (2018) argue that the events of phase resetting are unstable attractors: any perturbation during these events can therefore severely impact the synchronisation of the oscillators. While it is not clear the latencies of CV transitions represent attractor points for neural entrainment to speech, previous research has shown that the phase resetting of neural oscillations in the auditory cortex occurs in response to discrete events, such as the period of a musical beat (Szymanski et al., 2011). If CV transitions are indeed unstable attractors, this could explain why small perturbations at their latencies could lead to considerable disruptions (or enhancing) of entrainment, like we found in Experiment 3A. However, that remains to be determined, since the stability of

synchrony between coupled neural oscillations is not a trivial problem and remains debateable (Timme and Wolf, 2008).

Lastly, the present research has important applications in understanding dyslexia. It has been found that children with dyslexia have difficulty processing the rising amplitudes of the envelopes, across different languages (Thomson et al., 2013). Some researchers claim that envelope rise times are associated with phonemic spectro-temporal information (Tallal, 2004), while others suggest that they help convey stress patterns (Goswami & Leong, 2013). Indeed, dyslexic children seem to benefit from both phonemic and stress-training procedures Thomson et al. (2013. For example, one study by (Thomson et al., 2013) describes how both a phonemic intervention procedure, in which children matched the sounds of different syllables to a target one, and a rhythm identification procedure, in which they repeated the stress of certain words using non-word syllables such as “dee” (e.g., where the stress in Harry Potter was illustrated by “DEEdee DEEdee”) have beneficial effects on reading and writing. However, this study showed that the rhythm intervention had an advantage over the phonemic one in terms of the children’s envelope rise time discrimination. The importance of envelope rise time perception over other variables is nonetheless debateable.

It would be interesting to see if interventions tackling the CV transition more readily would also improve reading skills. While such interventions currently exist, they mainly concern altering the duration of the formant transition to be longer or shorter, and seems to lead to little improvement in dyslexic children’s reading performance (Menell, McAnally, & Stein, 1999). Perhaps our results could motivate the creation of improved methods concerning the perception of CV transition in dyslexic children. Nonetheless, the CV transitions as a main landmark for neural

entrainment to speech needs to be confirmed. Future studies could investigate multiple points in the envelope rise time as potential landmark candidates, and, as mentioned previously, this needs to be done for a variety of syllables, which span a range of different syllable-initial consonants, as well as different structures (e.g. CVC, CCV, CCVC, etc.). The P-centre literature indicates that formant transitions of syllables of other forms than CV are less reliable (R. C. Villing et al., 2011), but the implications of such findings remains unclear. Thus, the present research is but a starting point in understanding the role of landmarks in neural tracking and neural entrainment to speech.

Concluding remarks

In the present thesis, I described a series of experiments showing that different phonemes placed at the beginning of syllables lead to differences in neural phase locking to those syllables. Whereas we did not find this for continuous speech Experiment 1 replicated results from previous studies showing a higher level of phase coherence in the theta range compared to other frequencies. In Experiment 2, syllables starting with different consonants led to differences in a linear combination of the inter-trial phase coherence taken at the syllabic rate and its harmonics, with sibilants leading to the least entrainment and stops showing the highest amounts. Moreover, both phase locking and different edge markers of the envelope seemed to be explained by different articulatory features of the phonemes. In Experiment 3, differences in phase coherence between two streams of repeated syllables (“da” or “ta”) showed opposite trends when different types of noise were placed at the CV transitions of each syllable, with a click type sound enhancing entrainment for “da” streams, and a short snippet of white noise increasing entrainment for “ta” syllables.

We argue that, not only can different phonemes provide different edge markers for entrainment, but that landmarks for speech tracking may particularly be found at formant transitions. These markers may either represent speech edges which trigger phase resetting of neural oscillations or trigger higher evoked activity, depending on the type of entrainment which is considered, but these specific details remain to be confirmed. Considering phonemic aspects, particularly at the CV transitions, could have various implications for models of speech tracking and understanding speech development related conditions such as dyslexia. However, more research is needed to confirm the importance of CV landmarks for a variety of syllable-initial consonants, types of syllables and continuous speech.

References

Aeschbach, D., & Borbély, A. A. (1993). All-night dynamics of the human sleep EEG. *Journal of Sleep Research, 2(2),* 70–81. https://doi.org/10.1111/j.1365-2869.1993.tb00065.x

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences, 98(23),* 13367–13372. https://doi.org/10.1073/pnas.201400998

Apoux, F., & Bacon, S. P. (2008). Differential contribution of envelope fluctuations across frequency to consonant identification in quiet. *The Journal of the Acoustical Society of America, 123(5),* 2792–2800. https://doi.org/10.1121/1.2897916

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

Arsenault, J. S., & Buchsbaum, B. R. (2015). Distributed neural representations of phonological features during speech perception. *Journal of Neuroscience, 35(2),* 634–642. https://doi.org/10.1523/JNEUROSCI.2454-14.2015

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. (2014). Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science, 25(8),* 1546–1553. https://doi.org/10.1177/0956797614533705

Başar, E. (2013). Brain oscillations in neuropsychiatric disease. *Dialogues in Clinical Neuroscience, 15(3),* 291–300.

Berger, H. (1929). Über das elektrenkephalogramm des menschen. *Archiv für Psychiatrie und Nervenkrankheiten, 87(1),* 527–570. https://doi.org/10.1007/BF01797193

Besle, J., Schevon, C. A., Mehta, A. D., Lakatos, P., Goodman, R. R., McKhann, G. M., Schroeder, C. E. (2011). Tuning of the human neocortex to the temporal dynamics of attended events. *Journal of Neuroscience, 31(9),* 3176–3185. https://doi.org/10.1523/JNEUROSCI.4518-10.2011

Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer (Version 5.4. 08)[Computer program]. Retrieved October 13, 2016.

Bor, D., & Seth, A. K. (2012). Consciousness and the prefrontal parietal network: insights from attention, working memory, and chunking. *Frontiers in Psychology, 3,* 63. https://doi.org/10.3389/fpsyg.2012.00063

Bourguignon, M., Tiège, X. D., Beeck, M. O. de, Ligot, N., Paquier, P., Bogaert, P. V., Jousmäki, V. (2013). The pace of prosodic phrasing couples the listener's cortex to the reader's voice. *Human Brain Mapping, 34(2),* 314–326. https://doi.org/10.1002/hbm.21442

Buzsáki, G. (2005). Theta rhythm of navigation: link between path integration and landmark navigation, episodic and semantic memory. *Hippocampus, 15(7),* 827–840. https://doi.org/10.1002/hipo.20113

Buzsáki, G. (2006). *Rhythms of the Brain*. https://doi.org/10.1093/acprof:oso/9780195301069.001.0001

Capilla, A., Pazo-Alvarez, P., Darriba, A., Campo, P., & Gross, J. (2011). Steady-state visual evoked potentials can be explained by temporal superposition of transient event-related responses. *PloS one, 6(1,),* e14543.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., & Shamma, S. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *The Journal of the Acoustical Society of America, 106(5),* 2719–2732. https://doi.org/10.1121/1.428100

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America, 118(2),* 887–906. https://doi.org/10.1121/1.1945807

Cooper, A. M., Whalen, D. H., & Fowler, C. A. (1986). P-centers are unaffected by phonetic categorization. *Perception & Psychophysics, 39(3),* 187–196. https://doi.org/10.3758/BF03212490

Cummins, F. (2012). Oscillators and syllables: a cautionary note. *Frontiers in Psychology, 3,* 364. https://doi.org/10.3389/fpsyg.2012.00364

David, S. V., Mesgarani, N., Fritz, J. B., & Shamma, S. A. (2009). Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. *Journal of Neuroscience, 29(11),* 3374–3386. https://doi.org/10.1523/JNEUROSCI.5249-08.2009

Di Liberto, G. M., & Lalor, E. C. (2017). Indexing cortical entrainment to natural speech at the phonemic level: methodological considerations for applied research. *Hearing Research, 348,* 70–77. https://doi.org/10.1016/j.heares.2017.02.015

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology, 25(19),* 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Ding, N., Melloni, L., Tian, X., Zhang, H., & Poeppel, D. (2015). Cortical entrainment reflects hierarchical structure building in speech comprehension*. Nature Neuroscience*, In-press.

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience, 33(13),* 5728–5735. https://doi.org/10.1523/JNEUROSCI.5297-12.2013

Ding, Nai, Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage, 88,* 41–46. https://doi.org/10.1016/j.neuroimage.2013.10.054

Ding, Nai, Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in Human Neuroscience, 11,* 481. https://doi.org/10.3389/fnhum.2017.00481

Ding, Nai, Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience, 19(1),* 158–164. https://doi.org/10.1038/nn.4186

Ding, Nai, Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews, 81,* 181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011

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

Doelling, K., Arnal, L., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage, 85,* 761-768*.* https://doi.org/10.1016/j.neuroimage.2013.06.035

Drullman, R., Festen, J. M., & Plomp, R. R. (1994). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America, 95(5),* 2670–2680. https://doi.org/10.1121/1.409836

Edwards, E., & Chang, E. F. (2013). Syllabic (~2–5 Hz) and fluctuation (~1–10 Hz) ranges in speech and auditory processing. *Hearing Research, 305*, 113–134. https://doi.org/10.1016/j.heares.2013.08.017

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLOS Computational Biology, 5(3)*, e1000302. https://doi.org/10.1371/journal.pcbi.1000302

Fogerty, D., Humes, L. E., & Kewley-Port, D. (2010). Auditory temporal-order processing of vowel sequences by young and elderly listeners. *The Journal of the Acoustical Society of America, 127(4)*, 2509–2520. https://doi.org/10.1121/1.3316291

Fogerty, D., & Kewley-Port, D. (2009). Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *The Journal of the Acoustical Society of America, 126(2)*, 847–857. https://doi.org/10.1121/1.3159302

Friederici, A. D. (2004). Processing local transitions versus long-distance syntactic hierarchies. *Trends in Cognitive Sciences, 8(6),* 245–247.

