# MODULATORY EFFECTS OF LINGUISTIC ASPECTS ON CORTICAL TRACKING OF SPEECH

**SHWETA SONI**
**Master of Cognitive & Neuroscience, University of Rajasthan, 2012**

A thesis submitted
in partial fulfilment of the requirements for the degree of

**DOCTOR OF PHILOSOPHY**

in

**COGNITIVE NEUROSCIENCE**

Department of Neuroscience
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

© Shweta Soni, 2021

MODULATORY EFFECTS OF LINGUISTIC ASPECTS ON CORTICAL TRACKING OF SPEECH

SHWETA SONI

Date of Defence: November 29, 2021

| | | |
|---|---|---|
| Dr. M. Tata<br>Thesis Supervisor | Professor | Ph.D. |
| Dr. A. Luczak<br>Thesis Examination Committee Member | Professor | Ph.D. |
| Dr. L. Fangfang<br>Thesis Examination Committee Member | Associate Professor | Ph.D. |
| Dr. C. Ekstrand<br>Internal External Examiner<br>Department of Neuroscience<br>Faculty of Arts and Science | Assistant Professor | Ph.D. |
| Dr. T. Herdman<br>External Examiner<br>University of British Columbia<br>Vancouver, BC | Associate Professor | Ph.D. |
| Dr. S. Pellis<br>Chair, Thesis Examination Committee | Professor | Ph.D. |

# Dedication

To the Universe.

# Abstract

Comprehending speech is a very challenging problem that the human brain solves. Phase alignment between low-frequency cortical oscillations and amplitude modulations in speech (known as 'speech tracking') can resolve certain neurocomputational mechanisms of speech perception, e.g., syllable extraction and phonemic processing. Therefore, speech tracking may be a bottom-up, stimulus-driven mechanism that reflects the processing of speech acoustics. However, efficient speech perception requires the integration of both sensory information embedded in the speech stimulus and top-down influences such as attention and complementary visual information. Yet, the contribution of linguistic aspects in speech tracking responses is poorly investigated. We explored this by comparing speech tracking responses, measured by electroencephalography, from listeners having differential prior experience with the English language. The results suggest that speech tracking responses are not only resulted from bottom-up acoustical processing of speech input but are also modulated by top-down mechanisms learned through deep familiarity with a language.

# Acknowledgments

First, I am very thankful to the LIFE that has bestowed a lot of gratifying experiences upon me. In my countless moments of self-doubt and zero motivation, it always gave me something to hold and go on.

My family is my biggest support system. I thank my parents for guiding and loving me. Thank you my amazing husband Rajarshi Mukherjee, not just for loving me but for bringing in a lot of positive energy, joy and excitement in the monotonous and tiresome phases of my life, and most importantly, for inspiring me to become a better person. This work would not have been possible without your rock-solid support and patience.

I would like to acknowledge my indebtedness and express warmest thanks to my supervisor, Dr. Matthew Tata, for his immense support to make this possible. His friendly guidance and expert advice have been invaluable throughout all stages of the work. He held my back with a lot of patience and believed in me, even when I could not! I would also want to thank Dr. Dillon Hambrook, a friend and a colleague, who helped me to understand electroencephalography (EEG) and its complexities. With both Matt and Dillon, I have had extended and constructive discussions about scientific research and life in general. The time we spent together during lab meetings and executing our projects have contributed greatly to improve this thesis. This work has also been benefited from comments and valuable suggestions made by my committee members, Dr. Artur Luczak and Dr. Li Fangfang. I would like to extend my gratitude towards them.

I am extremely grateful to all those students who participated in my projects; the school of graduate studies (SGS), the Graduate student association (GSA), Amanda Mauthe-Kaddoura, Margaret McKeen and Naomi Cramer for navigating me through all administrative is-

sues. A big thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) and NSERC Collaborative Research and Training Experience (CREATE) program [Biological Information Processing: From Genome to Systems Level program] for supporting my finances.

I am also thankful to Dr. Edmund Lalor for providing me with an opportunity to spend my CREATE travel period in his lab at University of Rochester Medical Centre (URMC), and all the other members of his lab for their hospitality.

Kartik, special thanks to you. We have had our moments of gossips, arguments, laughs, indulgence, discussions that not only kept the heaviness of PhD life at bay, but also propelled me to explore another dimensions of life altogether.

In the end, many thanks to all those who directly or indirectly helped me to finish this 'once seemed impossible to finish' milestone of my life.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The use of language is a defining characteristic of our species, yet the psychological, biological, and computational processes of speech are not well-understood and are sometimes quite counter-intuitive. For example, deciphering spoken sentences is not as easy as reading words written on a page. One can easily identify words in the written text because they are separated by clear white spaces. By contrast, when listening to speech, the stimulus that reaches the ear is a stream of vibrations produced by coordinated movement of vocal cords, jaw, lips, tongue, and the respiratory system. Thus, a speech signal is fundamentally a pattern of acoustic energy bursts that changes with time at various frequencies. However, the consequent acoustic dynamics of the raw signal often do not directly and unambiguously align with the linguistic structure.

The information of speech is encoded in its temporal and spectral components. Each word or each phoneme for that matter, uttered by a speaker have a particular spectrotemporal profile which can also be varied across instances. Yet, despite spectrotemporal variability across speakers and utterances, our brain not only recognizes those sounds but also transforms them, seemingly with ease, into discrete meaningful units, such as phonemes, syllables, words and so on. Most importantly, the brain 'picks up' those units from a complex acoustic signal that often does not have clear acoustic demarcations corresponding to those units. Speech is perceived as a gestalt whole - with grammatical and syntactic structure that is not present in the raw acoustic input. Importantly, the apparent ease with which we parse and understand speech in our own language does not extend to unfamiliar lan-

guages. We do not feel equally comfortable recognizing the proper words in the acoustics of foreign speech. Instead, speech in an unfamiliar language mostly sounds like a train of intermixing sounds without grammatical or syntactic boundaries. Evidently, the brain encodes and represents familiar and unfamiliar speech in fundamentally different ways.

This thesis explores the underlying cortical mechanisms that support the remarkable computational act of making sense from the acoustic signal of speech; a process which is perhaps the most fundamentally challenging aspect of speech perception. Specifically, we test the theory that cortical tracking of speech temporal structure is the consequence of both bottom-up acoustic processing and top-down linguistic processing. The thesis primarily focuses on whether these cortical mechanisms are modulated by the prior experience of a language. Here, 'prior experience' of the language mainly refers to a listener's familiarity with the spoken language, which is achieved through repeated exposure over a long period of time. Prior experience entails the familiarity not only with vocabulary and grammatical structures, but also with acoustic-phonetic features such as co-articulations and prosody. This thesis is structured in the following way: Chapter-1 defines the concept of cortical tracking of the temporal fluctuations in speech, reviews the main related hypotheses and the theoretical and methodological approaches and existing evidence in support or contrary to them, and finally enumerates the objectives of this thesis. In Chapter-2, we investigate the influences of linguistic factors, acquired through prior experience with a language, on the brain's electrical dynamics while segmenting word boundaries in speech sentences. We approach this goal by comparing brain responses from native and non-native speakers. We also explore the contributions of linguistic aspects on the alignment between cortical oscillations and temporal fluctuations in speech acoustics. In Chapter-3, we examine how prior experience with the language affects the cortical tracking of speech features other than the speech envelope. In addition, we describe a longitudinal experiment that aimed to measure longitudinal changes in speech tracking responses as the function of time in non-native listeners. Finally, Chapter-4 summarizes the results and discusses them in the

context of the thesis. It concludes by providing suggestions for potential future research.

## 1.1 Speech and cortical oscillations

Speech is quasi-rhythmic in nature and acoustic events are transmitted in systematic patterns. For example, the duration of phonemes, which are the ”building block” sound components of speech, is on the scale of a few milliseconds to a few tens of milliseconds (mostly between 20-50 ms). Phonemes are nested within syllables which occur with a time-scale of around 125 to 300 ms. Along with syllabic-rate (3-8 Hz) and faster phoneme-rate modulations (>20 Hz), slower modulations (<2 Hz) are also present in speech that corresponds to information about the phrasal structure (duration 500-2000 ms)in speech [55][54][5]. The temporal structure of speech is therefore not strictly periodic, but rather quasi-periodic with variabilities in duration and energy patterns within and across acoustic cues. Moreover, syllables and the bundles of phonemes that comprise them are encoded within amplitude fluctuations in broadband energy (also called the temporal envelope) that mostly occur at rates below 20 Hz with a peak around 5 Hz [66][40]. Interestingly, acoustic fluctuations occurring in the range of 3-8 Hz coincide with the rate of syllables across languages and seems to be a key temporal variable in communication as articulatory gestures, generated by mouth and lip movements in humans while speaking [20][41] and facial expressions such as lip smacking in monkeys [51] exhibit a syllabic-like rhythm. Therefore, slow amplitude modulations at the syllabic rate (between 3-8 Hz) in speech are considered as an important aspect of acoustic speech signal.

Like speech, neural oscillations are also rhythmic. They correspond to periodic fluctuations in the excitability of synchronized neural populations at different timescales that can be measured using various electrophysiological techniques such as electroencephalography (EEG), magnetoencephalography (MEG), and intracranial electroencephalography. Due to the periodic nature, neural oscillations have been proposed to play crucial roles in the encoding of sensory events in our surroundings [42][46][123]. For example, phase

distributions of oscillations in the alpha range (6-12 Hz), shortly prior to presentation of target flashes of light, have been found to predict whether a participant will detect the target or not [18]. This suggests that the processing efficiency of stimulus events that show temporal regularities can be modulated by the phase of ongoing neural oscillations. In this view, information encoded in stimulus dynamics that aligns with a high-excitability phase of oscillations will be processed with more efficiency than information arriving at a low-excitability phase of oscillations [123][46]. Therefore, this phase-alignment of cortical oscillations (low-frequency mainly) with rhythms of incoming stimuli has been seen as a neural mechanism to optimize perceptual sensitivity to relevant information.

***Then, do neural oscillations similarly analyze spectrotemporally complex sounds such as speech? Are rhythmic fluctuations in the brain related to the temporal fluctuations in speech acoustics for speech perception and comprehension?***

A growing number of studies in the past two decades have argued that brain oscillations can track low-frequency amplitude modulations in the speech signal. In other words, the phase of neural oscillations is adjusted to align with the dominant phase of quasi-periodic fluctuations in the envelope of speech. Cortical circuits are thus believed to adjust and adapt to the frequency and timescale of external rhythms. Throughout this thesis, the term 'speech tracking' will refer to this systematic association between the phase of brain oscillations and the phase of the speech envelope. An early account to support this phenomenon comes from a MEG study by Ahissar et al. (2001)[1], who presented listeners with speech sentences with similar temporal envelopes and time compressed them to 20%, 35% 50% and 75% of their original duration to change both rhythms and intelligibility of speech. The results showed that the peak in MEG global field power spectra matched with the dominant modulation frequency of the speech envelope ($\sim$5 Hz) only for intelligible (less time-compressed) sentences, suggesting the close link between low-frequency amplitude modulations and neural oscillations during speech comprehension. Subsequently, Luo & Poeppel(2007) [90] provided additional significant evidence that brain oscillations follow

the temporal regularities of speech. The authors utilized phase and power information of MEG responses to discriminate individual sentences based on acoustic differences among them and found that the phase of theta-band oscillations (4-8 Hz), not power spectrum, reliably distinguished between intelligible sentences. Furthermore, the tracking of acoustic information present in speech inputs through auditory cortical activity was also demonstrated in other studies using EEG ([2]) and electrocorticography (ECoG) [102]. In addition, phase tracking of fluctuations in the speech envelope through low-frequency cortical oscillations has been noted to modulate the power of high-frequency gamma band oscillations in auditory cortical areas [100][109], indicating that cross-frequency coupling may function as an underlying process for communicating among neuronal ensembles to optimize the cortical response to external inputs [46][56].

## 1.2 Role of speech tracking in decoding speech at sub-lexical levels

It is remarkable that cortical oscillations at distinct frequency bands match with the average duration of various speech constituents. For example, high frequency gamma- (>50 Hz) and beta-band (15-30 Hz) activity has been related to rapidly varying phonetic features. Similarly, low-frequency theta-band (4-8 Hz) oscillations have been linked with syllables and mono-syllabic words, and oscillations in the delta-band (<4 Hz) have been associated with the processing of phrase-level speech [56]. Considering the close association between cortical oscillations and fundamental units of speech, it is reasonable to think that speech is analyzed at other timescales (slower and faster rate) along with syllabic timescale. Inspired by this idea, it has been suggested that cortical tracking of the speech envelope drives speech decoding by discretizing the speech signal into linguistic units (syllables and phonemes) in a hierarchical 'window' structure and that is reflected as changes in the phase of neural oscillations at corresponding frequencies [53][56]. This proposed computational principle was formally explained in the TEMPO model, given by Ghitza (2011) [52]. According to this model, sensory information of speech is processed, concurrently, by a parsing path

and a decoding path. Particularly, the brain builds up a nesting relation within temporal windows such that theta-band oscillatory activity tracking the speech envelope (syllabic rate) shapes the beta-band and gamma-band activity (at frequencies roughly 4x and 10x the theta frequency respectively), allowing it to optimally process the dyads of phonemes and spectrotemporal characteristics that differentiate among phonemes. In this arrangement of oscillators, the theta oscillator parses syllables, and thus is considered as the 'master' due to its capability of tracking robust presence of energy fluctuations in speech acoustics. Furthermore, these theta-, beta- and gamma-cycle long segments of sensory input are decoded by mapping them into internally stored linguistic patterns [53][56][54]. Based on this explanation, it can be inferred that the alignment of cortical 'oscillators' to the quasi-rhythm of an input speech signal (conveyed primarily by the acoustic envelope of speech) is the key to speech segmentation. This proposed role of envelope tracking successfully explains counterintuitive behavioral results by [55] that noticed the restoration of speech intelligibility when the inserted gaps of silence in compressed speech sentences fell under the range of syllables ($\sim$120 ms) and phonemes ($\sim$20 ms). Similarly, studies have also demonstrated that when gaps of silence were filled with smooth noise such that the low-frequency acoustic envelope was restored, the perception of speech was also restored[60].

Interestingly, the tracking of slow modulations in the sound envelope was not restricted to speech and found also for non-speech sounds [127]. Since this tracking phenomenon seems not to be specific to speech but rather related to any acoustic signal with suitable dynamics, speech tracking has mostly been believed to represent bottom-up sensory processing of speech acoustics. However, studies have found interesting attention-related effects on speech tracking responses [74][82][57][35][36][61][62], an enhancement of cortical tracking to speech if it is complemented by visual information [134], and changes in envelope-tracking responses in responses to changes in speech intelligibility [110]. These observations lead to a point that speech tracking responses in listeners do not only respond to acoustic information in speech signal. In the following sections, we will discuss whether

cortical tracking responses to speech also represent other information in the speech signal, especially linguistic information.

## 1.3 Is speech tracking acoustic-driven only?

Thus far, we have discussed that the cortical tracking of speech envelope plays an important role in speech processing. However, the processing of speech is not simply limited to recognizing an acoustic stream as 'speech' or identifying certain phonemes or syllables within it. Speech is for conveying meaning through an acoustic input, that is to parse the continuous speech signal into meaningful linguistic units such as words so that a listener can comprehend what the speaker is saying.

***Therefore, a key question is whether speech-entrained brain rhythms contribute directly to speech comprehension. Do speech tracking responses merely represent a perturbation of synchronized networks of oscillating neurons by a sensory input whose rhythmicities correspond to the resonances of those networks? We will refer here to this idea as "stimulus-driven" or "bottom-up". Alternatively, is there a relation between speech tracking responses and higher cognitive processes that drive the perception of an incoming stimulus? In this thesis, we focus on top-down influences due to the learned linguistic mechanisms of understanding speech and will refer to this as a "top-down" or "learned" hypothesis.*** We leave aside questions of other top-down modulations, such as due to selective attention, which have been broadly considered elsewhere, for example [62] [135].

Several studies have attempted to examine that relationship by manipulating the intelligibility of speech and measuring the corresponding changes in speech tracking responses. One of the basic methods that can be used to manipulate the intelligibility of speech is to change the acoustical properties of the speech signal. For example, speech can be made unintelligible when the speech is time-compressed [1], or the speech is reversed in time [67][31], or when fine spectrotemporal details are disrupted [90][38] or when speech is noise-vocoded [110][30]. These studies have demonstrated that speech tracking by low-

frequency oscillations was reduced when speech intelligibility was compromised using such manipulations as compared to those situations when speech intelligibility was not disturbed [59][33][129]. Interestingly, when the overall amplitude envelope, which is the primary acoustic feature of speech rhythm, is kept preserved but the speech is completely unintelligible, speech tracking responses can also be observed. However, these speech tracking responses are significantly enhanced in the case of intelligible speech that contains complex spectral details [110]. These observations imply that low-frequency speech tracking responses are also benefited by non-sensory information available in the intelligible speech, rather than being purely driven by acoustic characteristics of speech during speech comprehension.

Although it seems apparent that the intelligibility of speech can modulate speech tracking brain responses, the relationship between speech intelligibility and the cortical tracking of speech is not so straightforward. Other observations suggested no link between cortical speech-tracked activity and speech comprehension by reporting that the cortical tracking of speech was not affected by deterioration in speech intelligibility. For example, Howard & Poeppel (2010) [67] found comparable theta-band oscillations while discriminating both normal speech and unintelligible time-inverted auditory inputs that had preserved the slow envelope modulations of normal speech. Similarly, Zoefel & VanRullen (2016) [136] used normal speech and noise-mixed versions of original speech and found that speech tracking responses were not different for normal and noise-mixed speech, and even when the noise-mixed speech was played time reversed. These studies have argued that theta-band speech tracking activity is associated with acoustic information present in the stimulus and independent of whether the stimulus is comprehensible. Thus, the speech-tracking phenomenon might reflect brain mechanisms related to the purely acoustic features of speech sounds and/or the linguistic content of speech, and the evidence remained equivocal.

## 1.4 Do linguistic aspects matter to speech tracking?

### 1.4.1 Evidence from PRIMING-based manipulations

One approach to solve the problem of separating acoustic and linguistic contributions in speech tracking activity involves the use of a priming paradigm. In this method, a degraded version of the speech stimulus (unintelligible by itself) is first primed with the non-degraded, original version of that speech (thus, intelligible) so that listeners can perceive degraded, unintelligible sentence as an intelligible sentence [27]. This approach addresses one of the main constraints that frequently occurred with acoustic manipulations of speech inputs. Manipulating speech acoustics might distort the rich spectrotemporal structure of speech along with other structure of the speech signal. Therefore, it becomes difficult to come up with control stimuli for speech intelligibility experiments that could appropriately match the physical properties of manipulated stimuli. However, if *acoustically identical* speech stimuli are presented twice, before and after priming, they will differ only in their intelligibility, because only the second presentation of acoustic stimuli would be comprehensible through prior perceptual learning. Thus, this method would allow one to examine the changes in speech tracking responses due to the differences in speech intelligibility rather than due to differences in the physical properties of speech stimuli. Recent electrophysiological studies using noise-vocoded speech sentences and the above-mentioned priming paradigm have revealed perceptual enhancement of unprimed vocoded speech after the presentation of original version of speech that was correlated with an increase in the cortical tracking of speech-specific features, especially phonemes and phonetic features [30][29][32], indicating the role of higher-level linguistic aspects associated with the prior knowledge of acoustic sentences on low-frequency speech tracking responses.

Despite the sophistication of the priming paradigm, studies investigating the linguistic aspects of speech tracking activity have observed a lack of consistency among results. For example, Millman and colleagues (2015) [98] compared theta-band oscillatory responses, recorded using MEG, to tone-vocoded speech stimuli before- and after- the exposure of

original unprocessed sentences and found no enhancement of speech tracking responses even when the unintelligible vocoded speech was perceived as intelligible by listeners, suggesting that the relationship between cortical responses that track temporal fluctuations in speech and speech perception does not rely upon linguistic information. Similarly, Baltzell et al. (2017) [8] used 3-channel and 16-channel tone-vocoded speech sentences and manipulated their intelligibility using priming paradigm. The authors examined EEG-measured speech tracking responses to vocoded (target) sentences that were primed by non-vocoded natural speech that either matched the target (valid) or did not match the target (invalid). The results of the study demonstrated larger speech tracking in valid trials for both 3-channel and 16-channel tone-vocoded sentences, though these speech tracking responses did not differ between 3-channel and 16-channel vocoded speech. That suggested that any modulatory effect of prior information on speech tracking responses were sensory-driven, not linguistic-driven. The probable reason why these could not find the connection between speech intelligibility and speech tracking activity is the behavioral task they had chosen. In the Millman (2015) study, listeners had to rate the intelligibility of speech sentences by pressing the button after receiving acoustic cues while listeners in Baltzell (2017) study had to identify if the probe snippets were drawn from unintelligible vocoded sentences. Therefore, listeners were explicitly directed either to perceptually map the degraded version of speech acoustics onto either the original non-degraded speech or to the stimulus acoustics, something that would not require linguistic processing. Even if listeners were using linguistic information when comparing the primed sentence (after-exposure) to the prime (original), it might be restricted to the over-learning of clean speech features that could just facilitate perceptual restoration of degraded speech. Possibly, if Millman et al. (2015), Baltzell et al. (2017) and others, had chosen behavioral tasks that explicitly demanded the semantic and/or syntactic processing of speech stimulus rather than just detecting the repetition of syllables, recognizing tone-pip within speech signal, and so on, the contributions of linguistic information would have been more reflected in speech tracking responses.

### 1.4.2 Evidence from Linguistic manipulations

We have, thus far discussed that brain responses to slow temporal fluctuations in speech may or may not be related to the intelligibility of incoming speech input. The issue, therefore, remains to be resolved as there is no clear consensus whether speech tracking responses reflect neural processing of higher-level linguistic content while listening to natural speech. In the past couple of years, other approaches have been shown to cope with the problem more efficiently by directly changing the linguistic content of speech stimuli. The rationale is that the brain responds very differently to speech that it can comprehend, relative to what it cannot comprehend. For example, as we listen to speech in our native language, our brain continuously predicts and generates expectations about upcoming speech information. This is possible only because a deep language model is learned through prior experience. These predictions align to bottom-up processing of sensory information. These top-down predictions are known to affect cortical oscillations to enhance sensory perception of audiovisual stimuli [7][134]. Therefore, it is reasonable to expect different low-frequency speech tracking responses when the presented speech does not align with a learned linguistic model. Studies that used this kind of manipulation for variety of tasks have yielded fascinating results. For example, significant differences in speech tracking responses were observed in the delta-band band EEG when participants listened to sentences constructed with real words, but not pseudowords with valid syllables; thus showing that syntactic structure matters for speech tracking [92].

A few recent EEG/MEG studies have proposed that cross-linguistic investigations could also reveal language-related effects on speech tracking, especially with respect to speech envelope and other abstract linguistic representations. To illustrate this, Ding et al. (2016)[34] compared brain responses following speech rhythms from listeners with different levels of speech comprehension ability and found that cortical oscillations, measured by MEG, track hierarchical linguistic structures (words, phrases, and sentences) in artificially constructed connected speech at distinct timescales. Most importantly, the neural tracking of abstract

linguistic structures was found to rely upon whether the testing language was comprehensible. For example, when Chinese sentences were presented to native Chinese listeners or English sentences were presented to native English listeners, MEG responses in both groups followed words and phrases on their respective time scales along with a $\sim$4 Hz syllabic rhythm. However, when native English listeners listened to Chinese sentences, the cortical activity showed encoding of the acoustic (syllabic) rhythm only. A crucial point of this experiment was that Ding et al. (2016) artificially constructed speech sentences based on an isochronous sequence of syllables such that when these syllables were structured into higher-order linguistic constituents, no acoustic gaps were inserted between constituents. Due to the lack of these acoustic cues, the acoustic envelope only represented fluctuations corresponding to the rate syllables (4 Hz), and cortical activity at different timescales resulted from an internal understanding of grammar-based connections across multiple linguistic levels rather than encoding of acoustic cues. Put another way, the only temporal dynamics available in the acoustic stream was the 4 Hz syllable rhythm, yet brain electrical activity tracked slower linguistic structures as well. Furthermore, low-frequency cortical responses are not limited to track phonemes and syllables in speech sentences. Studies have observed cortical tracking at lexical-, phrasal- and sentential- levels [34]. Interestingly, different brain areas seem to contribute to the cortical tracking of distinct linguistic structures for comprehended sentences. For example, left premotor cortex strongly tracks phrases while left middle temporal cortex follows sentences at the scale of words [72]. Moreover, the oscillatory activity of several brain areas underlying linguistic processing has been noted to influence delta-band cortical tracking of auditory speech [73]. That suggests that cortical responses aligning with slow fluctuations in speech can interact with different linguistic mechanisms involved in speech perception. Similar results have been observed in a different yet related MEG study [99] which contrasted speech tracking responses to natural speech compared to amplitude-modulated white noise and spectrally rotated speech. They found coherence between brain oscillatory activity and the speech envelope in both theta

and delta band, but speech tracking activity was higher than other conditions only in the delta band. Based on these results, it can be argued that delta-band speech tracking entails internal modulatory computations subserving language comprehension, whereas theta-band speech tracking primarily reflects auditory processing of a speech signal.

Recently, a few other electrophysiological studies, examining the relation between speech tracking and language-related effects using different linguistic populations or stimuli in different languages, have produced interesting results. For example, Etard & Reichenbach (2019)[44] reconstructed speech envelopes from EEG activity, recorded from English speakers, in response to noise-mixed sentences spoken in English (native) and Dutch (foreign) language using linear backward modelling. The results of that study revealed that cortical tracking, measured as the accuracy of the reconstructed envelope, of Dutch sentences significantly differed from English sentences, supporting the idea that comprehension of speech modulates cortical tracking of speech against noisy backgrounds. Furthermore, Zou and colleagues (2019)[137], using the similar EEG-based envelope reconstruction approach, showed that both native and foreign listeners of Cantonese were able to reconstruct the temporal envelope of Cantonese speech at various signal-to-noise levels, representing the role of bottom-up auditory processing even for degraded, unintelligible speech. However, contrary to previous results [44], the strength of the correlation between reconstructed and original envelope was higher in the listeners who did not understand the language in which sentences were spoken. However, the results overall suggest that language processing affects how precisely the speech envelope is encoded in the brain.

Due to the lack of consensus, it is therefore of interest to further explore whether language-related features contribute to speech tracking in the EEG. The thesis considered here is that cortical tracking in response to natural speech reflects an interplay of acoustic aspects through bottom-up processing of auditory information of speech, and linguistic aspects through top-down linguistic mechanisms that are employed when the speech input is perceived as comprehensible. We further hypothesized that speech tracking activity corre-

sponds to acoustic and linguistic processing but in distinct frequency domains, pointing to their different functional roles during speech perception.