F.R.S, K. P. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11),* 559–572. https://doi.org/10.1080/14786440109462720

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

Ghitza, O. (2013). The theta-syllable: A unit of speech information defined by cortical function*. Frontiers in Psychology, 4,*138. https://doi.org/10.3389/fpsyg.2013.00138

Ghitza, O. (2014). Behavioral evidence for the role of cortical θ oscillations in determining auditory channel capacity for speech*. Frontiers in Psychology, 5,* 652*.* https://doi.org/10.3389/fpsyg.2014.00652

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence*. Phonetica, 66(1–2),* 113–126. https://doi.org/10.1159/000208934

Gini, C. (1921). Measurement of inequality of incomes*. The Economic Journal, 31(121),* 124–126. https://doi.org/10.2307/2223319

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

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech - a syllable-centric perspective. *Journal of Phonetics, 31(3)*, 465–485. https://doi.org/10.1016/j.wocn.2003.09.005

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLOS Biology, 11(12),* e1001752. https://doi.org/10.1371/journal.pbio.1001752

Harsin, C. A. (1997). Perceptual-center modeling is affected by including acoustic rate-of-change modulations. *Perception & Psychophysics, 59(2),* 243–251. https://doi.org/10.3758/BF03211892
Hasegawa, A., Okanoya, K., Hasegawa, T., & Seki, Y. (2011). Rhythmic synchronization tapping to an audio–visual metronome in budgerigars. *Scientific Reports, 1*, 120. https://doi.org/10.1038/srep00120
Howard, M. F., & Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *Journal of Neurophysiology, 104(5),* 2500–2511.
Howard, M. F., & Poeppel, D. (2012). The neuromagnetic response to spoken sentences: co-modulation of theta band amplitude and phase. *NeuroImage, 60(4)*, 2118–2127. https://doi.org/10.1016/j.neuroimage.2012.02.028
Howell, P. (1988). Prediction of P-center location from the distribution of energy in the amplitude envelope: I. *Perception & Psychophysics, 43(1)*, 90–93. https://doi.org/10.3758/BF03208978
Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., & Giraud, A.-L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. *ELife, 4,* e06213. https://doi.org/10.7554/eLife.06213
INTERSPEECH 2014 Abstract: Leong et al. (n.d.). Retrieved 8 September 2019, from https://www.isca-speech.org/archive/interspeech_2014/i14_2563.html
Kaiser, J., & Lutzenberger, W. (2005). Human gamma-band activity: a window to cognitive processing. *NeuroReport, 16(3)*, 207.
Keitel, A., Gross, J., & Kayser, C. (2017). Speech tracking in auditory and motor regions reflects distinct linguistic features. *BioRxiv.* https://doi.org/10.1101/195941
Khalighinejad, B., Silva, G. C. da, & Mesgarani, N. (2017). Dynamic encoding of acoustic features in neural responses to continuous speech. *Journal of Neuroscience, 37(8)*, 2176–2185. https://doi.org/10.1523/JNEUROSCI.2383-16.2017
Klimesch, W., Hanslmayr, S., Sauseng, P., & Gruber, W. R. (2006). Distinguishing the evoked response from phase reset: A comment to Mäkinen et al. *NeuroImage, 29(3)*, 808–811. https://doi.org/10.1016/j.neuroimage.2005.08.041
Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The tracking of speech envelope in the human cortex. *PLOS ONE, 8(1),* e53398. https://doi.org/10.1371/journal.pone.0053398
Lachaux, J.-P., Rodriguez, E., Martinerie, J., & Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Human Brain Mapping, 8(4)*, 194–208. https://doi.org/10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C
Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology, 94(3)*, 1904–1911. https://doi.org/10.1152/jn.00263.2005
Lalor, E. C. (2018). Neuroscience: The rhythms of speech understanding. *Current Biology, 28(3)*, R105-R108.
Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience, 31(1),* 189-193. https://doi.org/10.1111/j.1460-9568.2009.07055.x

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology, 102(1),* 349–359. https://doi.org/10.1152/jn.90896.2008

Lantz, G., De Peralta, R. G., Spinelli, L., Seeck, M., & Michel, C. M. (2003). Epileptic source localization with high density EEG: how many electrodes are needed?. *Clinical neurophysiology, 114(1)*, 63-69.

Leong, V., Kalashnikova, M., Burnham, D., & Goswami, U. (2017). The temporal modulation structure of infant-directed speech. *Open Mind, 1(2),* 78–90. https://doi.org/10.1162/OPMI_a_00008

Leong, V., Kalashnikova, M., Burnham, D., & Goswami, U. (2014). Infant-directed speech enhances temporal rhythmic structure in the envelope. In *Fifteenth Annual Conference of the International Speech Communication Association.*

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique*. MIT Press.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron, 54(6),* 1001–1010. https://doi.org/10.1016/j.neuron.2007.06.004

Luo, H., & Poeppel, D. (2012). Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex. *Frontiers in Psychology, 3*. https://doi.org/10.3389/fpsyg.2012.00170

Marcus, S. M. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception & Psychophysics, 30(3)*, 247–256. https://doi.org/10.3758/BF03214280

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition, 122(3)*, 346–362. https://doi.org/10.1016/j.cognition.2011.11.003

Menell Peter, McAnally Ken I., & Stein John F. (1999). Psychophysical sensitivity and physiological response to amplitude modulation in adult dyslexic listeners. *Journal of Speech, Language, and Hearing Research, 42(4)*, 797–803. https://doi.org/10.1044/jslhr.4204.797

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343(6174)*, 1006–1010.

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. *European Journal of Neuroscience, 48(7)*, 2609–2621. https://doi.org/10.1111/ejn.13748

Michalewski, H. J., Prasher, D. K., & Starr, A. (1986). Latency variability and temporal interrelationships of the auditory event-related potentials (N1, P2, N2, and P3) in normal subjects. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 65(1)*, 59–71. https://doi.org/10.1016/0168-5597(86)90037-7

Molinaro, N., & Lizarazu, M. (2018). Delta(but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience, 48(7)*, 2642–2650. https://doi.org/10.1111/ejn.13811

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

Notbohm, A., Kurths, J., & Herrmann, C. S. (2016). Modification of brain oscillations via rhythmic light stimulation provides evidence for entrainment but not for

superposition of event-related responses. *Frontiers in human neuroscience, 10*, 10.
Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., Brugge, J. F. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *Journal of Neuroscience, 29(49),* 15564–15574. https://doi.org/10.1523/JNEUROSCI.3065-09.2009
Nozaradan, S., Peretz, I., Missal, M., & Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *Journal of Neuroscience, 31(28)*, 10234-10240.
Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in cognitive sciences, 23(11),* 913-926*.*
Oganian, Y., & Chang, E. F. (2018). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *BioRxiv*, 388280. https://doi.org/10.1101/388280
Pasley, B. N., & Knight, R. T. (2013). Decoding speech for understanding and treating aphasia. *Progress in Brain Research, 207*, 435–456. https://doi.org/10.1016/B978-0-444-63327-9.00018-7
Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology, 3,* 320*.* https://doi.org/10.3389/fpsyg.2012.00320
Peelle, J. E., Gross, J., & Davis, M. H. (2012). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex, 23(6),* 1378–1387.
Penny, W. D., Duzel, E., Miller, K. J., & Ojemann, J. G. (2008). Testing for nested oscillation. *Journal of Neuroscience Methods, 174(1)*, 50–61. https://doi.org/10.1016/j.jneumeth.2008.06.035
Penttonen, M., & Buzsáki, G. (2003). Natural logarithmic relationship between brain oscillators. *Thalamus & Related Systems, 2(2)*, 145–152. https://doi.org/10.1016/S1472-9288(03)00007-4
Pérez, A., Carreiras, M., Dowens, M. G., & Duñabeitia, J. A. (2015). Differential oscillatory encoding of foreign speech. *Brain and Language, 147*, 51–57.
Pérez, A., Dumas, G., Karadag, M., & Duñabeitia, J. A. (2019). Differential brain-to-brain entrainment while speaking and listening in native and foreign languages. *Cortex, 111*, 303–315. https://doi.org/10.1016/j.cortex.2018.11.026
Picton, T. W., John, M. S., Dimitrijevic, A., & Purcell, D. (2003). Human auditory steady-state responses: Respuestas auditivas de estado estable en humanos. *International Journal of Audiology, 42(4)*, 177–219. https://doi.org/10.3109/14992020309101316
Pikovsky, A., & Rosenblum, M. (2007). Synchronization. *Scholarpedia, 2(12)*, 1459. https://doi.org/10.4249/scholarpedia.1459
Port, R. F. (2003). Meter and speech. *Journal of Phonetics, 31(3–4)*, 599–611. https://doi.org/10.1016/j.wocn.2003.08.001
Power, Alan J., Colling, L. J., Mead, N., Barnes, L., & Goswami, U. (2016). Neural encoding of the speech envelope by children with developmental dyslexia. *Brain and Language, 160*, 1–10. https://doi.org/10.1016/j.bandl.2016.06.006
Power, Alan James, Mead, N., Barnes, L., & Goswami, U. (2013). Neural entrainment to rhythmic speech in children with developmental dyslexia. *Frontiers in Human Neuroscience, 7,* 777. https://doi.org/10.3389/fnhum.2013.00777

Prendergast, G., Johnson, S. R., & Green, G. G. R. (2010, November 1). Temporal dynamics of sinusoidal and non-sinusoidal amplitude modulation. *European Journal of Neuroscience, 32(9),* 1599-1607.https://doi.org/10.1111/j.1460-9568.2010.07423.x

Riecke, L., Sack, A. T., & Schroeder, C. E. (2015). Endogenous delta/theta sound-brain phase entrainment accelerates the buildup of auditory streaming. *Current Biology, 25(24)*, 3196–3201. https://doi.org/10.1016/j.cub.2015.10.045

Rosen Stuart, Carlyon Robert P., Darwin C. J., & Russell Ian John. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 336(1278)*, 367–373. https://doi.org/10.1098/rstb.1992.0070

Rosenblum, M. G., Pikovsky, A. S., & Kurths, J. (1996). Phase synchronization of chaotic oscillators. *Physical Review Letters, 76(11),* 1804–1807. https://doi.org/10.1103/PhysRevLett.76.1804

Schadow, J., Lenz, D., Thaerig, S., Busch, N. A., Fründ, I., & Herrmann, C. S. (2007). Stimulus intensity affects early sensory processing: Sound intensity modulates auditory evoked gamma-band activity in human EEG. *International Journal of Psychophysiology, 65(2)*, 152–161. https://doi.org/10.1016/j.ijpsycho.2007.04.006

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

Shower, E. G., & Biddulph, R. (1931). Differential pitch sensitivity of the ear. *The Journal of the Acoustical Society of America, 3(2A)*, 275–287. https://doi.org/10.1121/1.1915561

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

Steinschneider, M., Nourski, K. V., & Fishman, Y. I. (2013). Representation of speech in human auditory cortex: is it special? *Hearing Research, 305*, 57–73.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America, 111(4),* 1872–1891. https://doi.org/10.1121/1.1458026

Srinivasan, R. (1999). Spatial structure of the human alpha rhythm: global correlation in adults and local correlation in children. *Clinical Neurophysiology, 110(8)*, 1351-1362.