## 1.5 Thesis Objectives

In this thesis, we show that low-frequency cortical tracking of natural speech reflects top-down linguistic aspects acquired through prior knowledge of the language. This tracking is beyond that of the lower-level acoustic representations. The findings of this thesis will assist further researchers in understanding the functionalities of cortical rhythms in speech perception and their reliance upon deep familiarity with the spoken language. The tracking of slow fluctuations in the amplitude envelope of a speech signal through cortical oscillations has been studied extensively and proposed as a neural mechanism to facilitate speech perception. However, top-down mechanisms are also required for efficient speech perception, which are likely to be reflected in speech tracking responses. Though several studies have revealed the effects related to higher-level processes such as attention [134][61][62] and intelligibility [38][34] on speech-related cortical responses, the contributions of non-acoustic information provided by a learned internal language model in auditory cortical tracking are relatively poorly understood. In Chapter-2, we examine cortical tracking responses to the speech envelope from native and non-native listeners of the English language. Previous studies have mainly demonstrated changes in speech tracking responses in paradigms that altered speech intelligibility via acoustic manipulations or priming. Here, we present a novel experimental design that manipulates speech intelligibility in two ways: by priming speech using multiple repetitions of the same acoustic input and by considering listeners' prior experience with the testing language in a between-groups design. We also consider the effects of a prior language model on phase dynamics of cortical responses while processing the speech at the time scales of words and phonemes. In Chapter-3, we investigate the cortical encoding of other speech features in addition to the acoustic envelope and show that prior linguistic knowledge modulates how well they are encoded in brain

electrical dynamics. Second, we examine if acoustic and linguistic aspects are reflected at distinct timescales in speech tracking activity. Third, we report the results of an attempt to gauge longitudinal changes in speech-following cortical activity in non-native listeners of English. We tested the hypothesis that changes in the familiarity with the foreign language due to a prolonged immersion in that foreign language are accompanied by the changes in the cortical tracking of speech in that language.

# Chapter 2

# Brain electrical dynamics in speech segmentation depends upon prior experience with the language

The present chapter explores the hypothesis whether the process of speech tracking, which facilitates speech segmentation, reflects top-down mechanisms related to prior linguistic models or stimulus-driven mechanisms, or possibly both. The specific aims of the present study were as follows: 1) To investigate the neural correlates of speech segmentation, specifically word-segmentation that rely on prior experience of the language. 2) To investigate whether cortical responses following temporal fluctuations in speech envelope are modulated by listener's prior experience with the language. To address these, we recorded high-density EEG responses from native and non-native speakers of English that had different prior experience with the English language while they heard acoustically identical sentences. Despite a significant difference in the ability to segment and perceive speech, our EEG results showed that theta-band tracking of the speech envelope did not depend significantly on prior experience with language. Furthermore, native and non-native speakers showed different phase dynamics at word boundaries, suggesting differences in segmentation mechanisms. Finally, we found that the correlation between higher frequency dynamics reflecting phoneme-level processing and perceptual segmentation of words might depend on prior experience with the spoken language. Our results suggest a possible correlate of word boundary segmentation in the coupling of faster (i.e., 12–18 Hz) EEG dynamics with perceptual segmentation of words.

## 2.1 Introduction

To a native speaker of a language, a spoken sentence sounds like a train of meaningful words. However, a non-native listener might perceive this same sentence as merely a time-varying sequence of meaningless sounds. In fact, it is the non-native listener who has a more true percept of the actual acoustics: the perceptual boundaries between linguistic units such as words and phrases in the speech as understood by the native listener often do not reflect boundaries in the amplitude envelope of that speech. The resulting difficult computational problem of finding where syllables, words, and phrases begin and end in time has been called the segmentation problem [15][21]. Recent theories have proposed that the tracking of low-frequency dynamics of speech by oscillatory or quasi-periodic brain electrical activity reflects the neurocomputational mechanisms by which the brain transforms acoustic events into linguistic events [2][29][30][31][37][52][56][60][72][135]. This is because cortical electrical activity measured by electroencephalography (EEG), magnetoencephalography (MEG), and electrocorticography (ECoG), at distinct rates, aligns with the timescales of various linguistic structures. For example, brain dynamics in the delta ( 1–4 Hz) band has been shown to track the temporal dynamics of words and phrases [34][79][96], whereas, activity in the theta-band ( 4–8 Hz) band has been found to track the syllabic rate [38][53][54][56][61][111].

Despite clear evidence that brain electrical dynamics can track acoustic fluctuations and linguistic features of speech, and a strong theoretical basis to believe that this tracking phenomenon reflects brain mechanisms of parsing linguistics units from the acoustic stream, there is an outstanding question about the relative contributions of top-down mechanisms related to prior linguistic models versus stimulus-driven mechanisms related to features of the acoustic stimulus. It is possible that speech tracking in at least some frequency bands of the EEG reflects purely stimulus-driven mechanisms that do not depend on familiarity with the language. On the other hand, the speech tracking phenomena might reflect predictive mechanisms dependent on prior knowledge of the language. One way that this question has

been addressed by previous research has been to present speech stimuli that do not align with the listener's prior linguistic models, for example by degrading or manipulating the physical properties of speech stimulus. Using this approach, some studies have found that speech tracking by low frequency activity was enhanced when speech was intelligible as compared to conditions in which speech was made unintelligible by compressing the sentences in time [1], by removing spectrotemporal details [33][110], by distorting temporal cues from the envelope [38], by adding noise [90][129] or by disturbing phonological information [92]. On the contrary, studies have also noticed speech tracking even when the low-frequency envelope and spectral features were obscured [136], when speech was time-reversed [67], or when phonemic features were removed but the low-frequency envelope was preserved [60]. Each of these studies revealed important insights about how the brain tracks the acoustic and linguistic dynamics of speech by changing, in various ways, the acoustic signal presented to the listener.

Using a different approach, some previous studies have made significant progress in disentangling aspects of speech tracking responses related to speech acoustics from other aspects related to speech comprehension. This approach uses various paradigms to modulate speech intelligibility while keeping low-level acoustics either masked or unchanged. For example, vocoding speech into a few frequency channels reduces its intelligibility [124]. Subsequently, priming that speech using either hearing or vision with the original version of the same speech restores intelligibility [27]. This perceptual improvement in speech intelligibility due to priming occurs presumably because top-down mechanisms are provided with prior knowledge about the phrase and can generate expectations of speech content. Importantly, such enhancement of perception of vocoded stimuli by priming also enhanced speech evoked responses in EEG and increased source power in inferior frontal gyrus revealed by source reconstruction analysis [125]. Furthermore, this experimental manipulation also revealed better encoding of phonetic and phonemic features in both EEG and MEG recordings [30][29]. More recently, a study [16] using natural continuous speech

stimuli, considered the predictability of words based on their semantic similarity with their sentential context and showed enhanced cortical tracking of the word envelope, especially in the theta-band, associated with better predictability given the context. This suggests that semantic expectation of upcoming speech, provided by contextual cues that link to a prior language model, has some influence on the neural encoding of speech.

In the present study we took a somewhat different approach. Rather than using acoustic manipulations to modulate speech intelligibility, we considered brain responses to physically identical speech stimuli in two different groups: native and non-native speakers of the language. The advantage of this approach over the above-mentioned studies is that manipulating intelligibility by changing the stimulus itself necessarily modulates both low-level acoustics and higher-level linguistic processing [110]. By leaving the stimuli acoustically identical, but considering different prior language experience of the speakers, we were able to reveal aspects of speech tracking that can be due only to higher-level linguistic processing, rather than low-level stimulus acoustics. In particular, although speakers were presented with identical spectrotemporal dynamics, they varied in their ability to parse the boundaries of language-specific structural units, particularly words.

The few prior studies exploring speech tracking to linguistic features have reported the important result that neural response patterns in the theta-band were no different for speech in native as well as in unknown language [112][113]. Using MEG, Ding et al. (2016)[34] showed that theta-band brain responses entrain to the acoustic syllabic rhythm of the speech, regardless of the familiarity with the language [34]. By contrast, delta, but not theta, signals followed the phrasal and sentential structure of continuous speech, but critically only among native speakers of that language [34]. Although this study attempted to differentiate cortical tracking of acoustic and linguistic features in a sophisticated way, the artificially designed stimuli make these results somewhat difficult to generalize to more natural speech. In a recent study [44], Etard and Reichenbach (2019) used continuous natural speech in both English and Dutch, and changed both acoustic features (by varying

degree of background noise) and comprehension (by presenting Dutch stimuli to native English speakers). They similarly found that delta-band tracking was regulated by language comprehension, whereas speech clarity determined the cortical tracking of speech in the theta-band. Taken together, it seems that the brain can parse the 5 Hz syllable structure of speech based purely on the acoustic envelope, and thus can do so with or without a prior language model. However, the mechanisms by which the brain segments linguistic features at other time scales such as slower (i.e., in the delta range) grammatical units or faster (i.e., in the 12–18 Hz range) phonemic units might depend on top-down mechanisms that engage only for speech that the listener fluently understands.

The aim of the present study was, therefore, to investigate the neural correlates of speech segmentation, specifically identifying correlates that rely on prior experience of the language. In a simple listening task, we presented continuous English sentences to two distinct groups of participants that prominently differed in their familiarity with English, and we simultaneously recorded high-density EEG signals from these speakers. We predicted that both native-speakers and non-native English speakers would be able to track speech acoustical dynamics in the theta-band, in line with previous research demonstrating stimulus-driven processing in the theta-band for incomprehensible, time-reversed, or foreign speech [34][67][112]. However, we expected different phase dynamics related to language-specific processing in native English speakers relative to their non-native counterparts. Our results supported our predictions and showed that non-native English speakers do track the syllabic level dynamics in the theta-band. However, higher frequency dynamics reflecting phoneme-level processing depended on prior experience with the spoken language. Furthermore, we report here a possible correlate of word boundary segmentation in the coupling of phoneme-rate (i.e., 12–18 Hz) EEG dynamics with perceptual segmentation of words.

## 2.2 Methods and Materials

### 2.2.1 Participants

A total of 15 native English speakers (Native: 3 males & 1 left-handed) and 13 non-native English speakers (Non-native: 2 males & 2 left-handed) participated in this study. Of the non-native English speakers (mean age in years (SD): 19.5 (1.09), range = 17–21 years), 10 were native Japanese speakers along with one native Korean, one native Ukrainian and one native Spanish speaker. All participants were recruited at the University of Lethbridge. They provided written informed consent at the beginning of the study and received either credits for an academic course or monetary remuneration for their participation.

At the time of testing, non-native English speakers had been taking an "English for academic purposes" course for 3–6 weeks in the international services centre at the university. They had been exposed to English in their educational system at an average age of 10.7 years (SD: 2.21, range: 6–13 years) but reported little fluency in English as they used it infrequently for regular conversation. The linguistic profile of these participants was acquired on the basis of a custom-made language questionnaire in which they reported about their language acquisition history and self-rated on a ten point scale (1=least proficient, 10=most proficient) for speaking, listening, reading and writing skills in each language (Table 3.1). The self-reported ratings revealed higher scores for above-mentioned skills in their native language relative to English. The frequency with which they watched TV shows or movies in English and their native language ranged from rarely to everyday. They also reported having spent no more than 3–6 months in Canada or any other English-speaking country, except one participant who had spent  2 years at the time of the study. All native English speakers (mean age in years (SD): 21.47 (3.02), range = 19–30 years) were Canadians and recruited from an undergraduate course in the university.

In this study, the behavioral data from all 28 participants were analyzed; electrophysiological data from one non-native English-speaking participants (native Ukrainian) was excluded in the final data analysis due to extremely noisy EEG signals. All participants had

Table 2.1: Mean (SD) of non-native participants' linguistics self-rated proficiency on the scale of 1-10 point

| Language skills | In native language | In English language | Significance |
|---|---|---|---|
| Speaking | 9.50 (0.53) | 5.30 (1.06) | 1e-07 |
| Listening | 9.80 (0.42) | 5.00 (1.05) | 1e-07 |
| Reading | 9.10 (0.88) | 6.20 (1.32) | 2e-07 |
| Writing | 8.70 (1.16) | 5.95 (1.67) | 2e-07 |

normal or corrected-to-normal vision and reported no history of any neurological or psychiatric disorders. The study was in accordance with the Declaration of Helsinki and that all procedures were approved by the Human Subjects Ethics Committee of the University of Lethbridge.

### 2.2.2 Stimuli

Fifty continuous speech sentences were taken from TIMIT Acoustic-Phonetic Continuous Speech Corpus [49]. The corpus contains time-aligned orthographic, phonemic, and word transcriptions of over 600 speech samples of different dialects of American English. All chosen speeches sampled at 16 KHz were spoken by male speakers from two dialect regions across the United States. Speech samples varied between 3–4 s in length and normalized to the equal root mean square (RMS) amplitude. For each stimulus, two speech samples were concatenated in a way that they were no shorter than 5.5 s and no longer than 6.5 s in length, resulting in 25 unique speech sentences. Speech stimuli contained an average of 16.24 (SD: 2.28, range: 12–21) words, an average of 24.28 (SD: 3.13, range: 19–31) syllables and an average of 70.20 (SD: 8.72, range: 52–89) phonemes per stimulus and were presented for each participant in random order. We were interested in the relationship between EEG dynamics and the time course of linguistic events in these sentences. To identify frequency bands of interest in the EEG, we evaluated the time-scales of these linguistic units based on the statistical regularities of our speech stimuli [72]. We found the presentation rate of words and phonemes by taking inverse of the median difference

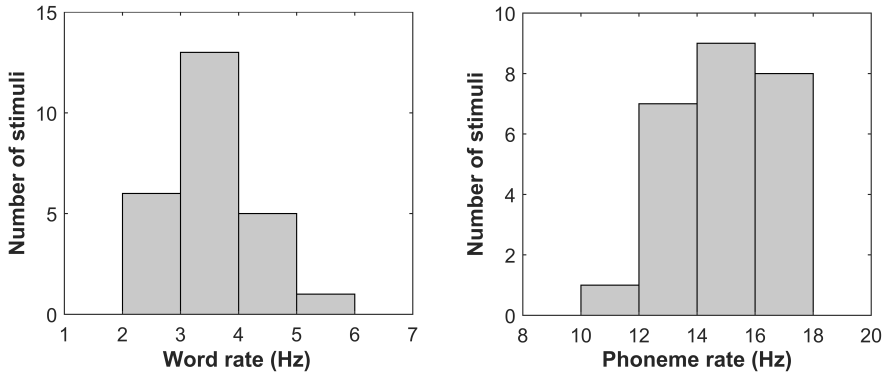


Figure 2.1: **Histograms showing the rate of occurrence of words (left) and phonemes (right) in speech sentences used in the current study**. Words and phonemes occurred at the rate of 2–5 Hz and 12–18 Hz respectively.

between onset times of consecutive words and phonemes in the available text transcripts (Figure 2.1).

Since the TIMIT transcripts do not include syllable boundaries, syllables in the presented sentences were counted manually and the syllabic rate was calculated as the total number of syllables per stimulus divided by the duration of that stimulus. Therefore, we defined three frequency bands of interest for subsequent EEG analysis, specific to each linguistic unit, as the minimum and maximum frequencies across all stimuli: 2–5 Hz corresponding word rate, 3.2–4.9 Hz corresponding syllabic rate and 12–18 Hz corresponding phoneme rate.

### 2.2.3 Experimental Procedure

An Apple Mac Pro with a firewire audio interface (M-Audio Firewire 410) was used to present all stimuli in free-field and send triggers to Electrical Geodesics Inc. Net Station data acquisition software. Participants sat 1 metre from a studio-grade audio monitor (Mackie HR624 MK-2) located on the auditory front midline in a sound attenuated room. It was ensured that EEG recording was free of any speaker or other electromagnetic interference to avoid artifacts. Participants used a keyboard on a table close to them to report behavioral responses. The presentation of stimuli was controlled by customized MATLAB

scripts (The MathWorks Inc., Natick, MA, USA) and Psychophysics Toolbox functions [12], which were running on Apple Computer's Core Audio Framework (Mac OS 10.6).

Participants completed a total of 100 trials as each speech stimulus was presented four times consecutively, with one break after 50 trials. Participants were asked to listen to stimuli while focusing eyes on a '+' sign displaying in the centre of a computer monitor directly in front of them. After each trial, participants were required to type all the words they heard in that trial and press the "Enter" key. The limited time of 30 s was given to complete the trial to control the writing speed variability. The design of the task necessarily involves both speech perception and working memory. By providing an excess of time to respond, and assuming equivalent working memory capacity between native and non-native listeners, we can know that the results will reflect speech perception or possibly an interaction between speech perception and working memory mechanisms. Before starting the experiment, all participants were given proper instructions and practice stimuli to make them familiar with the task.

### 2.2.4 EEG data recording and preprocessing

High-density EEG was recorded throughout the experiment with Ag/AgCl 128 electrode net (Electrical Geodesics Inc., Eugene, OR, USA) fitted on participants' head. The EEG was sampled at 500 Hz sampling rate and electrode impedances were kept below 100 KΩ. The EEG data was pre-processed first with Brain Electrical Source Analysis (BESA; Megis Software 5.3, Grafelfing, Germany). The data was first filtered at 0.5 to 30 Hz pass band to avoid unnecessary low- and high-frequency fluctuations. Channels which looked visibly noisier or flat than surrounding channels, and/or had signal from less than 10 channels were identified as 'bad channels'. Recordings of eye movements and blinks were obtained in the data due to the length of trials and subsequently corrected using ocular artifact correction spatial filters as described in [69] and implemented by the BESA analysis software. Bad channels were then replaced using spline interpolation after all artifact cor-

rections. The data was thoroughly checked if it had trials with more than 10 bad channels to be discarded from the analysis. However, qualitatively, after correction we did not observe any trials with substantial artifacts, so we kept all trials for both participants groups. The data was then re-referenced to an average reference and transferred from BESA to MATLAB (MATLAB version 9.1.0; The Mathworks Inc., 2016, Natick, MA, USA) for all further analysis using EEGLAB functions [28] and customized code. We extracted 8.5 s epochs of EEG data, beginning 700 ms before the onset of each trial after filtering data between 1 to 20 Hz pass band using FIR.

### 2.2.5 Statistical tests

The Statistical analysis was done in SPSS and MATLAB. Because of the between and within-subject design of the task, we mostly performed mixed ANOVA (analysis of the variance) test. The assumptions for homogeneity of variance with Levene's test were checked before interpreting ANOVA's test statistics. If data violated the assumption of sphericity (p-value ¿ 0.05 for Mauchly test), degrees of freedom were corrected using Greenhouse-Geisser method. The one sample/paired t-test was also used to compare with null/surrogate distributions. To correct for multiple comparisons, the Benjamini-Hochberg method or Bonferroni method was used for controlling the type-1 errors.

## 2.3 Data analysis

### 2.3.1 Behavioral Data analysis

The behavioral performance was assessed based on two perceptual measures. First, the ability to segment the speech on each trial was measured as the percentage of reported words, regardless of their correctness, out of total words presented in the speech stimulus. Second, we calculated the percentage of correctly written words by matching response words with actual words in the speech stimulus. We did this while both including and excluding articles (‘a’ ‘an’ and ‘the’) in the speech text. We predicted that non-native

speakers would report lesser words at a given presentation level, when the accuracy of words were not considered and when articles were considered in the speech text or not. Whereas native speakers would perform better considering their better ability to segment and identify words. In general, second-language learners of English tend to face difficulties in acquiring English articles, regardless of the number of years spent in learning English (Andersen, 1984). The learning of the English article system particularly becomes more complex in the absence of an equivalent article system in L2 learners' native language [70][116]. In the case of our non-native participants, Japanese contains nothing equivalent to English articles [3], whereas Korean does not have any definitive article system like English does [85]. Therefore, as an additional measure, we took the difference between correct words with and without articles to see whether non-native speakers could identify and segment English articles when articles were included in the speech text. We predicted that native speakers would show more difference between these two scores (correct words with and without articles) considering their ability to identify and segment articles. Whereas non-native speakers would not show much difference at a given presentation level when articles were considered in the speech text or not.

### 2.3.2 EEG Data Analysis

We restricted all our EEG analysis to a cluster of 14 frontocentral electrodes covering the frontocentral mid line scalp. We chose them based on our observations which is consistent with previous work in our lab [60][61]. The criteria for selecting particular frequency bands and lags for each EEG analysis has been explained in respective data analysis and result sections.

#### 2.3.2.1 Neural tracking of speech acoustics

To measure how the brain responds to the acoustic envelope modulations related to syllables, we employed a cross-correlation technique used previously to assess cortical speech tracking [61][62]. For both groups, we extracted the speech-tracked brain responses as the

cross-correlation between the first derivative of the envelope of each speech stimulus and corresponding time-aligned EEG signal at each channel. The acoustic envelope of each speech stimulus was calculated by computing the absolute value of Hilbert Transform followed by low-pass filtering at 20 Hz using a zero-phase finite impulse response filter and then resampling at 500 Hz to match the sample rate of EEG. The acoustic envelope was then processed by calculation of its first time derivative, half-wave rectification and normalization in such a way that the sum of signal across whole epoch equaled 1 (Figure 2.2). Shorter trials were zero-padded up to the max length of the signal (6.5 s) to keep the trials of same duration. The resulting envelope emphasizes energy fluctuations corresponds to syllabic and phonetic information to which auditory neurons are known to be sensitive [62]. The periodograms of envelopes of all stimuli demonstrated that variations in the amplitude of speech stimuli occur approximately at the rate of 3–8 Hz (see Supplementary Figure A.1). Since brain is highly sensitive to the onsets of stimuli, brain responses to sentence onsets were expected to be larger than envelope-tracking responses [2]. To account for this, the initial 500 ms of both the EEG and the acoustic envelope from each trial were discarded before cross-correlation. Then, the cross-correlation function was normalized by the activity in the [-700 0] ms lags, averaged across trials, separately at each channel. Finally, they were averaged over all chosen frontocentral electrodes and then across all participants for each presentation in both groups to get "grand mean cross-correlation" functions.

The cross-correlation function reflects the overall phase entrainment between acoustic envelope and EEG but does not display any information about the frequency components involved. To analyze phase-tracking within frequency bands, we further analyzed this cross-correlation based phase-locked activity in frequency domain using morlet wavelet decomposition for particular lags [-300 700] ms. For this, we used *timefreq()* function from the EEGLAB toolbox. Finally, these time-frequency representations were trial-averaged, normalized by the power in the [-300 -100] ms lag segment and grand-averaged over participants within a group.

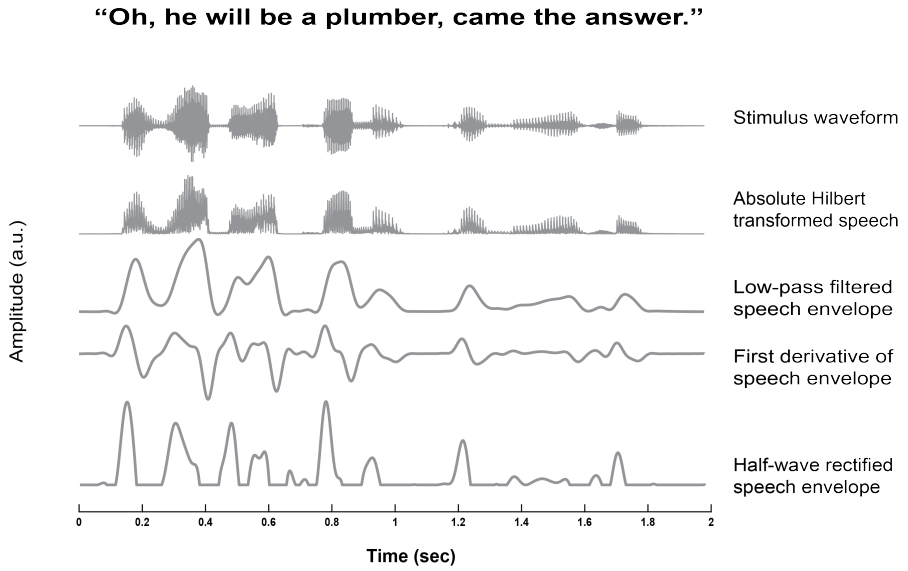


Figure 2.2: **Process for the calculation of the speech envelope and its first derivative.**

To test the statistical significance of this observed speech-entrained brain activity, we compared it with the surrogate distributions at individual level as well as at group level. These distributions were constructed by cross-correlating the acoustic envelope of each trial with the brain response of a randomly chosen trial. By repeating this simulation 200 times for each subject, we obtained a distribution of cross-correlations and their time-frequency representations, which can be seen as a null hypothesis that there was no correlation between low-frequency dynamics of speech and neural responses. For group-level significance, surrogate distributions were then averaged to make grand-averaged surrogate distributions and compared with original grand-averaged function for each time/frequency bin.

#### 2.3.2.2 Phase-locked responses to word events in continuous speech

Finding the boundaries between words in a fluent speech input is crucial for speech perception. Therefore, we next sought to answer how differently native and non-natives' brains would respond to higher-level abstract linguistic structures such as words in continuous speech. To assess phase-locked neural responses related to words in continuous speech, we first obtained the starting and ending timepoints for each word event in each speech sen-

tence from the TIMIT text transcripts. Then, we extracted brain responses between 350 ms before and after each word onset event in the speech sample for all presentations of every sentence. Notably, the initial two words were discarded from the analysis to avoid onset transient responses. Word-locked EEG responses were then normalized by subtracting the mean of entire epoch and then averaged across words and presentations, separately in both native and non-native speakers. The resultant four responses corresponding to four presentations represented time and phase-locked cortical fluctuations evoked by word boundaries, which we compared across native and non-native English speakers. It is noteworthy here that unlike behavioral analysis, these brain responses to word boundaries included English articles also.

#### 2.3.2.3 Assessment of within-sentence phase dynamics

We calculated within-sentence phase coherence (WSPC), which reflected the phase consistency among responses from all participants for each sentence within a group (native or non-native English speakers). The goal was to investigate the effect of prior knowledge on the phase-dynamics within a particular sentence. For each speech sentence, EEG responses from native and non-native English speakers were separately pooled across presentations. These responses were then transformed into time-frequency domain using morlet wavelet convolution (frequency: 1–20 Hz, wavelet cycles: 4–10, time: -2:1/sampling rate:2 sec). Each EEG response was convolved with the morlet wavelet in frequency domain and then transformed into time-domain using inverse fast Fourier transform (IFFT). This provides a complex valued analytical signal $Z_{n,t,f}$ at each time $t$ and frequency $f$ which is expressed as

$$Z_{n,t,f} = r_{n,t,f}.e^{i\theta_{n,t,f}} \tag{2.1}$$

where r is the magnitude of $Z_{n,t,f}$ and $\theta$ is the phase of the response corresponds to $n^{th}$ subject within a group at time $t$ and frequency $f$. The phase of this analytical signal is

computed as

$$\theta_{n,t,f} = Arg(Z_{n,t,f}) \tag{2.2}$$

The phases of analytical signals were then averaged to compute the coherence using the following equation,

$$WSPC_{t,f} = |\frac{1}{K}\sum_{n}^{K} e^{i\theta_{n,t,f}}| \tag{2.3}$$

where $K$ is the total number of responses for one particular sentence within a language group. The values of within-sentence phase coherence varied between 0 (no phase consistency) and 1 (perfect phase consistency). We specifically focused our analysis in 12–18 Hz frequency band corresponding phoneme rate as previous studies have shown that cortical measures of speech tracking are sensitive to phoneme-level processing [31][29]. Further, we counted the number of peaks in the phoneme-related WSPC time series for each speech stimulus using the *findpeaks()* function in MATLAB. The *findpeaks* function returns a vector containing the local maxima (peaks) in the signal.