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

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science, 150(3700),* 1187–1188. https://doi.org/10.1126/science.150.3700.1187

Szymanski, F. D., Rabinowitz, N. C., Magri, C., Panzeri, S., & Schnupp, J. W. H. (2011). The laminar and temporal structure of stimulus information in the phase of field potentials of auditory cortex. *Journal of Neuroscience, 31(44),* 15787–15801. https://doi.org/10.1523/JNEUROSCI.1416-11.2011

Tallal, P. (2004). Improving language and literacy is a matter of time. *Nature Reviews Neuroscience, 5(9)*, 721–728. https://doi.org/10.1038/nrn1499

Ten Oever, S., Schroeder, C., Poeppel, D., van Atteveldt, N. M., & Zion-Golumbic, E. (2014). Neural entrainment to auditory rhythm prior to detection is linked to perception.

Thomson, J. M., Leong, V., & Goswami, U. (2013). Auditory processing interventions and developmental dyslexia: a comparison of phonemic and rhythmic approaches. *Reading and Writing, 26(2)*, 139–161. https://doi.org/10.1007/s11145-012-9359-6

Thut, G., Schyns, P., & Gross, J. (2011). Entrainment of perceptually relevant brain oscillations by non-invasive rhythmic stimulation of the human brain. *Frontiers in Psychology, 2*, 170.

Vaughan Jr., H. G. (1969). The relationship of brain activity to scalp recordings of event-related potentials. In *Average Evoked Potentials: Methods, Results, and Evaluations* (pp. 45–94). https://doi.org/10.1037/13016-002

Villing, R. (2004). Automatic blind syllable segmentation for continuous speech. https://doi.org/10.1049/cp:20040515

Villing, R. C., Repp, B. H., Ward, T. E., & Timoney, J. M. (2011). Measuring perceptual centers using the phase correction response. *Attention, Perception, & Psychophysics, 73(5)*, 1614–1629. https://doi.org/10.3758/s13414-011-0110-1

Vos, P. G., Mates, J., & van Kruysbergen, N. W. (1995). The perceptual centre of a stimulus as the cue for synchronization to a metronome: evidence from asynchronies. *The Quarterly Journal of Experimental Psychology Section A, 48(4)*, 1024–1040. https://doi.org/10.1080/14640749508401427

Włodarczak, M., Sˇimko, J., & Wagner, P. (2012). Syllable boundary effect: temporal entrainment in overlapped speech. In *Speech Prosody 2012.*

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials (ERPs) in studies of perception and attention*. Attention, Perception & Psychophysics, 72(8)*. https://doi.org/10.3758/APP.72.8.2031

Wulff, P., Ponomarenko, A. A., Bartos, M., Korotkova, T. M., Fuchs, E. C., Bahner, F., Monyer, H. (2009). Hippocampal theta rhythm and its coupling with gamma oscillations require fast inhibition onto parvalbumin-positive interneurons. *Proceedings of the National Academy of Sciences, 106(9),* 3561–3566. https://doi.org/10.1073/pnas.0813176106

Xue, J., Lee, C., Wakeham, S. G., & Armstrong, R. A. (2011). Using principal components analysis (PCA) with cluster analysis to study the organic geochemistry of sinking particles in the ocean. *Organic Geochemistry, 42(4)*, 356–367. https://doi.org/10.1016/j.orggeochem.2011.01.012

Zarco, W., Merchant, H., Prado, L., & Mendez, J. C. (2009). Subsecond timing in primates: comparison of interval production between human subjects and rhesus monkeys. *Journal of Neurophysiology, 102(6),* 3191–3202. https://doi.org/10.1152/jn.00066.2009

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,* 274. https://doi.org/10.3389/fnhum.2016.00274

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron, 77(5),* 980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. *Current Biology, 28(3),* 401-408.
Zoefel, B., ten Oever, S., & Sack, A. T. (2018). The involvement of endogenous neural oscillations in the processing of rhythmic input: more than a regular repetition of evoked neural responses. *Frontiers in neuroscience, 12*, 95.
Zoefel, B., & VanRullen, R. (2015). Selective perceptual phase entrainment to speech rhythm in the absence of spectral energy fluctuations. *Journal of Neuroscience, 35(5)*, 1954–1964.
Zoefel, B., & VanRullen, R. (2016). EEG oscillations entrain their phase to high-level features of speech sound. *NeuroImage, 124*, 16–23. https://doi.org/10.1016/j.neuroimage.2015.08.054

Appendix 1

A1.1.1 A brief account of electrical activity in the brain

The majority of animals are able to receive information about the world and subsequently take actions through the means of a nervous system. Nervous systems usually centralise in dense control units, such as brains. The cells responsible for the brain's transmission of information are neurons, which communicate with each other through electric impulses and the release of chemicals called neurotransmitters.

An action potential occurs when the negatively-charged cell membrane depolarizes, or becomes more positive, than a specific threshold level. For neurons, this is usually around -55 mV. This is due to the influx of positive sodium ions into the neuron's membrane, which is normally caused by external stimulation. Once the membrane's voltage reaches threshold, the neuron fires, which implies that the membrane continues to depolarise towards 0 mV even in the absence of continued stimulation, after which the voltage decreases again. This spike in voltage is called the action potential, and is described in Figure A1.1.

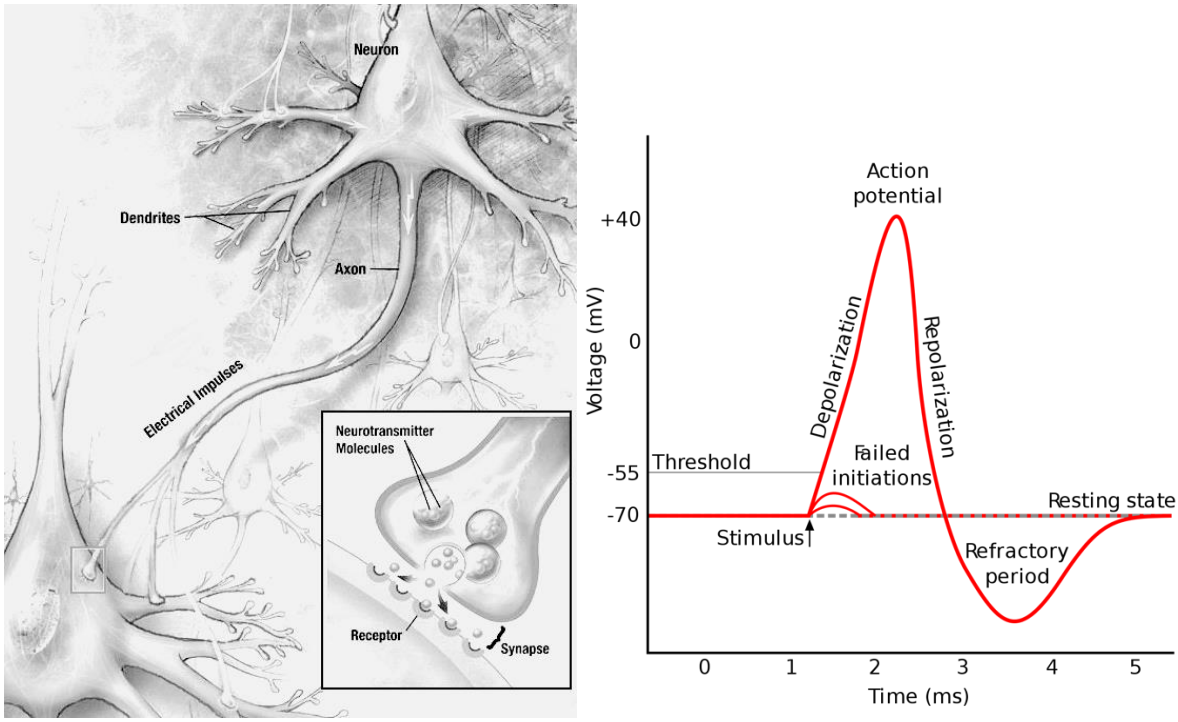


Figure A1.1. A. Starting with the top neuron, arrows show how electrical impulses enter the dendrites, into the cell body, and travel down the axon until they reach the dendrites of a connecting neuron, located at the bottom left. The picture on the bottom right illustrates schematically the neurotransmitter release associated an action potential, at the synaptic site formed by the axon terminal of the top neuron and the dendrite receptors of the bottom neuron. B. The voltage of a neuron before, during and immediately after an action potential. This is initially at -70 mV during the resting state, increases when stimulation (e.g., an external current) is applied, and only continues to increase when a threshold (-55 mV) is reached. After the spike there is typically a refractory period, often about 5 ms during which the neurons cannot fire; often there is also a short hyperdepolarization when the voltage is lower than the resting voltage. This slowly increases back to its resting state. (Both pictures are downloaded from Wikipedia.org)

The action potential travels from the cell body to the synapses, where the axon terminates and where neurotransmitters are released into the synaptic cleft (the space between neurons). The release of neurotransmitters leads to another electrical impulse which can be recorded from the extracellular space, called the postsynaptic potential. The neurotransmitters bind onto the membrane receptors of the neighbouring neuron. This can either depolarise the membrane, subsequently leading to an action potential being fired by the postsynaptic neuron, or, on the contrary, it can hyperpolarise it, or prevent it from firing. In the first case, the postsynaptic potential is excitatory, and in the second case, it is inhibitory.

Action potentials can be recorded both directly from the neuron or from the extracellular space. By placing a probe so that it is close to multiple neurons, one can record multi-unit activity (Luck, 2014). Local field potentials (LFPs) can also be recorded from the extracellular space between neurons. These are sustained, low-frequency currents which are believed to arise from the synchronised input of multiple cells into the recorded area. It is thought that summed postsynaptic potentials from neuronal assemblies give rise to LFPs and can also be recorded from distal locations, such as the scalp (Luck, 2014).

Action potentials are short-lived and hardly occur in different neurons at the same time, implying that they will cancel out when one records from a neighbouring location, and not directly from the neurons. However, postsynaptic potentials last longer and are instantaneous, which increases the chances that such potentials coming from different units will summate, and it is this summed activity which will be easier to record from a distance (Luck, 2014). Therefore, scientists believe that it is not the action potentials, but the postsynaptic potentials of neurons which give rise to scalp electrical activity.