## 2.4 Results

### 2.4.1 Native English speakers perceived speech better than their non-native counterparts

The group of non-native English speakers reported fewer words, regardless of their correctness, after listening to speech in English relative to native English speakers (Figure 2.3A). A 2x4 mixed-way ANOVA with language group (Native and Non-natives) as between-subject factor and presentation number (pre#1, pre#2, pre#3 and pre#4) as within-subject factor revealed a significant interaction for the number of words responded between language group and presentation number ($F(1.93, 50.27) = 16.43, p < 0.001, \eta^2 = 0.39$, Greenhouse-Geisser adjusted, $\varepsilon = 0.64$), the main effect of presentation number ($F(1.93, 50.27) = 141.22, p < 0.001, \eta^2 = 0.85$, Greenhouse-Geisser adjusted, $\varepsilon = 0.64$) and a significant main effect of language group ($F(1, 26) = 58.32, p < 0.001, \eta^2 = 0.69$).

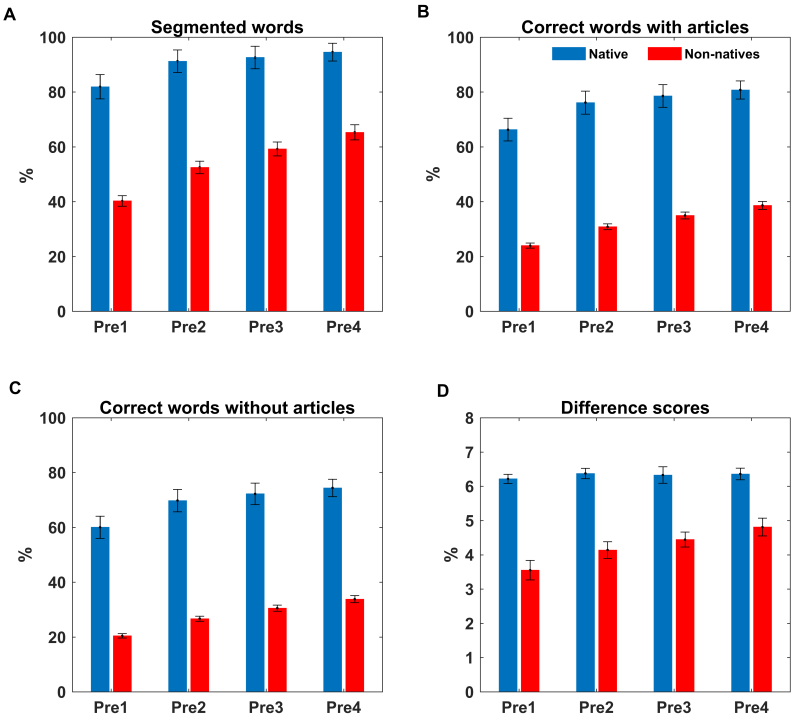


Figure 2.3: **Behavioral responses of native (blue) and non-native (red) English speakers.** A) Mean percentage of reported words, regardless of their accuracy, B) Mean percentage of correctly-written words with consideration of English articles, C) Mean percentage of correctly-written words without consideration of articles, and D) Mean percentage of difference between correctly-written words, with and without consideration of articles. This difference indicates the mean number of articles in the listener's responses. For all measures, native speakers scored significantly higher than non-native speakers ($p < 0.01$, FDR corrected) for each sentence presentation. Error bars represent standard error of the mean.

Analysis of the simple main effects showed significant effects of presentation number for both native $(F(3,24) = 18.04, p = 0.004$ Benjamini-Hochberg adjusted, $\eta^2 < 0.69)$ and non-native $(F(3,24) = 64.11, p < 0.001$ Benjamini-Hochberg adjusted, $\eta^2 < 0.89)$ English speakers. There was also a significant simple main effect of language group for each level of presentation $(F(1,26) < 67.68, p < 0.001$ Benjamini-Hochberg adjusted, $\eta^2 < 0.72)$, suggesting that native English speakers reported more words after each presentation.

Further, we performed a 2x4 mixed ANOVA with language group (Native and Non-natives) as a between-subject factor and presentation number (pre#1, pre#2, pre#3 and pre#4) as a within-subject factor and the percentage of correct words as dependent variable.

For Figure 2.3B, the ANOVA did not show any significant interaction between presentation number and language group ($F(3,78) = 1.37, p = 0.259, \eta^2 = 0.05$). However, the results showed the main effect of language group ($F(1,26) = 102.58, p < 0.001, \eta^2 = 0.80$), suggesting that native English speakers wrote more correct words than non-native speakers when articles were included in the speech text. The ANOVA also identified the main effect of presentation number ($F(3,78) = 100.08, p < 0.001, \eta^2 = 0.79$). The post-hoc analysis of pairwise comparisons revealed that listeners gradually reported more correct words as the percentage of correct words at a given presentation level was significantly different from remaining presentations ($p < 0.0001$ Benjamini-Hochberg adjusted). We found similar results for the percentage of correct words when English articles were not considered in the speech text (Figure 2.3C). The ANOVA showed non-significant interaction between language group and presentation numbers ($F(3,78) = 1.54, p = 0.211, \eta^2 = 0.06$), but the main effect of language group ($F(3,78) = 101.45, p < 0.001, \eta^2 = 0.80$) and the main effect of presentation number ($F(3,78) = 100.64, p < 0.001, \eta^2 = 0.80$). Pairwise comparisons found that the percentage of correctly-written words was significantly improved as a function of presentation number ($p < 0.0001$ Benjamini-Hochberg adjusted).

Further, native speakers did show a bigger difference between the number of correct words with and without consideration of articles (Figure 2.3D). We performed a 2x4 mixed ANOVA with language group (Native and Non-natives) as a between-subject factor and presentation number (pre#1, pre#2, pre#3 and pre#4) as a within-subject factor and the difference score as dependent variable. The results showed a significant interaction between presentation number and language group ($F(2.64, 68.63) = 6.54, p = 0.001, \eta^2 = 0.20,$ Greenhouse-Geisser adjusted $, \varepsilon = 0.88$), the main effect of language group ($F(1,26) = 65.45, p < 0.001, \eta^2 = 0.72$), and the main effect of presentation number ($F(2.64, 68.63) = 9.92, p < 0.001, \eta^2 = 0.28,$ Greenhouse-Geisser adjusted $, \varepsilon = 0.88$). The analysis of simple interaction effects identified a significant simple main effect of language group for each level of presentation ($F(1,26) < 78.49, p < 0.001$ Benjamini-Hochberg adjusted $, \eta^2 < 0.75$)

suggesting that native English speakers scored more for all presentation levels when articles were considered in the speech text. Pairwise comparisons revealed that the non-native language group revealed a significant difference among presentation levels ($F(3,24) = 9.84, p = 0.004$ Benjamini-Hochberg adjusted $, \eta^2 = 0.55$), except pre# 2 and 3 ($p = 0.065$ Benjamini-Hochberg adjusted) and pre# 3 and 4 ($p = 0.065$ Benjamini-Hochberg adjusted). However, native English speakers did not show significant differences among presentation levels ($F(3,24) = 0.25, p = 0.94$ Benjamini-Hochberg adjusted $, \eta^2 = 0.03$).

### 2.4.2 Speech tracking in the theta-band was modulated by prior expectations of the acoustics but not the prior experience of the language

We observed the robust tracking of the temporal dynamics of speech through cortical fluctuations in both native and non-native English speakers (Figure 2.4A). Both groups showed the typical positive-negative-positive pattern of speech-locked EEG activity consistently across all presentations between [0 300] ms lags as we also observed in Figure 6A and as is typically found in studies of theta-band tracking. In the native English-speaking group (Figure 2.4A, left), the speech-locked EEG activity was visibly similar across all presentations at observed peaks ( 40 ms, 132 ms and 220 ms lags), however the correlation between speech dynamics and EEG seemed to vary with presentation in the non-native group (Figure 2.4A, right). The scalp topographies of grand-averaged cross-correlation functions in both groups exhibited similar bilateral frontocentral activation for above-mentioned peaks (Figure 2.4B).

To investigate these observed differences between native and non-native English speakers and the effect of presentation on cortical tracking of speech acoustics, we looked into the time-frequency decomposition of the grand-mean cross-correlation function (Figure 2.4C). We calculated speech-locked power within the $\sim$3–6 Hz band (Figure 2.5). We selected this frequency band because: 1) it has been consistently shown to be involved in the neural tracking of speech acoustics [91]; 2) it falls under the natural range of syl-

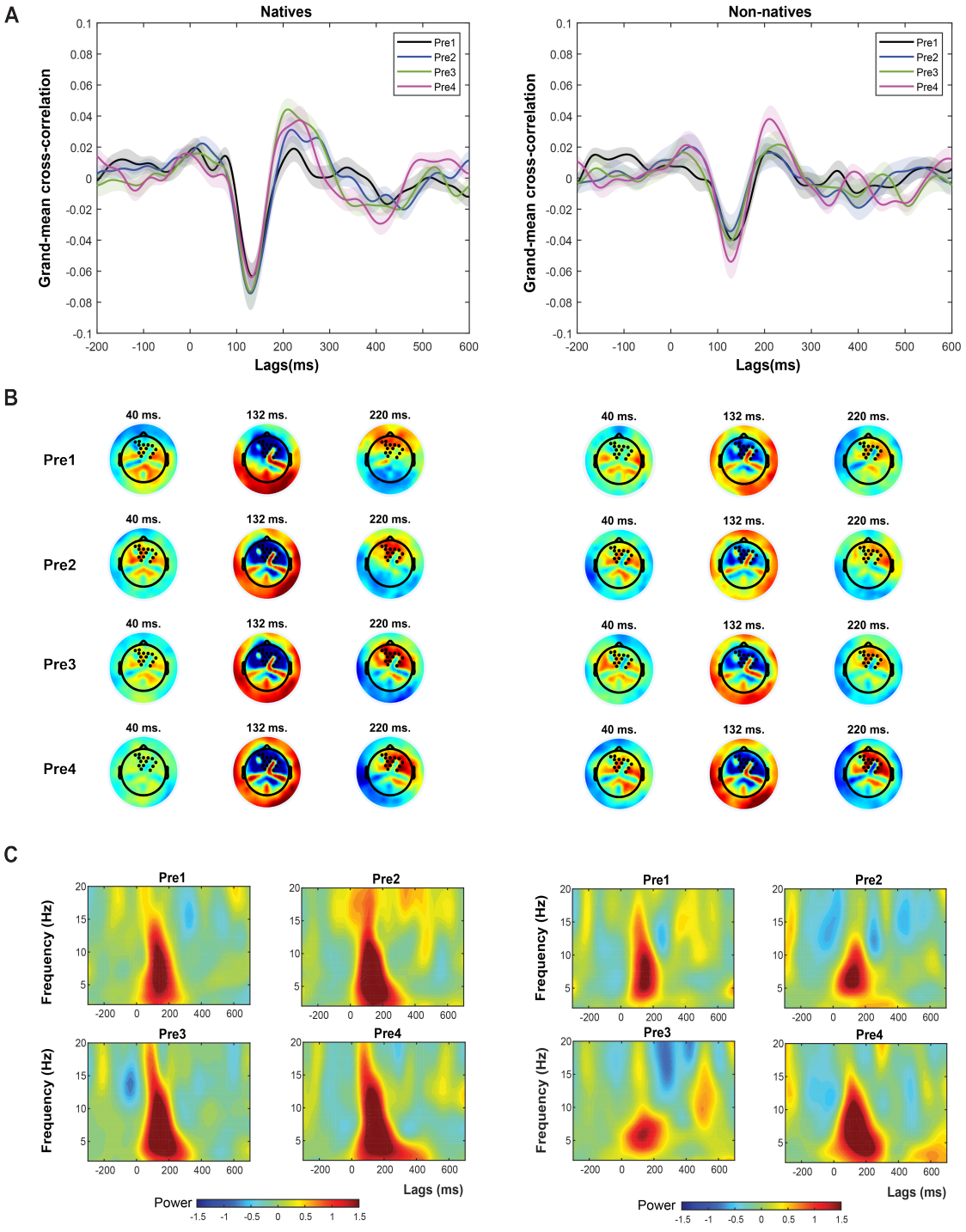


Figure 2.4: **EEG x Envelope cross-correlational results for native (left) and non-native (right) English speakers as a function of sentence presentation.** A) Grand-mean cross-correlation function plotted at different lags. The first derivative of speech stimulus was cross-correlated with corresponding EEG and averaged over frontocentral electrodes, trials, and participants within each group. Shaded area indicates standard error of the mean. B) Topoplots of grand-mean cross-correlation function corresponding to peaks at 40, 132 and 220 ms for all presentations. Black dots represent the chosen frontocentral electrodes. C) Time-frequency representations of the speech x EEG cross-correlation function.

labic rhythm across all languages [58] and spans the 3.2 Hz to 4.9 Hz estimate of syllable rate determined for our selection of TIMIT sentences; and 3) periodograms of our stimuli envelopes showed peaks concentrated within this band (see Supplementary Figure A.1). We found that the speech-locked power in this frequency band was significantly increased across participants at [0 300] ms lags, as compared to the surrogate null distributions ($p < 0.01$, one-tailed paired t-test, FDR corrected), both at the group level (Figure 2.5B) and at the individual level (Figure 2.5C).

We considered the relationship between native and non-native speakers using a 2x4 mixed-way ANOVA with language group (Native and Non-natives) as a between-subject factor and presentation number (#1, #2, #3 and #4) as a within-subject factor performed on the speech-evoked theta activity at [0 300] ms lags. This ANOVA failed to reveal a significant interaction between language group and presentation number ($F(3,75) = 1.36, p = 0.26, \eta^2 = 0.05$). The ANOVA also did not demonstrate a significant main effect of language group ($F(1,25) = 2.04, p = 0.17, \eta^2 = 0.08$), suggesting that prior experience of the language has little or no effect on the tracking of the speech envelope per se. Native and non-native speakers did not differ over all in this measure of speech tracking. However, the ANOVA did show a significant main effect of presentation number on speech-evoked theta activity at [0 300] ms lags ($F(3,75) = 3.21, p = 0.03, \eta^2 = 0.11$). Post-hoc pairwise comparisons revealed that speech-evoked theta activity in presentation #4 was significantly more than presentation #1 ($p = 0.04$, Bonferroni corrected) but not different from other presentations ($p > 0.05$, Bonferroni corrected). The other combinations such as presentation #1 and #2, presentation #1 and #3, and presentation #2 and #3 did not show any significant difference ($p > 0.05$, Bonferroni corrected).

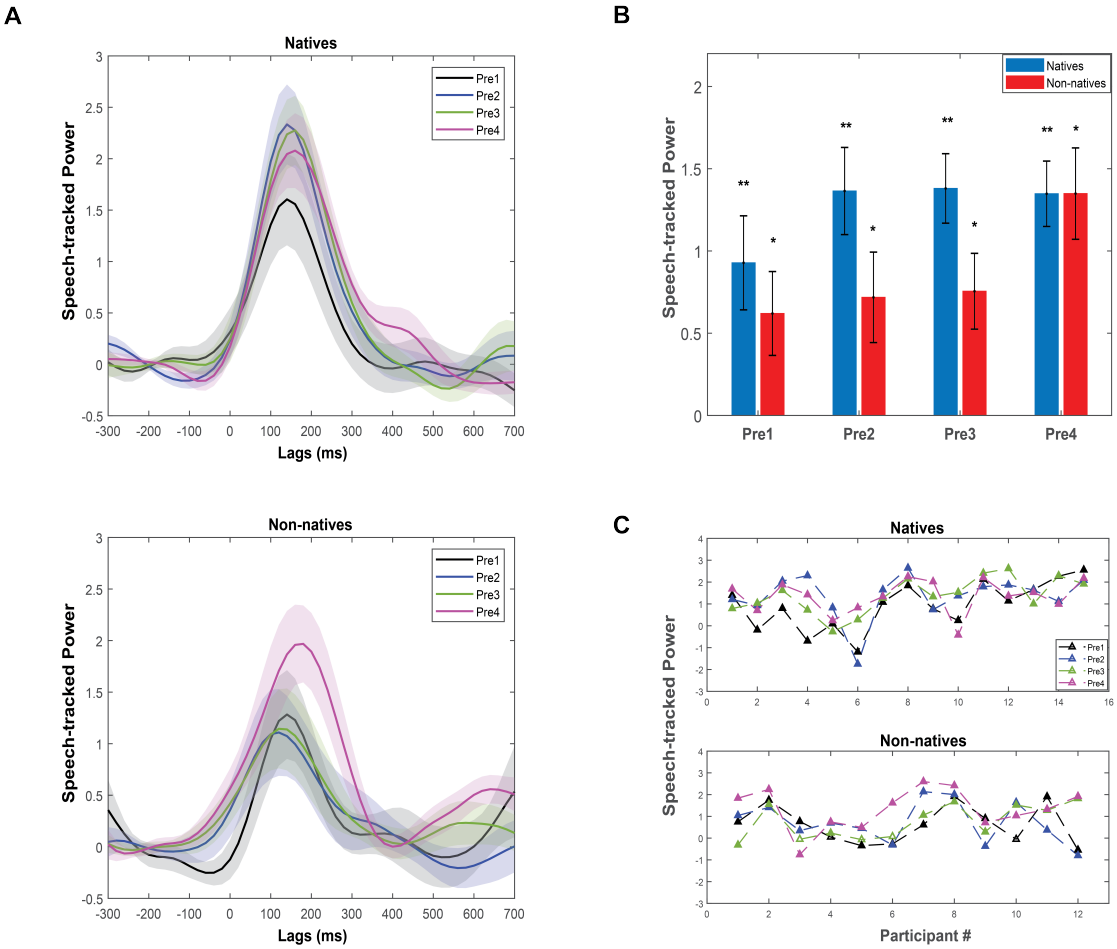


Figure 2.5: **Speech-tracked power in 3–6 Hz band.** A) Time-domain representation of speech-tracked power in 3–6 Hz band represents the highest peak between [0 300 ms] lags in native (top) and non-native (bottom) speakers. Shaded area indicates standard error of the mean. B) Speech-evoked power averaged between [0 300] ms lags in native (blue) and non-native (red) speakers. Asterisks indicate group-level statistical significance from surrogate data (one-tailed t-test, $**p < 0.001 and *p < 0.01$, FDR corrected). C) Speech-evoked power averaged between [0 300] ms lags at individual level. Filled triangles indicate statistical significance as compared to individual surrogate distributions (one-tailed t-test, $p < 0.01$, FDR corrected).

### 2.4.3 Phase dynamics related to word boundaries showed no effect of repeated presentations of speech stimuli but showed an effect of prior experience

The fluctuations in cortical responses following word events looked very different for the two groups as non-native English speakers showed a large deflection at 120 ms before word events (Figure 2.6). Noticeably, brain electrical responses before word event onsets cannot be considered here as 'prestimulus activity' in the conventional sense of the classical Event-Related Potential (ERP), for example as shown in previous studies quantifying ERPs related to the cognitive processing of temporally isolated words [9][105]. ERPs are calculated by averaging EEG activity in response to a number of similar and discrete events of interest. This process highlights brain activity that is synchronized in both time and phase to those events of interest and reduces incoherent and random non-event related activity, before and after the presentation of events [131]. We did not present our events of interest (i.e., words) in isolation. Instead, words appeared as sequences of events in natural continuous speech at semi-regular intervals and of variable durations (*average length of words*: 314.88 ms, *SD*: 37.40 ms). Therefore, the pre-word EEG responses also include time and phase-locked activity related to the previous word rather than random brain responses before a word begins (Figure 2.6).

Nevertheless, the EEG responses time-locked to word boundaries contains important information regarding phase consistency of periodic electrical activity following words in the stimulus sequence. These phase-locked signals might reveal important differences between native and non-native speakers. We therefore employed a series of analytical tests to explore these phase dynamics both within and between language groups. First, a 2x4 mixed-way ANOVA with language group type (Native and Non-natives) as a between-subject factor and presentation number (#1, #2, #3 and #4) as a within-subject factor and EEG responses evoked by word boundaries at approximately 84–90 ms latency as dependent variable was performed. The particular latency was chosen based on the visual inspection of our data. The results revealed a significant main effect of group ($F(1,25) =$

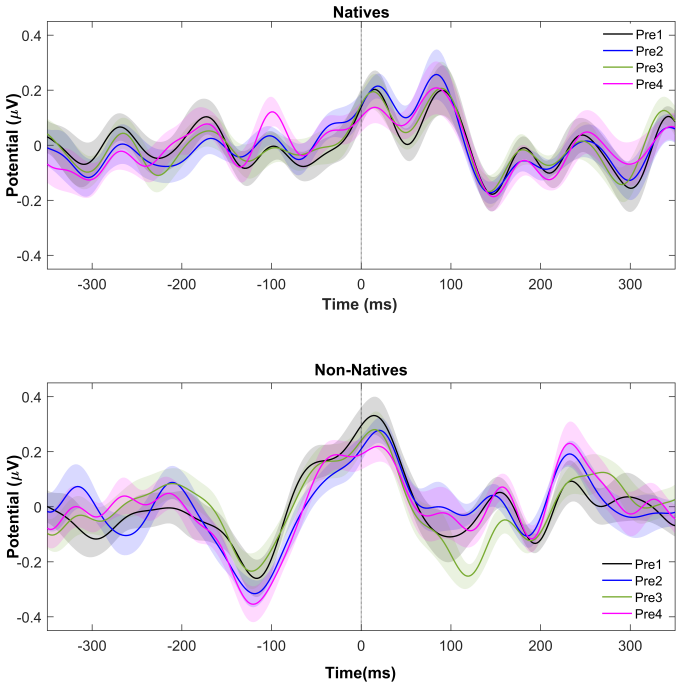


Figure 2.6: **Cortical responses phase-locked to word boundaries in native (top) and non-native (bottom) speakers.** Shaded area represents standard error of the mean.

$4.88, p = 0.037, \eta^2 = 0.16$) showing higher evoked signals in native than non-native English speakers. However, there was neither a main effect of presentation number ($F(3,75) = 1.99, p = 0.123, \eta^2 = 0.74$) nor an interaction between group type and presentation number ($F(3,75) = 0.50, p = 0.682, \eta^2 = 0.02$).

Next, to explore phase dynamics within each subject but across presentations of the same sentence, we performed pairwise cross-correlation within like stimuli. We cross-correlated phase-locked word responses corresponding to one presentation with other presentations for all possible combinations such as pre# 1&2, 1&3, 1&4, 2&3, 2&4 and 3&4. The within-subject phase similarity across presentations was confirmed by these pairwise cross-correlations which showed a peak around zero lag in both non-native and native English speakers (see Supplementary Figure A.2 & A.3).

Importantly, we attempted to find out whether the phase dynamics related to word events differed in native and non-native English speakers. Because within-subject brain responses

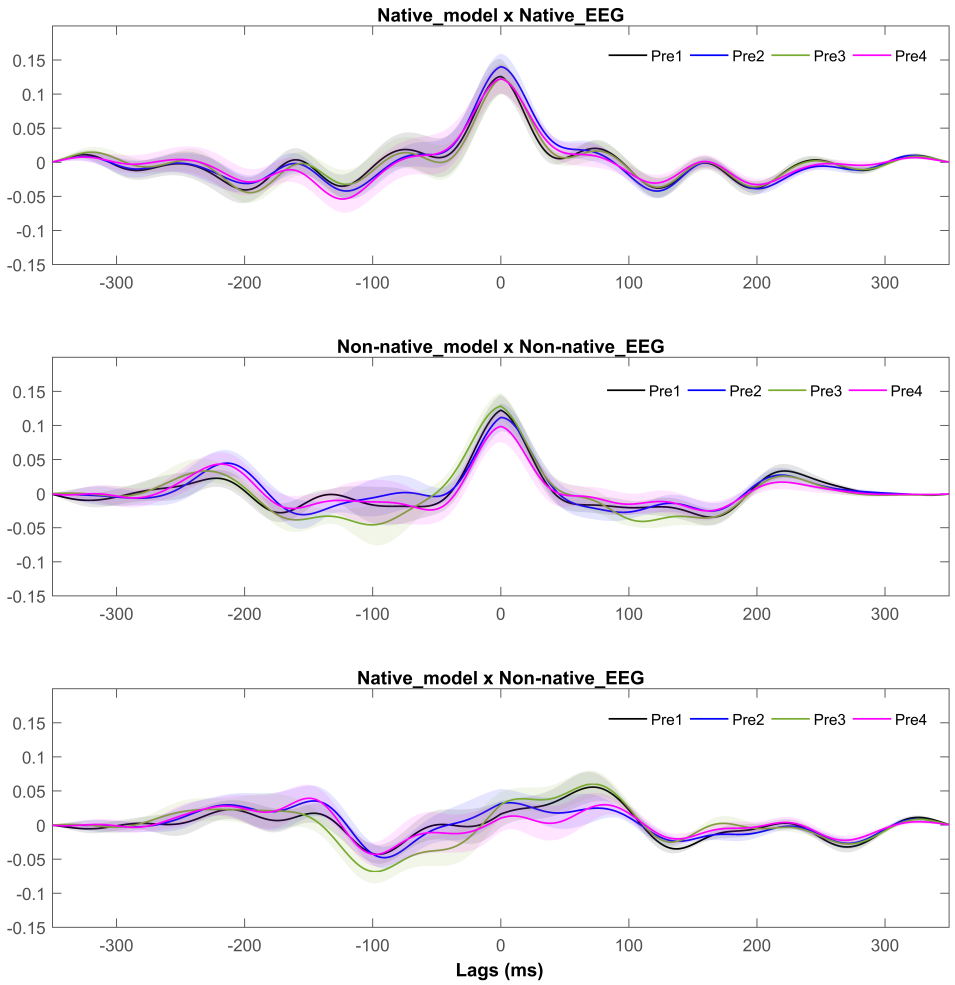


Figure 2.7: **Analysis of model fit between native and non-native groups.** Within- and between-group dynamics of phase-locked responses to word boundaries for each presentation of the sentences. (top panel) Cross-correlation between the native model and native individual EEG responses. The peak at zero lag indicates that the mean response among native participants is a good model for individual trials. (middle panel) Cross-correlation between the non-native model and non-native individual EEG responses. (bottom panel) Cross-correlation between the native model and the non-native individual EEG responses. The absence of a peak at zero lag indicates that the native model is not a good representation of the individual EEG responses of non-native listeners.

were phase-locked to each-other across presentations of the same sentence (Supplementary Figure A.2 & A.3), we averaged these responses across all presentations and subjects within a group to create a model of brain electrical responses time-locked to word boundaries. This model behaves as a template representing cortical processing of word events in those who are familiar with English language ('Native model') or are not ('Non-native model'). We first checked for within-group similarity by cross-correlating the native EEG model to each of the native speaker EEG responses (Native_model x Native_EEG). As expected, this cross-correlation resembled an autocorrelation (top panel, Figure 2.7), reflecting the fact that the responses of native speakers converge on the average "model" response. Likewise, we cross-correlated the non-native model with each of the non-native EEG responses (Non-native_model x Non-native_EEG). This cross-correlation (middle panel, Figure 2.7) revealed a similar autocorrelation pattern reflecting between-subjects similarity among non-native responses. One might concern that cross-correlating a specific subject's EEG to its own group-model which includes that specific subject would give autocorrelation pattern. To address this, we performed the split-group analysis. In this, we split the group into two subsets, averaged responses to word boundaries of one subset and then cross-correlated it to each trial and in each subject of another subset. We found similar autocorrelation patterns (Supplementary Figure A.4) as we observed in top and middle panels of the Figure 2.7.