The frequency of a neural rhythm depends on the size of the neural assembly which is responsible for it, with small neuronal groups giving rise to higher oscillations and large networks eliciting slower rhythms. However, single cells which are part of a single oscillatory network need not have the same intrinsic frequency: during neural oscillations, individual neurons don't exhibit the full oscillation but their activity becomes entrained to the rhythm. Thus, if a cell displays a tempo which is either behind or ahead of the global frequency of oscillation, the other cells in the

population will force it to either keep up or slow down with the common rhythm (Buzsáki, 2006).

It is believed that oscillatory patterns correspond to temporal windows of neural information processing, which are used by the brain as means of self-organisation, in order to predict periodic events (Arnal & Giraud, 2012). Information patterns could originate either internally, or in the surrounding environment, as stimuli. However, the exact mechanisms that lead to the oscillations recorded at scalp level are not fully known. For example, the multiple oscillatory rhythms which have been discovered empirically, in both humans and animals, seem to be attributed with different behaviours or neural functions. Furthermore, the synchrony between separate rhythms, which has also been found consistently on certain occasions, may be suggestive, perhaps, of hierarchical neural processes (Penttonen & Buzsáki, 2003).

Different parts of the brain can also give rise to the same rhythm, with the same brainwaves being seen across multiple regions of the scalp (Srinivasan, 1999). Most often than not, the rhythms recorded at electrodes placed in different areas are correlated with one another, indicating that rhythms occurring in different areas of the brain are synchronised even across long distances. Scientists believe that this happens due to multiple cell populations communicating with one another, through rhythmic patterns of excitation and inhibition in the post-synaptic activity of numerous cells (Buzsáki, 2006).

In sum, the brain's electrical activity is highly complex and happens at different scales, both spatial and temporal. This is illustrated by short-lived, cell-to-cell communication in the form of action potentials, as well as prolonged postsynaptic potentials, both excitatory and inhibitory, which are synchronised

across multiple cell populations, giving birth to neural oscillations. Neural oscillations span different rhythms, which can be seen as successive temporal windows, whose regularity helps the brain in self-organisation and information prediction. In the following section, we will refer how neural oscillations arising at scalp level can be recorded, with a focus on electroencephalography (EEG).

A.1.1.2 Recording the brain's electrical activity

Electrical activity was first recorded from the scalp of humans in 1929, by Hans Berger, who invented the method known as electroencephalography (EEG). By placing several clay electrodes at the occipital site of the skull, he recorded two distinct rhythms: one which was visible when the eyes were closed, and one when the eyes were open (Haas, 2003). He named these two different rhythms alpha and beta, respectively. Initially, the currents were measured using a galvanometer, which is an electromechanical instrument. The EEG method is still very widely used today, but EEG caps are made from up to hundreds of light-weight electrodes, and the signals they record are amplified and digitally processed by computers.

The advantage of EEG is its fine temporal resolution, which helps measure neural processes at the level of millisecond. However, the EEG does not boast very good spatial resolution, i.e., it is not easy to identify which brain regions trigger the activity measured at the scalp. The difficulty in establishing the sources of the electric signals is partly due to the orientation of the neurons in the brain, which leads to some currents cancelling out, even though some of them may be responsible for the observed activity (Luck, 2014). This is called the inverse problem and, which is difficult to solve because combinations of different orientations are virtually limitless. Secondly, neural processes are known to be nonlinear, which

means, the outcome is more than simply the sum of its parts (Buzsáki, 2006). Even if the summation of postsynaptic potentials is possible, the complex interplay between excitation and inhibition, which cannot be inferred by merely looking at an oscillation, as well as the interference of different rhythms, makes it difficult to determine how, or what, is responsible for the measured signals.

Source localisation can be attempted by means of complex algorithms which take into consideration the orientation of multiple cells. However, these require many electrodes, (e.g., Lantz, Grave de Peralta, Spinelli, Seeck, & Michel, 2003), which can cause additional problems such as bridging, etc. A way of recording directly from the neural location of interest is through electrocorticography (ECOG), which involves attaching electrodes to the open cortices of patients undergoing brain surgery. A more sophisticated, non-invasive method is achieved through magnetoencephalography (MEG). MEG uses magnometers such as SQUIDS (superconducting quantum interference devices) to detect the weak magnetic fields which are triggered at the same time as the brain's electrical activity. MEG can target fields of a specific orientation, which makes it better than EEG at source localisation, but only slightly so, and is more expensive than EEG. Nonetheless, EEG, MEG and ECOG have all helped shed light on numerous brain processes and continue to do so, and we referred to research involving such techniques extensively throughout this thesis.

A.1.1.3 Oscillatory rhythms in the brain

Oscillations in the brain arise as the a consequence of the synchronised activity of multiple neurons, occurring in the form of temporal patterns of excitation and inhibitions across neuronal assemblies (Buzsáki, 2006). When taking place over

large groups of neurons, oscillations are macroscopic and can be recorded using techniques such as EEG. Neuronal rhythms happen at various time scales and therefore seem to come in separate frequency bands: 0.5 – 4 Hz (delta), 4 – 8 Hz (theta), 8-12 Hz (alpha), beta (12 – 30 Hz) and gamma (> 30 Hz), with the boundaries between these bands being somewhat arbitrary and having been determined empirically over time (Buzsáki, 2006). The first discovered neural rhythm is the alpha rhythm, which was measured by Hans Berger (who also coined the term 'electroencephalography') at occipital electrode sites, when the eyes of the person were closed (Berger, 1929). He also discovered the beta rhythm, during states of alertness, when the alpha rhythm was supressed. In general, it seems that different neural rhythms may serve different cognitive functions (e.g., memory, perception) or states (e.g., alertness or consciousness). Roughly, it seems that high frequency oscillations (e.g., gamma) are observed during cognitive processing (e.g., Kaiser & Lutzenberger, 2005), while non-alert states such as sleep are marked by a change towards slow rhythms (i.e., alpha or delta) (Aeschbach & Borbély, 1993).

Different neural rhythms may also be involved in the same general psychological function, but are known to be generated by different mechanisms. Both hippocampal theta and gamma are observed during memory encoding in the rat, but the theta rhythm also occurs at the same time as faster oscillations during rapid-eye movement sleep, when the gamma rhythm is not present (Penttonen & Buzsáki, 2003). Furthermore, it has been shown that the theta rhythm is crucially dependent on inhibitory interneurons but gamma activity can still be noticed if the inhibitory behaviour of the same neurons is blocked (Wulff et al., 2009).

Not all rhythms have known mechanisms, however. In fact, most of the neural processes leading to different oscillations are not known. Furthermore, different

mechanisms can give rise to the same neural rhythms: these can be recorded from different parts of the brain in both awake and anaesthetised animals (Penttonen & Buzsáki, 2003). If these rhythms are evoked by external stimulation, their superposition leads to the appearance of event-related potentials (Başar, 2013). When they occur concomitantly, they can become synchronised by aligning their phases, and are coherent. We will briefly cover event-related potentials (ERPs) in the section below.

A1.1.4. Event-related potentials

The term ‘ERP’ was first proposed by Vaughan Jr. (1969) in order to describe an EEG event which can be reliably associated with a “specific time reference”, such as the onset of external stimulation. The most widely investigated are visual and auditory evoked potentials, which show as responses to stimuli in their respective modalities. In both cases, early components appear as positive or negative changes in voltage shortly after stimulus onset (approximately 100 ms). These are sometimes named as N1 (negative) or P1 (positive) and are thought to indicate the perception of attended stimuli (Woodman, 2010).

ERPs do not just reflect stimulus-response type of processes. Top down mechanisms such as attention crucially impact the size and duration of ERPs and some of them appear as a result of cognitive interpretations, including, possibly, predictions about the outside world (Luck, 2014). A classical such experiment involves the discovery of the P3 component, which occurs 300 ms after stimulus onset, by Sutton, Braren, Zubin, and John (1965). The P3 was associated with an inability to predict whether the stimulus was auditory or visual, because its size was

greater than when the modality of the stimulus mismatched the participants' expectations.

The electrical activity of the brain recorded at scalp level can reflect both responses to attended external stimuli, in the form of early ERP components, as well as more complex cognitive processes, at later post-stimulus latencies. ERPs are also illustrative of the fact that the brain uses information in order to predict events, but, as we shall see more in depth in the next sections, prediction, especially that of events which occur with temporal regularity, can be achieved at an oscillatory level through synchronisation between the brain and the stimulus.

A1.1.5 Oscillations and synchronisation

An oscillation describes something that as it varies, repeats the same values. A single repetition is called a period. Oscillations show regularities across successive periods, which can be described as a rhythm. If the rhythm is constant, we can say that the oscillation is periodic. However, most oscillations, such as naturally occurring ones, either change their rhythm with time or do not show perfect regularities, and are known as non-periodic or quasi-periodic. Neural oscillations, as well as speech, belong to this group of natural, quasi-periodic type.

A simple example of an oscillation is a sine wave, or a sinusoid, like the one in Figure A1.4.A. The sinusoid is periodic because it varies constantly around the equilibrium, i.e., the variation around zero can be seen to repeat in identical cycles. The amplitude of the sine is given by the distance from equilibrium and its phase is represented by it position. The period, amplitude and phase of an oscillation are important features of oscillations and will be mentioned frequently in the present

thesis, and can be used when referring to periodic and non-periodic oscillations alike.

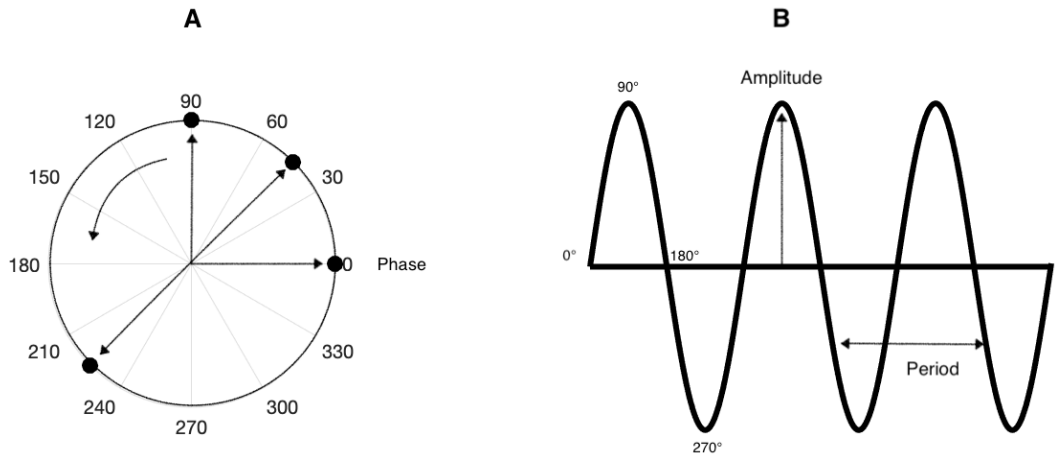


Figure A1.4. A simple example of an oscillation can be a dot moving constantly around a circle, as in A. The phase of the movement is the angle the dot makes with respect to the centre of the circle. The movement of the dot can be described by a sinusoid, represented in B. The amplitude of the dot's movement is the amplitude of the sinusoid. The period represents the amount of time which the dot takes to go around the entire circle, completing one sinusoidal cycle. The sinusoid above has three different cycles. The phases in A correspond to different points of a sinusoidal cycle. The phases also repeat with every cycle.

A.1.1.6 Properties of oscillations

The Fourier Transform

Any time-varying signal can be decomposed into multiple sinusoids at each of the frequencies existing in the signal, by means of the Fourier transform. The Fourier transform rewrites the function f(t) as in described in equation 1.1, where $e^{ikt} = \cos(kt) + i\sin(kt)$ is a periodic function with frequency $k/2\pi$ and $\tilde{f}(t)$ can be thought of as `amount' of $f(t)$ which corresponds to frequency $k/2\pi$.

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{f}\,(k)\, e^{ikt}\, dt \quad (1.1),$$

Since the decomposition is over different frequencies where the frequency is a real number, the decomposition involves an integral rather than a sum. In fact $\tilde{f}(t)$ is a complex number, its phase is related to the phase of the frequency $k/2\pi$ part of the oscillation, and the magnitude represents the amplitude of the sinusoid of that particular frequency, or the amount of the respective frequency in the signal (Picton et al., 2003). The magnitude is known as the power of the Fourier transform. If the sinusoid in Figure A1.4.A completes a cycle within 1 second, it will have a period T of 1 second and a frequency 1/T of 1 Hz. Because this is a perfectly periodic oscillation, its power spectrum will show a single peak at 1 Hz, as its only discernible frequency. This can be seen in in Figure A1.5.

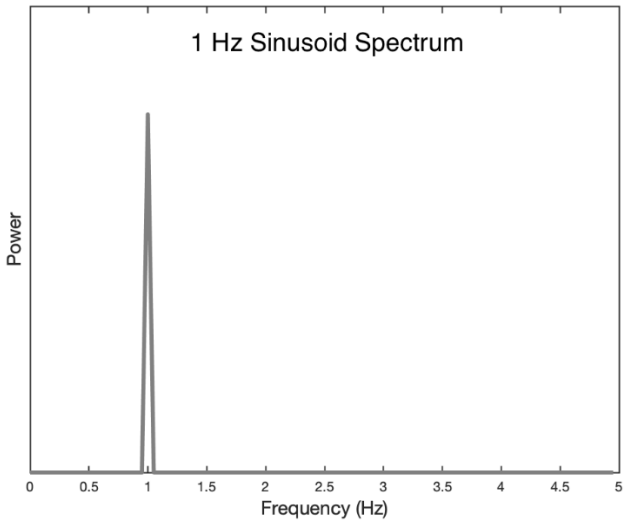


Figure A1.5. The power or magnitude of the Fourier transform of a 1 Hz sinusoid shows a single peak at 1 Hz. Reproduced after Zhou et al. (2016).

Zhou, Melloni, Poeppel, and Ding (2016) illustrate, in a series of examples, the power spectrums of different combinations of oscillations. Most temporal signals

are not perfect sinusoids, exhibiting a wide range of frequencies. The lowest frequency at which such signals oscillate is called the fundamental frequency, or f0. When looking at the power spectrum of complex periodic signals, one notices peaks not only at f0, but also at multiples of f0, such 2f0, 3f0, etc., or even 1/2f0. These peaks are called harmonics. Importantly, harmonic peaks in the power spectrum can sometimes take higher values than that of the fundamental frequency, which may, in certain cases, be absent altogether.

This may not be perfectly intuitive, but Zhou et al. (2016) explain this using a simple example. The signal in Figure A1.6.A describes a single cycle of a 5 Hz oscillation, which is repeated every second. Its fundamental frequency is thus 1 Hz, because it is the lowest in the signal. The power spectrum of this signal, shown in Figure A1.6.B, shows peaks at 1, 2 and up to 6 Hz, the tallest being at 4 Hz. Therefore, the harmonic peak at 4 Hz is taller than the one at the fundamental frequency of 1 Hz.

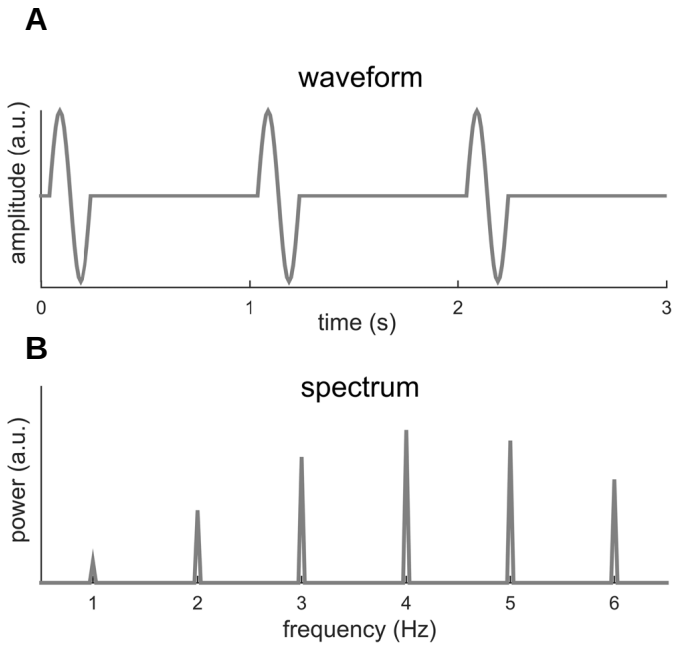


Figure A1.6. A. A single cycle of a 5 Hz sinusoid is repeated every 1 Hz (every second), for three seconds. The fundamental frequency F0 of this oscillations is 1 Hz. B. The power spectrum of the waveform in A. There is only a small peak at F0, but harmonics of this frequency show higher peaks, at 2-6 Hz, peaking at 4 Hz. In both A and B, 'a.u.' on the y axes stands for 'arbitrary units'. (Image reproduced from Zhou et. al (2016), as found on frontiersin.org)

Zhou et al. (2016) indicate that sometimes, a higher frequency oscillation can be amplitude modulated at a slower frequency, i.e., the amplitude of the peaks in the higher frequency oscillation rise and fall following a rhythmic pattern imposed by the slower oscillation. This can be seen in Figure A1.7.A, where a continuous 20 Hz oscillation is modulated at 1 Hz. However, the power spectrum of this waveform does not contain a peak at 1 Hz, but only a large peak at 20 Hz, and smaller ones at 19 and 21 Hz. In order to see the peak at 1 Hz, one needs to extract the envelope of the signal, which contains its low frequency amplitude modulations. This would lead to obtaining another 1 Hz sinusoid with the same power spectrum as illustrated in Figure A1.5.

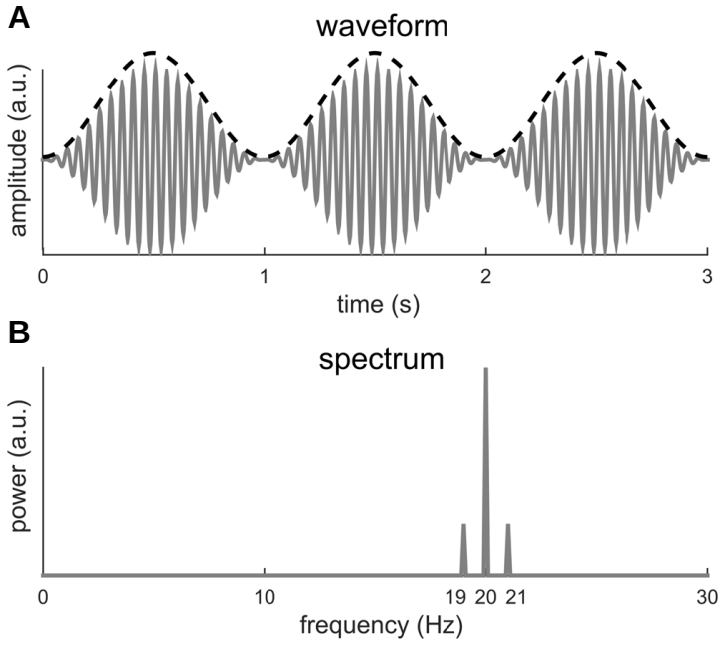


Figure A1.7. A. A 20 Hz waveform (continuous line) is amplitude modulated at 1 Hz (dotted line), implying that the amplitude of the 20 Hz signal rises and falls every second.
B. The power spectrum of the waveform in A. This only shows peaks at 20 Hz, then 19 and 21 Hz. In both A and B, 'a.u.' on the y axes stands for 'arbitrary units'. (Image reproduced from Zhou et. al (2016), as found on frontiersin.org)

Sometimes, the peaks in power of neural oscillations, especially as recorded by EEG and MEG, are difficult to notice. This is because the lower frequencies are better represented in the brain than higher ones. While this may seem confusing, lower frequencies are thought to correspond to the slower communication across

larger cell assemblies, and higher frequencies, to the fast oscillations of local populations (Buzsáki, 2006). Depending on the size of the population, it makes sense for some oscillations to have greater intensity than others. When these oscillations are measured, they show a spectrum where the power decreases from the lower frequencies to the higher ones.

Interestingly, the power decreases proportionally to the inverse of the frequency, or, mathematically, $A \sim 1/f^{\alpha}$, where A is the power, f is the frequency, and α is an arbitrary exponent (Buzsáki, 2006). This is known as the $1/f$ behaviour of EEG (Figure A1.8). When α is particularly large (the value of the exponent can vary based on individual differences), the peaks in power due to stimulation, especially lower harmonic peaks, can be especially difficult to see. The $1/f$ behaviour is nevertheless considered to be internal noise, and can be removed using linear regression, after taking the logarithm of both the power and frequency (this linearizes their relationship).