The most important comparison, however, was the cross-correlation between the native model and non-native listener's EEG response (Native_model x Non-native_EEG). The resultant function provides us information about how the prior knowledge of a language influences the way speakers responded to word boundaries; that is, whether the EEG responses of non-native speakers were similar to the average response of the native speakers. If the word-locked EEG response was stimulus-driven and dependent on acoustics, it should be independent of language group and thus resemble an autocorrelation, whereas if prior experience affects word-locked responses there should be little or no autocorrelation in this analysis. The native_model x non-native_EEG cross-correlation (bottom panel, Figure 2.7)

revealed almost no autocorrelation at the zero lag. The average word-locked EEG response of native speakers did not provide a good model fit for the word-locked EEG responses of non-native speakers. To check the statistical significance of autocorrelation within-group (top and middle panel, Figure 2.7) and no-correlation between groups (bottom panel, Figure 2.7), we tested cross-correlation function at zero lag for all panels in Figure 2.7 against their respective null distribution. The distribution of null hypothesis (little or no autocorrelation between groups) was estimated using 5000 permutations. The cross-correlation between native model and non-native EEG responses at the zero lag was not significant ($p > 0.01$ for all presentations, two-tailed t-test, family-wise error corrected). However, the autocorrelations when the native model was cross-correlated to native listener's EEG response or the non-native model was cross-correlated to non-native listener's EEG response were found significant at the zero lag ($p < 0.01$ for all presentations, two-tailed t-test, family-wise error corrected).

### 2.4.4 Tracking of phoneme-related dynamics (12–18 Hz) correlated with the perception of words and might depend upon the prior experience of the language

We more deeply investigated phase dynamics in the 12–18 Hz frequency band corresponding to phoneme-rate dynamics of speech. We were motivated by two observations of different patterns in comparing native with non-native speakers: First, in cross-correlating the word-locked EEG model with individual word-locked responses (see above) we obtained an autocorrelation which revealed components of the EEG that were consistent across individuals.

The periodogram of this autocorrelation (Figure 2.8) revealed prominent peaks at the word ( 2–5 Hz) and phoneme ( 12–18 Hz) frequency ranges. However, the phoneme range peak was less prominent in non-native speakers ($F(1,25) = 6.17, p = 0.02, \eta^2 = 0.20$). Second, in exploring EEG responses time-locked to word boundaries, we also plotted long-duration brain responses averaged over all stimuli and individually for each presentation

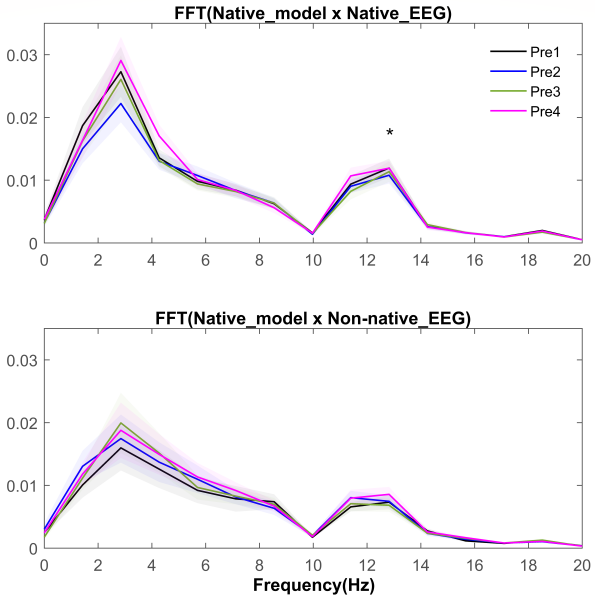


Figure 2.8: **Periodograms of the cross-correlations in (A) when native model was cross-correlated to native group (top) and non-native group (bottom).** This reveals a prominent phoneme-rate (12–18 Hz) component in the word-locked dynamics among native but not non-native listeners. Shaded area represents standard error of the mean. The asterisk sign indicates a significantly better phoneme peak in natives than non-natives.

(see Supplementary Figure A.5). We noticed in these plots a degree of transient coherence across successive presentations mainly in the native English speakers group, particularly in the first few hundred milliseconds, and approximately in the 12–18 Hz phoneme-rate range. These These observations prompted a more thorough investigation of the EEG dynamics in the 12–18 Hz band. We performed an analysis of within-sentence phase coherence (WSPC) at the 12–18 Hz phoneme-rate to provide a cortical measure of tracking phoneme-related dynamics in each spoken sentence (see section 2.3.2.3). This resulted in a time series of WSPC phase-locking values for each sentence. We predicted that peaks in this 12–18 Hz WSPC time series would be related to the occurrence of words in the stimuli. However, since words appear at different times in different sentences, speech stimuli necessarily differed in long-term dynamics. It would thus not be appropriate to average these PLV time series across different stimuli. Instead, we counted the number of peaks in the phoneme-related WSPC time series for each speech stimulus. Since the potential number of peaks,

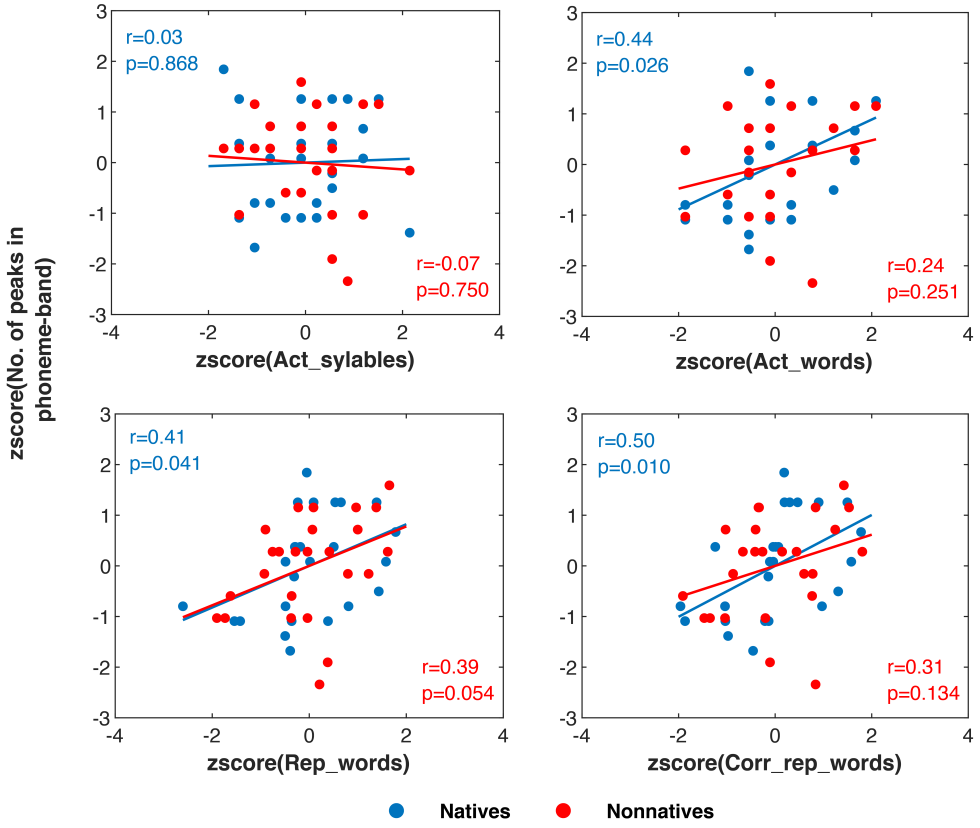


Figure 2.9: **The correlation between the number of peaks in within-sentence phase coherence at phoneme rate (12–18 Hz) and the number of:** top-left) number of actual syllables presented in the associated sentence; top-right) number of actual words presented; bottom-left) number of reported words regardless of accuracy; and bottom-right) number of correctly reported words in both native (blue) and non-native (red) speakers.

calculated by the *findpeaks()* function is unbounded, we constrained the 'Minimum Peak Prominence' to 0.0028. We correlated the number of peaks in each stimulus with both physical aspects of the speech stimuli and behavioral measures of speech perception: actual number of syllables (Act_syllables), actual number of words (Act_words), number of words reported (Rep_words), and number of words correctly reported (Corr_rep_words). The number of reported words and correctly reported words were measured by counting the total number of written words and correctly written words with articles respectively. We expected to see the effect of prior experience of English revealed in the correlation between these measures and the WSPC peaks.

In line with our hypothesis, we found in the Figure 2.9 that phoneme-related dynamics in native English speakers positively correlated with both physical and perceptual measures (actual number of words: Pearson's correlation coefficient $r = 0.44, p = 0.026$; reported words: Pearson's correlation coefficient $r = 0.41, p = 0.041$; correctly reported words: Pearson's correlation coefficient $r = 0.50, p = 0.010$). As expected, there was no significant correlation for non-native English speakers between WSPC peaks and perceptual measures (actual number of words: Pearson's correlation coefficient $r = 0.24, p = 0.251$; reported words: Pearson's correlation coefficient $r = 0.39, p = 0.054$; correctly reported words: Pearson's correlation coefficient $r = 0.31, p = 0.134$). We also did not find any correlation between WSPC peaks and actual number of syllables in any group (Native speakers: Pearson's correlation coefficient $r = 0.03, p = 0.868$; Non-native Speakers: Pearson's correlation coefficient $r = -0.07, p = 0.750$). Although the correlation coefficients were significant and non-significant in the native group and the non-native group respectively, we further checked the differences in the correlation coefficients. We used the Fisher z-transformation and found no significant differences in the strength of the relationship between WSPC peaks and perceptual measures ($z < 1.96, p > 0.05$). However, based on these results, we speculate a nesting of bursts of 12–18 Hz coherence within the slower periodicity of words, and that the process of unifying bursts of phonemes into segmented words during speech listening might depend upon the prior experience of that language. Additionally, there was qualitatively no similar pattern of correlation between peaks in WSPC series and word or syllable units when the WSPC time series was restricted to word (2–5 Hz) and syllable (3.2–4.9 Hz) frequency bands (see Supplementary Figure A.6) for word-band and A.7 for syllable-band).

## 2.5 Discussion

The goal of the present study was to investigate the modulatory effects of top-down linguistic mechanisms on neural correlates of speech segmentation. We investigated this

by comparing the scalp-recorded electrical activity of two groups that had different knowledge and experience of the English language, while they listened to natural English speech stimuli. Both native and non-native speakers heard four presentations of each sentence. By presenting identical stimuli to the two groups, we ensured that any differences in speech-tracking responses between the native and non-native speakers would be possibly due to the differences in their prior experience of English, not due to acoustic properties of stimuli. Our results clearly demonstrated the differential effects of prior experience of the language in cortical tracking of acoustic and linguistic features. We found that theta-band tracking of the acoustic envelope increased with repetitions of the same sentence, which aligned with the behavioral result. However, theta-band tracking was not significantly different between groups, despite our finding that native speakers performed better. However, we did observe different patterns of word/sentence phase dynamics among non-native speakers (Figure 2.7 and supplementary figure A.5), suggesting that prior experience of language influences cortical tracking of language-specific features, beyond simple theta-band tracking of the acoustic envelope. Also, the tracking of phonemic features, measured by within-sentence phase coherence in the 12–18 Hz frequency range, correlated with the segmentation of words, but only by native English speakers (Figure 2.9). We discuss our key findings below in detail in the light of current literature.

### 2.5.1 Prior experience with the language and speech perception

Our behavioral task revealed the differences between linguistic profiles of native and non-native English speakers. We measured behavioral performance on two perceptual measures: the ability to segment word boundaries regardless of their correctness, and the ability to recognize correct words when articles were presented in the speech text or not. Our behavioral results revealed that native English speakers outperformed non-native speakers on both measures. This difference in processing spoken sentences can be attributed to the differential prior experience of the English language. Relative to native English speakers, our

non-native participants started to learn English in their home countries, where more emphasis was given to teaching grammatical rules, and passive skills of reading and writing than actual conversational aspects of language. These results are well-aligned with suggestions in previous literature that prior experience provides statistical cues to identify and parse the boundaries of structural components of speech [120][128][133]. Moreover, the repetition of the same sentence offered gradual perceptual improvement as we found the main effect of presentation number. A similar pattern of results was also found in previous studies [68][106].