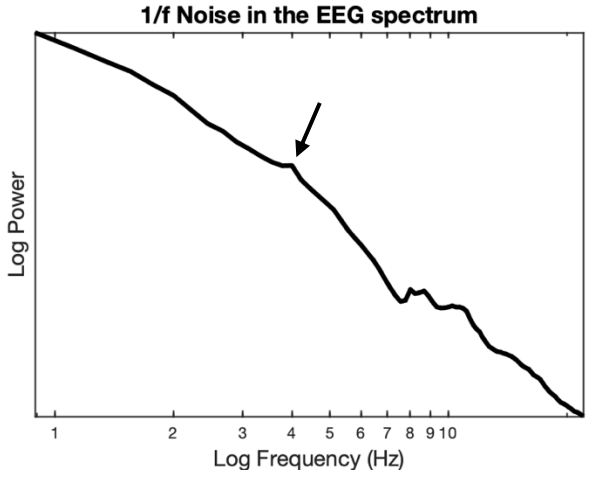


Figure A1.8. A log-log plot showing the power spectrum of EEG, obtained by taking the logarithms of both the power and the frequency. Note the overrepresentation of low frequencies in the spectrum. The power decreases almost linearly with respect to the log frequency. The arrow points to a small peak at 4 Hz, which can be made visible after linear regression. Reproduced after Buzsáki (2006).

Synchronisation and Coherence

In nature, a periodic force which acts upon an oscillator can cause this to change its rhythm (or phase-reset it) according to the one exhibited by the enacting force (Thut et al., 2011). Entrainment, phase locking and synchronisation are all terms which describe this phenomenon (Pikovsky & Rosenblum, 2007).

For example, two oscillators can exhibit the same rhythm but be initially out of phase. In time, if weak force acts between the two, they will start oscillating in phase with one another, or their phase difference will gradually become smaller, and remain constant. The two oscillators are now coupled. Sometimes, coupled oscillators will also exhibit the same amplitude but the synchronisation of natural, nonlinear oscillators can be explained more robustly by consistencies in phase than in amplitude (Pikovsky & Rosenblum, 2007).

Entrainment could happen in the case of two pendulums which are placed on the same wall: the vibrations through the wall produced by each pendulum, which fluctuate rhythmically, act as weak forces between the two pendulums, leading to the coupling of the pendulums. In humans, someone tapping their finger to a regular sound beat can be considered evidence of entrainment. This can only happen if first, neural oscillations phase lock to the rhythm of the stimulus, in order to be able to predict the timing of its occurrences.

Coherence is used to measure the amount of synchronisation between two oscillators, such as two neural rhythms, or a neural oscillation and a periodic stimulus. This is calculated based on the phase information at each frequency, which can be obtained from the Fourier transform. The phase represents the argument of the Fourier transform, and can be obtained means of the function $atan2$, or the two-argument arctangent function. Thus, the phase is calculated as the argument of the

complex number $e^{ikt} = \cos(kt) + i\sin(kt)$ in equation 1, which is $atan2(\sin(kt), \cos(kt))$. The phase coherence between two oscillators measures their phase difference over time, with more constant phase offsets leading to more coherence. A picture of two oscillators becoming coherent is illustrated in Figure A1.9, where one can notice the phase between them becoming zero and staying entirely constant. However, such a case of perfect coherence does not happen for natural, chaotic oscillators, in the sense the phase difference does not remain perfectly constant over time (see Rosenblum, Pikovsky, & Kurths, 1996).

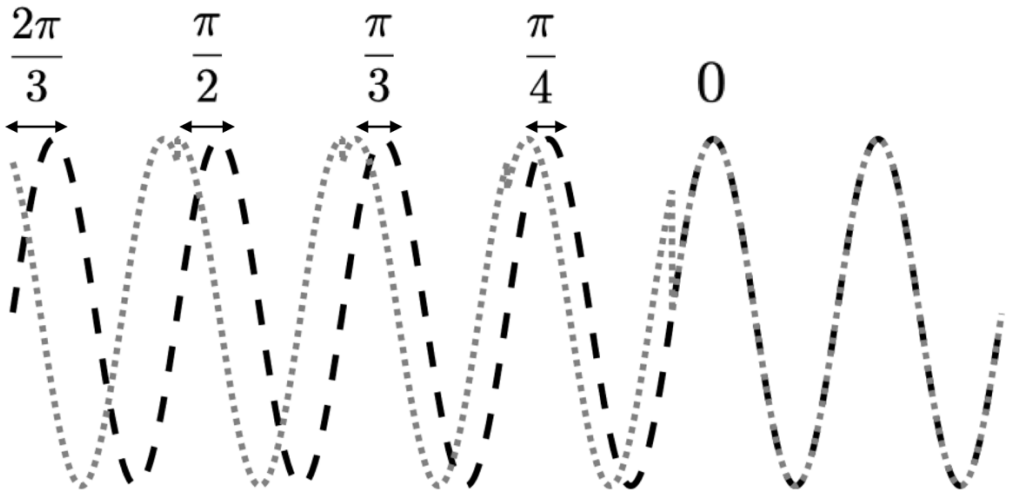


Figure A1.9. Depiction of two oscillators becoming coherent. Dashed black line represents an oscillating force which remains constant. Dotted grey line represents an oscillators which becomes phase locked to the external force. Note there is some jitter at moments of phase resetting: this is merely a consequence of the fact that this example concerns periodic oscillators. However, this could be observed when recording continuous EEG activity (Klimesch et al., 2006).

Furthermore, the phase difference between oscillators does not need to be zero in order for these to be coherent. Such is the case of the oscillators depicted in Figure A1.10, where the phase difference between the two oscillators remains the same at various angles (for the sake of simplicity, we chose to illustrate here only constant phase differences).

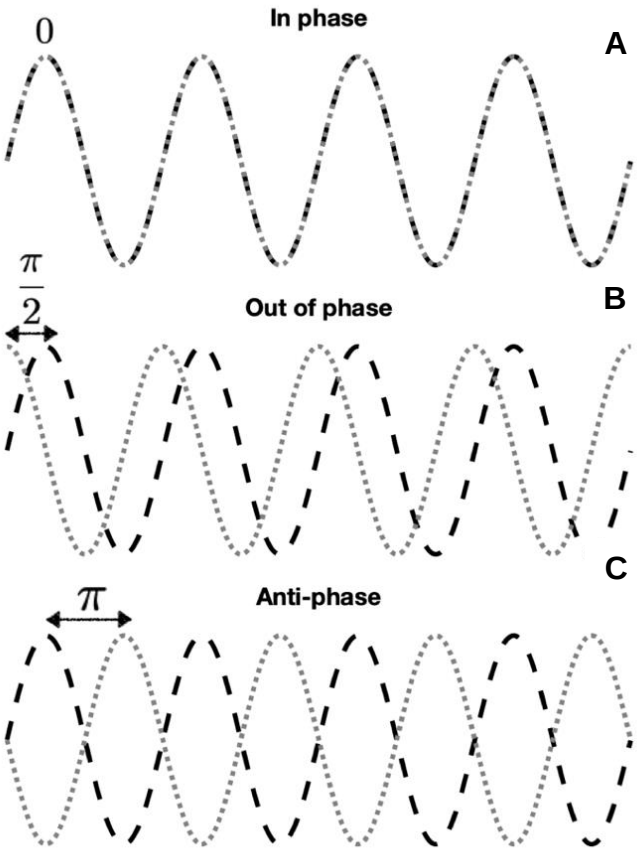


Figure A1.10. A. Two sinusoidal oscillators are in phase – they are imposed over one another and their phase difference is 0°. B. Oscillators are 90° out of phase. C. The two oscillators are in anti-phase, or 180° out of phase. The phase difference in constant in A, B and C, therefore the two oscillations are coherent in each of the three cases.

Entrainment of neural oscillations

We mentioned before that in the brain, two rhythms of the same frequency can be generated by different parts of the brain, i.e., by different neural assemblies. In general, these rhythms are synchronised with one another: EEG electrodes positioned at different locations on the scalp can show phase locking between their respective oscillations at a given frequency (Rosenblum, Pikovsky, & Kurths, 1996). Brainwaves can also become entrained to periodic stimuli which act as external forces. Synchronisation to an external stimulus leads to an alteration in the phase, as well as an increase in the amplitude of the neural oscillations (Thut et al., 2011).

Importantly, the rules of entrainment do not imply that the phase of brainwaves will exactly match that of the external force (Pikovsky & Rosenblum, 2007).

When perfectly periodic visual or auditory stimuli are presented, the brain responds with steady-state oscillations: these are evoked potentials of constant amplitude and phase and can be observed beyond the period of stimulation (Picton et al., 2003). Their steady-state nature makes them easy to study. For example, potentials in response to periodic visual stimuli are easily recognisable in an ongoing EEG (Picton et al., 2003). However, that is not always the case. For example, in the case of the well-known auditory-steady state potential or the ASSR, the background activity needs to be subtracted from the studied waveforms through a variety of averaging methods, in order for evoked responses to periodic stimuli to be observed (Geisler, 1960).

While the phase-resetting of endogenous oscillations to an external stimulus may be noticeable in the ongoing EEG activity (Klimesch, Hanslmayr, Sauseng, & Gruber, 2006), time-frequency analyses, for example based on the Fourier transform, spectrogram or wavelet analyses, such as the power or phase coherence are also able to quantify the extent of synchronisation between neuronal oscillations, or between these and the stimulus. The brain also shows what is known as nested oscillations, by which increases in the amplitude of a faster rhythm occur at the frequency of a slower rhythm, i.e., the fast oscillations are coupled to the slow oscillations (Penny et al., 2008). This also seems to be the case for the neural processing of auditory stimuli, including speech, which is characterised by a hierarchical coupling of the delta, theta and gamma rhythms (Hyafil et al., 2015; Lakatos et al., 2005). The existence of hierarchical nested oscillations in speech is thought to be a consequence of the brain employing discrete windows of perception

for the successful tracking of different speech components, from auditory to semantic ones (Giraud & Poeppel, 2012).

In the present thesis, I explored the entrainment of neural oscillations to speech, predominantly to the syllabic rhythm. The experiments described here involved analyses methods such as the power or phase coherence calculated using time-frequency analyses based on the spectrogram or the Fourier transform. Using these, I showed the effects of different phonemes placed at the onsets of syllables on neural speech tracking.

A1.2. Examples of phonemic features

This section summarizes some of the consonantal features proposed by Stevens (2002). Manner of articulation and voicing are covered in the main chapters of the thesis. Stevens (2002) also proposes that consonant segments can be divided into [- continuant], if the oral cavity is completely closed during production, or [+ continuant], if the closure if partial. He then divides [- continuant] segments into [- sonorant], if there is an increase in intraoral pressure during closure, or [+ sonorant], if there is no increase. On the other hand, [+ continuant] segments are divided into [+ sonorant] and [- sonorant]. The first of these applies to fricatives whose spectrum amplitude at higher frequencies is higher than that of the neighbouring vowel, whereas a [- sonorant] consonant contains weak high frequencies. Most of the consonants provided in Table A1.2.1 were used in the construction of the stimuli for Experiments 1 and 2.

Table A1. Examples of consonants grouped into different phonemic features (manner of articulation, voicing, articulator-free features).

	/b/	/p/	/m/	/l/	/f/	/s/	/z/
Manner of articulation	plosive	plosive	nasal stop	liquid	fricative	sibilant fricative	sibilant fricative
Voicing	voiced	unvoiced	voiced	voiced	unvoiced	unvoiced	voiced
Articulator-free features	- continuant - sonorant	- continuant - sonorant	- continuant + sonorant	+ continuant - strident	+ continuant - strident	+ continuant + strident	+ continuant + strident
Similar consonants	/d, g/	/k, t/	/n/	/r/	/v/	/ʃ/	/ʒ/

Appendix 2

The sound files for the Experiment 1 can be found on the Open Science Framework website at https://osf.io/v78dm/. Below are the written sentences used as stimuli in this experiment, for both English and Russian. For the English condition, we specify how which "weak", "strong" and "filler" sentences were matched in terms of stress and number of syllables. The latter is also provided between brackets.

English

1. **Weak:** Sa-ra re-viewed se-ven Phy-sics les-sons. (10)
   Strong: Bob-by com-pared bar-gain da-ta pack-ets. (10)
   Filler: Han-nah re-turned ma-ny wol-len jump-ers. (10)

2. **Weak:** La-ry fal-si-fied lea-ses for fe-lons. (10)
   Strong: Peg-gy dic-ta-ted pa-pers to ty-pists. (10)
   Filler: Bar-ry im-por-ted chi-cken from trades-men. (10)

3. **Weak:** Vin-cent re-lieved Li-ly's se-vere fears. (10)
   Strong: Dun-can be-came Pat-ty's con-tent pal. (10)
   Filler: Car-la fos-tered Kei-ra's ma-rooned child. (10)

4. **Weak:** Rus-sell found love-ly sil-ver fai-ry lights. (10)
   Strong: Tuc-ker bought tac-ky gol-den coa-ting pens. (10)
   Filler: Jer-ry caught two mas-sive spot-ted ri-ver fish. (10)

5. **Weak:** San-ford re-ceived five la-zy sul-len fe-lines.(11)
   Strong: Can-dace dis-guised big dow-dy gar-den pat-terns. (11)
   Filler: Cas-sie re-claimed long stud-ded cop-per ear-rings. (11)

6. **Weak:** Lau-rence re-fused Ve-ra's sin-cere feel-ings (10).
   Strong: Pe-ter dis-dained Ted-dy's bol-der tac-tics (10).
   Filler: Da-niel car-ried Jas-mine's hea-vy suit-case. (10)

7. **Weak:** Fear-some sav-age liz-ards sur-round Phi-lip. (10)
   Strong: Po-tent bor-der keep-ers de-tain Bec-ky. (10)
   Filler: Lone-some whis-key drin-kers con-fuse Mad-dy. (10)

8. **Weak:** Lu-cy re-leased va-lid sur-vey re-sults. (10).
   Strong: Ca-dy be-gan daun-ting par-ty de-bates (10).
   Filler: To-ny ex-plained cor-rect dril-ling me-thods. (10)

9. **Weak:** So-phie re-vealed sev-er-al fal-la-cies. (10)
   Strong: Bet-ty de-bunked com-pe-tent dic-ta-tors. (10)
   Filler: Fan-ny re-moved il-le-gal sub-stan-ces. (10)

10. **Weak:** Li-lah re-vived sick fe-ral lo-ri-ses. (10)
    Strong: Pi-per de-bugged poor da-ted com-pu-ters. (10)
    Filler: Ri-ta ap-proved fair mo-dern ser-vi-ces. (10)

11. **Weak:** Fi-fi lost se-ve-ral vel-vet la-ces. (10)
    Strong: Ted-dy cooked dis-gus-ting tur-key bur-gers. (10)
    Filler: Moi-ra solved cum-ber-some jig-saw puz-zles. (10)

12. **Weak:** Ro-ry liked Su-zie's fa-vou-rite so-fa. (10)
    Strong: Co-dy took Pop-py's de-ca-dent ba-gel. (10)
    Filler: Fred-dy met Ri-chard's an-no-ying pa-rents. (10)

13. **Weak:** Li-za va-lued Ro-sie's self-less re-search. (10)
    Strong: Kit-ty tas-ted Deb-bie's bit-ter cab-bage. (10)
    Filler: So-ny ac-cessed Co-ra's hid-den re-cords. (10)

14. **Weak:** Ru-fus sold Ram-sey's large sai-ling ves-sel. (10)
    Strong: Co-by built Can-dy's tall bot-tom cup-board. (10)
    Filler: Da-ni cleared An-drew's full di-ning ta-ble. (10)

15. **Weak:** Le-land saw few li-ver fai-lure sur-vi-vors (11)
    Strong: To-by passed ten bit-ter coun-ty de-tec-tives. (11)
    Filler: Da-vid read most wo-men's fa-shion ma-ga-zines. (11)

16. **Weak:** Sam ra-vaged Syl-vie's fa-la-fel sa-lad. (10)
    Strong: Tom by-passed Dar-by's di-dac-tic gui-dance. (10)
    Filler: Jack men-tioned Car-rie's de-ci-sive ac-tion. (10)

17. **Weak:** La-cey re-ferred li-censed ci-vil ser-vants. (10)
    Strong: Pad-dy com-posed gau-dy de-tailed pain-tings. (10)
    Filler: Ti-na un-locked se-cret ac-count pass-words. (10)

18. **Weak:** Ray searched for Su-san's la-vish fo-rest vil-la. (11)
    Strong: Dean came to Ga-by's dain-ty tim-ber cot-tage. (11)
    Filler: Val cared for Stel-la's fus-sy lit-tle bro-ther. (11)

19. **Weak:** Za-ra saved Lo-ri's fear-ful lone-ly fer-ret. (11)
    Strong: Pip-pa kept Tab-bi's poin-ty tur-quoise pen-dant. (11)
    Filler: Til-da felt Chris-ta's last-ing pain-ful an-guish. (11)

20. **Weak:** Ral-phie fought Va-le-rie's sel-fish land-lord. (11)
    Strong: Ber-tie got Da-ko-ta's bul-ky text-book. (11)
    Filler: Ti-na stole Ca-ro-line's fra-grant hand-cream. (11)

Russian

1. К**а**тя потр**а**тит коп**е**йки к об**е**ду.
   Л**ю**ся зас**о**лит лос**о**ся в расс**о**ле.
   Н**а**дя устр**о**ит пров**е**рку в суббо**о**ту.

2. П**е**тя подт**о**пит п**е**чку для дет**е**й.
   В**а**ся разв**е**сит сп**и**сок за стол**о**м.
   Н**а**стя ув**и**дит сн**и**мки на стене.

3. Б**а**ба к**а**тит баг**а**ж по троп**и**нке.
   Лёва л**о**вит вор**о**н за зал**и**вом.
   П**е**тя в**ы**нул лист**ы** из карм**а**на.

4. Богд**а**н подобр**а**л дом к **а**вгусту.
   Рав**и**ль завар**и**л вар з**а**светло.
   Тит**о**в настилал пол дв**а**жды.

5. Трист**а**н копт**и**л треск**у** тр**и**жды.
   Фил**и**пп свар**и**л фас**о**ль с р**и**сом.
   Максим привез прибор с дачи.

6. Д**е**ти бег**у**т по крут**о**й троп**е**.
   В**о**ва зал**е**з за **ржавый** ларёк.
   Т**а**ня беж**и**т за мл**а**дшей сестр**о**й.

7. Вав**и**лов р**е**зво связ**а**л Л**и**зе в**а**зу.
   Пот**а**пов б**о**йко плетёт д**е**ду к**е**пку.
   Кор**о**вьев т**и**хо чит**а**л Р**и**те ск**а**зку.

8. Ф**и**ля ср**а**зу завёл св**е**тлый фрак.
   Д**я**дя б**о**йко прод**а**л г**и**бкий брус.
   Д**а**ня б**ы**стро над**е**л лёгкий шарф.

9. Бог**а**тый банк**и**р прокат**и**л дет**е**й по п**а**рку.
   Весёлый стил**и**ст рисов**а**л фас**о**н со см**ы**слом.
   При**я**тный шофёр пригласил люд**е**й в авт**о**бус.

10. Под д**у**бом б**о**йко б**е**гал П**е**тя Поп**о**в.
    За в**я**зом р**е**зво л**а**зил В**о**ва Фрол**о**в.
    По д**о**му д**о**лго б**е**гал С**а**ша Мил**о**в.

11. Ревиз**о**р разрез**а**л ст**а**рый рул**о**н.
    Бригад**и**р подкоп**а**л твёрдый кирп**и**ч.
    Самолёт соверш**и**л тр**у**дный манёвр.

12.П**е**тя гр**о**мко буд**и**л п**а**пу к об**е**ду.
В**а**ся ср**а**зу свал**и**л в**и**лы за ст**у**лом.
П**а**па б**ы**стро подн**я**л ребя**т** на пл**е**чи..

13.Б**ы**ков гр**у**бо прибил гв**о**зди к д**о**скам.
Ф**е**ликс р**аз**ом вонз**и**л вёсла в з**е**млю.
Боцман ловко зашил деньги в пояс.

14.Пет**у**х, д**и**ко кряхт**я,** пробег**а**л по тропе.
Вол**ы**, ср**а**зу с утр**а**, залегл**и** за сел**ом**.
Свинья, громко визжа, вбежала в загон.

15.Т**о**карь под б**а**лкой подкрут**и**л г**а**йки.
Сл**е**сарь со сп**о**ром рассверл**и**л ф**а**ры.
Ф**е**льдшер со зл**о**стью поброс**а**л м**а**рлю.

16.Бр**а**гин тайк**о**м подкуп**и**л г**и**да.
Вл**а**сов с утр**а** завоз**и**л Л**и**зу.
С**о**мов в об**е**д заскочил к др**у**гу.

17.Капит**а**н б**о**йко подкат**и**л кат**а**лку.
Фавор**и**т ср**а**зу разозл**и**л Лар**и**су.
Трактор**и**ст д**о**лго провер**я**л мот**о**ры

18.Д**и**кий б**е**ркут покид**а**л гнезд**о** т**о**лько днём.
Ст**а**рый ст**о**рож завяз**а**л ласс**о** в**о**семь р**а**з.
Бедный п**а**рень провож**а**л Л**е**ну к**а**ждый день.

19.Д**я**дя д**о**лго кат**а**л П**е**тю по п**а**рку.
Лёва ср**а**зу вел**е**л Л**ю**се раз**у**ться.
М**и**ша сп**е**шно отд**а**л С**о**не свой св**и**тер.

20.Б**е**лки доб**у**дут прок**о**рм без труд**а.**
Л**и**сы застр**я**ли в силк**а**х за сел**о**м.
К**о**зы слом**а**ли заб**о**р у вор**о**т.

Appendix 3

A3.1. The sound files of the stimuli can be found on Open Science Framework at https://osf.io/3c6tv/. Below are the three streams used for the /b/ condition, where “ba” (/ba/), “be” (/be/), “bee” (/bi/), “bo” (/bo/) and “boo” (/bu/) are the syllables which were repeated and pseudo-randomised for each stream.

Stream 1: “ba boo be ba bee bo ba bo ba bo be boo be bo ba be ba be ba bee”
Stream 2: “be ba bee boo be boo be boo bee be bee bo ba bee boo bee bo boo be boo”
Stream 3: “bee bo boo bo ba bee boo bee be boo ba bee bo boo be bo boo bee bo ba”

For each of the other 14 conditions (vowel-only, or where syllables started with /d, g, k, p, t, s, z, m, n, l, r, f, or v/), the order of the vowels was the same as in each of the /b/ streams.

An example of a filler stimulus is: “nee no noo no na nee noo nee ne **foo** na nee no noo ne no noo nee no na”, where “foo” (in bold) is the different syllable which participants were asked to identify.

A3.2. The code used in for pre-processing and time-frequency analyses of the EEG data can be found at https://github.com/phonemes-and-speech-entrainment/phoneme_isochronous.

A3.3. One-way repeated measures ANOVAs were conducted on the second PCA component of the ITC as well as the 4 Hz evoked power. For this, we calculated the means of each measure, first for sibilants, fricatives, nasals, liquids and stops and secondly, for sibilants/fricatives, nasals/liquids and stops. In the 4 Hz evoked power, both the five groups and the three groups ANOVAs were significant (five group: $F_{4,96}$ = 4.62 p<.01, three group: $F_{2,48}$ = 4.91, p<.05). Subsequent post-hoc T-tests revealed that only sibilants had significantly lower 4 Hz evoked power than stops (p<.05), and this was also the case when sibilants were averaged together with fricatives (p<.05). In the Compound ITC2, only the one-way repeated measures ANOVA conducted for five consonant groups was significant ($F_{4,96}$ = 2.56, p<05). Post-hoc T-tests revealed that the Compound ITC2 was greater for nasals than for liquids (p<.05).

Appendix 4

A4.A.1. The written example of the "da" control stimulus is: "da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da da", where the 250 ms "da" syllable is repeated 40 times.

An example of a filler stimulus is: da da da da da da da da da da da da da da da da da da da da da da da da da da da da da **see** da da da da da da da da da",

where the 250 ms syllable "see" is the 30th syllable in the stream, and is what participants were asked to identify as different, if this was perceived.

The stimulus files can be found online on the Open Science Framework website, following the public link https://osf.io/8hc5p/.

The custom scripts used in the pre-processing and time-frequency analyses of the data can be found at https://github.com/phonemes-and-speech-entrainment/landmarks_eeg.

A4.A.2.

In the 4 Hz ITC, a noise x consonant x location two-way, repeated measured ANOVA revealed a main effect of location ($F_{2,30}$= 5.103, p<.05). The effect of location was maintained for “da” experimental groups ($F_{1,15}$ = 5.91, p<.05), with posthocs showing that onset-altered stimuli triggered higher 4 Hz ITC than stimuli with modified amplitude peaks (p<.001, Bonferroni). When comparing individual groups, however, ITC at ‘Da click CV’ was significantly higher than at ‘Da Click Maximum amplitude (p<.05, Bonferroni). An effect of location was also obtained for ‘click’ but not ‘white noise’ groups ($F_{1,15}$ = 4.31, p<.05), showing a similar trend of stimuli with altered onsets eliciting higher 4 Hz ITC than the ones with altered peaks (p<.05, Bonferroni). Lastly, when investigating noise and consonant effects at separate altered locations, we found a marginal effect of consonant ($F_{1,15}$ = 6.76, p<.05) and a significant effect of noise type ($F_{1,15}$ = 4.49, p = .051) at CV locations only. Paired t-tests revealved that there was no statistical difference between control “da” and “ta” groups. The means of the 4 Hz ITC are given in Figure A4.1.

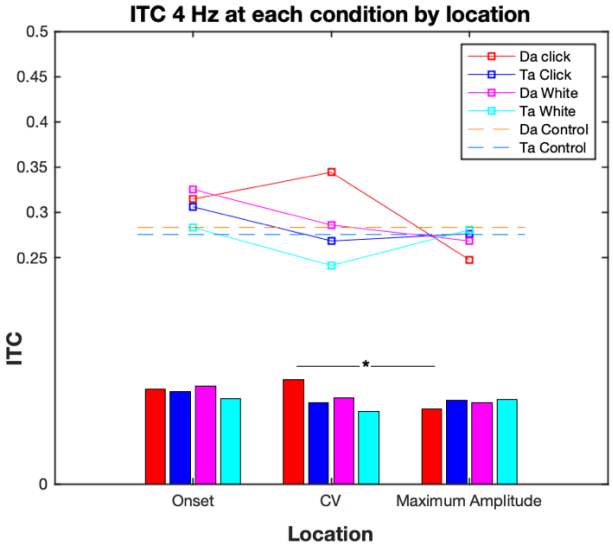
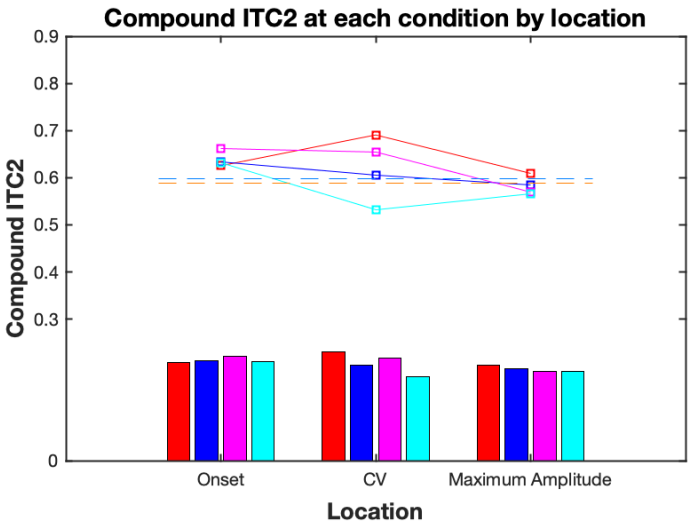


Figure A4.1. The bars represent the values of ITC 4 Hz at each syllable-altered condition and are delimited depending on the “Location” factor. Lines above the bars represent each condition, with solid lines for altered syllables, and also dotted lines for controls. The colour codes for each condition are provided in the legend. The asterisk represents significance level (<.05) between the two conditions.

We also conducted statistical analyses on the second and third components of the PCA run for the ITC at 4, 8, 12 and 16 Hz. We named these Compound ITC2 and Compound ITC3, which explained 22.79% and 9.48% of the PCA variance, respectively. These revealed an effect of noise in the noise type x consonant x location ANOVA, where the means of the 'Click' conditions were higher than the means of the 'White noise' conditions (Compound ITC2: $F_{1,15}$ = 9.04, p<.01; Compound ITC3: $F_{1,15}$ = 5.39, p<.05). Subsequent repeated measures ANOVAs established that this effect was seen for each consonant, but only when the noise was placed at the CV (Compound ITC2: $F_{1,15}$ = 6.89, p<.05, Compound ITC3: $F_{1,15}$ = 9.36, p<.01) and maximum amplitude locations (Compound ITC2: $F_{1,15}$ = 6.67, p<.05, Compound ITC3: $F_{1,15}$ = 7.36, p<.05). In the Compound ITC2, the mean of 'Ta white noise CV' was also significantly smaller than that of "Ta control", as revealed by a paired two-tailed T-test (t(15) = -2.82, p<.05). The means of the two PCA components of the ITC are given in Figure A4.2. No other comparisons were significant.



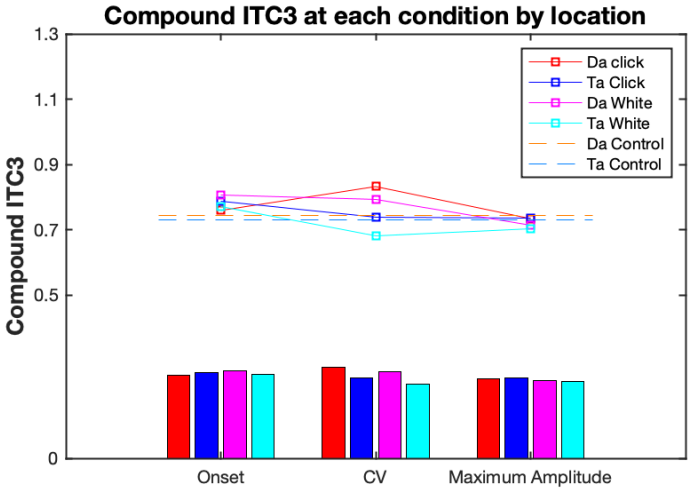


Figure A4.2. The bars represent the values of Compound ITC12 and Compound ITC2 at each syllable-altered condition and are delimited depending on the "Location" factor. Lines above the bars represent each condition, with solid lines for altered syllables, and also dotted lines for controls. The colour codes for each condition are provided in the legend in B.

A. Compound ITC1.

B. Evoked Power.

A4. B. The Python scripts used to run the behavioural experiment and the Matlab code for the binomial probability analyses can be found at https://github.com/phonemes-and-speech-entrainment/landmarks_behavioural.