We also replicated the interesting finding that non-native English speakers often face difficulties in perceiving English articles ('a', 'an', and 'the'). This is reflected as the lesser difference in non-native speakers, at a given presentation level, between the percentage of correctly written words when articles were and were not presented in speech texts (Figure 2.3D). This is because non-native English speakers were less likely to include articles in their responses and thus do not seem to show much difference between scores when articles were and were not considered in the speech text. One possible explanation of this result is that articles and the nouns they determine might be co-articulated and thus lack clear acoustic boundaries between them. In that case, our non-native speakers who have no prior linguistic model for English, or have comparatively sparse exposure to English, would have difficulty segmenting the boundaries between articles and nouns. By contrast, our native speakers were more likely to segment the boundaries between articles and nouns and therefore scored more when articles were considered in the text. In addition, multiple presentations of the same speech sentence improved the difference score in our non-native subjects. However, it is not clear that non-native speakers learn to segment articles per-se with repeated presentations. They could be identifying and segmenting other words or articles, and probably both. On the contrary, the robust difference between correct words with and without articles in native speakers does not show any effect of repeated presentation of the stimulus. This indicates that native speakers could recognize and segment the arti-

cles co-articulated into the acoustic stream immediately after the first presentation of the sentence.

### 2.5.2 Prior experience with the language and the tracking of the speech envelope

There has been uncertainty and some degree of confusion in the literature about two, possibly related, influences on the occurrence of theta-band tracking: intelligibility and the availability of a prior language model. Typically, in the literature, intelligibility is modulated by acoustic manipulations or priming, whereas manipulating a prior language model as an independent variable requires a design using either two different languages or two different linguistic groups. Our design used both approaches by comparing two different language groups across four repetitions of the same sentence, thus modulating both prior experience with the language and intelligibility due to repetition. In agreement with previous studies, our electrophysiological results demonstrated that theta-band speech tracking does not depend upon whether the listener has prior experience with the language [34][112][113]. Among our non-native speakers, every individual exhibited robust theta-band tracking of the speech envelope (Figure 2.5C), and the group-level theta-band tracking was significantly different from a null distribution for every repetition (Figure 2.5B). Therefore, the non-native speakers were able to track the speech envelope. This is despite the finding that English speech was less intelligible for the non-native group, relative to native English speakers. This robust speech-tracking in the theta band matched previous observations of theta-band tracking below 10 Hz [37][54][90] and with a positive delay between 0–300 ms [22][107]. Consistent with this idea, studies have also shown envelope tracking responses in the primary auditory cortex of animals while listening to natural stimuli [127]. Moreover, human research has observed robust envelope-tracking responses for non-speech sounds [83][130].

If language-related mechanisms acquired through familiarity with the language contribute to the tracking of the speech envelope through interactions with lower-level acous-

tic processing, then we would expect to see different envelope tracking responses in the theta-band between language groups. In line with this, other studies have reported that the tracking of the speech envelope is reduced in situations when the intelligibility of speech is compromised as compared to intelligible speech [1][31][38][90][110]. We did not find a statistically significant group difference between native and non-native speakers with respect to theta-band tracking, but we did find a significant enhancement in theta tracking with repetition of the sentences. One possible explanation is that the tracking of the temporal envelope is mainly driven by domain-general auditory processes which are not dependent upon mechanisms that are shaped by language learning and experience [127]. Therefore, although the sparser experience with English among the non-native speakers might affect their ability to segment speech into higher-level structural units, they nevertheless robustly track the acoustic envelope itself. However, we do not entirely rule out the possibility that access to a prior language model might indeed modulate theta tracking of the envelope to some degree. The linguistic processing might reduce the envelope-tracking cortical activity for native language in competition with cortical tracking of linguistic content [137]. In contrast, the current study observed qualitatively higher envelope tracking responses among native relative to non-native speakers across repetitions, although this main effect was not significant. The discrepancy between both observations might be attributed to the difference between experiment design and the methods to analyze cortical tracking of envelope. It is also possible that the between-subjects design and limited sample size of our study did not yield enough statistical power to reveal this main effect.

Prior experience with the specific sentences presented in each trial did modulate the neural tracking of speech in the theta-band (Figure 2.5) as we did find a significant main effect of repeated presentations of the sentences. The theta-band envelope tracking responses were generally greater in the last presentation than in the first presentation, with non-native listeners reaching the same speech-tracked theta power as native listeners by the fourth presentation. This finding is similar to the results of a study [8] which showed that

the cortical tracking of degraded yet perfectly intelligible speech increases when primed by the non-degraded version of the same speech. However, the range of lags at which the effect of prior acoustic knowledge on envelope tracking responses was observed in that study (25–125 ms lag) was little narrower than the lags in the present study (0–300 ms). The difference might be attributed to the variations in stimuli and the experimental manipulations. Notably, one might suppose that prior expectations of the speech stimulus, built after repeated presentations, might cause a priming-like effect to be more pronounced among native English speakers. However, we would then expect to see a significant interaction between language group (Native and Non-native) and presentation number (#1,#2,#3 and #4) which we did not. This further suggests that theta-band envelope tracking was not driven by linguistic mechanisms related to a familiarity with the language, but rather priming with repeated speech input modulated the strength of the tracking of speech acoustics. This interpretation leaves unanswered the question of how priming actually does affect theta-band tracking: one possibility is that priming improves intelligibility (which we see in our perceptual results in Figure 2.3), and intelligibility modulates tracking. Alternatively, priming might be envelope-related such that the listener learns to expect the timing of amplitude dynamics in the repeated sentence. In either case, having a prior model of the language seems not to be necessary. It is also possible that language-related mechanisms modulate cortical responses in frequency bands other than the theta-band. Studies in which EEG/MEG responses were recorded for known and unknown languages or from native and foreign speakers of the testing language, have found that the envelope tracking remains unchanged for foreign language, but in the theta band only [34][44]. In contrast, envelope tracking is increased by language-related mechanisms in the delta band [34][44]. Therefore, both the way in which speech intelligibility is manipulated and the frequency band in which the EEG/MEG is modulated, affect the observation of envelope tracking.

### 2.5.3 Prior experience with language modulates brain dynamics to word boundaries during speech segmentation

Although theta-band tracking of the acoustic envelope, roughly at the time scale of syllables, seems not to be strongly related to prior experience of language (i.e. there was no significant main effect of language group on theta-band tracking), there is reason to expect effects of prior language experience on brain electrical dynamics related to phonemic processing and/or word segmentation. For example, using an analytical framework based on ridge regression and the prediction of EEG, Liberto et al. [30] found that cortical measures sensitive to speech-specific features improved after priming by non-degraded speech. Also, this improvement in the tracking of speech features, especially phonetic features, correlated with the perceptual clarity of degraded speech [29]. This suggests that a prior exposure of speech sounds can modulate brain dynamics as revealed by EEG. With this in mind, we considered a more detailed investigation of the EEG dynamics looking specifically for differences between native and non-native speakers related to phoneme-level processing and word-boundary segmentation.

The effect of language experience on word-boundary segmentation was clearly visible in Figure 2.6 and Figure 2.7. The presence of early brain response between 84-90 ms in native English speakers (see Figure 2.6) is consistent with previous observations showing N100 word-onset effect while listening to native language [118]. Interestingly, the effect was absent in non-natives [119], which suggests that early word-onset responses depends upon language experience rather than acoustic characteristics of the stimulus. If it only reflects the physical differences in words or in the sounds that precede initial or medial syllables of words, then both groups would show this early onset responses. However, these early onsets do not seem to be affected by multiple presentations of the stimulus (no main effect of presentation, p¡0.05). A similar experimental design where N100 ERP responses to same physical stimuli were recorded and compared before and after they learn to segment words through training [121]. In contrast to our finding, this study reported

learning effects on N100 ERP responses to word onsets in continuous stream. There might be two possibilities for different observations. First, consecutive presentations of words in a sentence could not create the similar learning effects as the 20 minutes long training did in [121]. Second, we used real words than non-words which native speakers already might have segmented during the first presentation. It is known that listeners eventually learn to recognize a stream of random syllables as word when they are presented repeatedly in an artificially constructed, continuous speech [17]. Therefore, it is possible that the cortical activity in non-natives initially tracks syllables (mainly the initial syllable) and later searches for 'words'. It is also possible that native listeners would be better at processing words with their syntactic-semantic binding as soon as they hear them, whereas non-native listeners would be familiarizing with the acoustics of words, rather than binding them to meaning. This might be why the effect of repeated presentation to syllabic boundaries looked more prominent for non-native listeners (Figure 2.5A).

The results in the Figure 2.6 also shows different pattern of phase dynamics and inter-presentation coherence. Therefore, in further analysis, we created a model of phase-locked responses to word boundaries within both native and non-native speaker groups. The model captured the phase similarities across individuals within the groups. To visualize this within-group phase similarity, we cross-correlated the model with individual EEG responses within the group. If individuals within a group exhibit a high degree of phase similarity, then this analysis is equivalent to an autocorrelation. However, if individuals exhibit unique phase-locked responses to word boundaries, then no pattern of autocorrelation would be apparent. Indeed, the correlation between the 'native model' and native speakers' EEG, resembles an autocorrelation with a clear peak at zero ms of lag (top panel, Figure 2.7). Thus, speakers who had learned English as their native language appear to segment word boundaries in a similar fashion. If native and non-native speakers exhibit similar EEG dynamics in response to word boundaries, then the correlation between the native model and the non-native EEG should also resemble an autocorrelation. However, the bottom

panel of Figure 2.7 clearly shows no autocorrelation pattern between the native model and the non-native EEG. The dissimilarity between the groups, indexed by no correlation at zero lag, suggests that non-native speakers who do not have prior experience with language are not using cues related to the segmentation for word boundaries in exactly the same ways and/or in the same time windows as native speakers [119]. Although there was a slight difference in autocorrelation patterns across groups, these patterns were virtually identical across repetitions of the same speech input within groups. This suggests that repetition did not change EEG dynamics regardless of the language experience.

### 2.5.4 Prior experience with language and phoneme-level dynamics during speech segmentation

Our finding that there is phase consistency related to word boundaries among native but not non-native speakers prompted deeper exploration of the relationship between phase-locking in the EEG and the perception of word boundaries. To this end, we computed the phase-locking value for each sentence across speakers (named as within-sentence phase coherence or WSPC) for three different frequency bands of EEG (related to word, syllable, and phoneme time scales). We found that the WSPC time series in the phoneme band (12–18 Hz) exhibited peaks and troughs, and that the number of peaks was related to word segmentation. Specifically, Figure 2.9 shows correlations between the number of peaks in the WSPC series at the phoneme rate and perceived and actual word boundaries. These correlations were significant for actual number of words, reported words, and correctly reported words, but only among native speakers. Importantly, the simpler average magnitude of phoneme-level phase-coherence could not be the appropriate measure to compare language groups due to possible confounds of inter-subject variability. Differences in terms of their experience with the language, listening abilities, phonetic awareness and perceptual competence would all tend to modulate absolute phase coherence. Our results do not show unequivocally that the peaks in the WSPC time series directly correspond to phonemes in

the speech stimulus. However, because we applied the same minimum peak-prominence threshold for both groups in counting peaks in the WSPC time series, we believe that this approach is a valid and novel potential method to compare the effect of prior language experience on speech segmentation across different groups of speakers.

An important consideration in comparing our WSPC analysis to prior work is that we identified EEG frequency bands for word-, syllable-, and phoneme-level dynamics using the actual text transcripts of the sentence stimuli. This approach is different from that of other studies, which typically identify frequency bands based on characteristics of the EEG itself (i.e., delta, theta, and alpha) (however, see [72] which takes a similar approach to the present study). Thus, the 12–18 Hz frequency range of the WSPC analysis described above aligns with the actual rate of phonemes in our speech stimuli. It is intriguing that peaks of phase coherence revealed in the WSPC time series at 12–18 Hz should correspond with word segmentation, which occurs with much slower dynamics in the range of 2–5 Hz. Native and non-native speakers received the same acoustic input, but showed different patterns in this WSPC correlation with word dynamics. Thus, it is possible that these periodic fluctuations of within-sentence phase coherence might have been due to either their prior experience of the language or their successful perception of the speech stimuli as a consequence of that prior experience. This correspondence between phoneme-level phase dynamics and word segmentation is consistent with models of spoken word recognition. Such models suggest that words should be represented as sequences of phonemes [39][94], and that word recognition is a process of matching incoming phoneme sequences with predicted sequences based on lexical candidates that are activated from the listener's mental lexicon [48][132]. This mechanism of matching top-down and bottom-up codes falls within the more general framework of predictive coding [47] and helps in word processing. The predictions that match with linguistic input increase the activation of relevant candidates and reduce the activation of irrelevant lexical items. For example, the spoken word 'starch' initially also matches 'sale', 'steel', and 'star', but only matches 'steel' and 'star' after the

initial ‘s’ and ‘t’ phonemes have been presented and resolves to only match ‘starch’ once the ‘ch’ phoneme has occurred. This process of ‘top-down’ predictions for upcoming segments and comparison with incoming phonemes until a word unit has been matched might naturally lead to phoneme-rate cortical dynamics that nests within word-rate groupings.

Interestingly, these ‘top-down’ predictive-coding mechanisms have been proposed to operate in the frequency range of 13–22 Hz [6][43][87][88]. Consistent with this idea, a recent study [111] showed higher power in the 14–21 Hz frequency band for comprehensible speech compared to incomprehensible time-compressed speech. The authors of that study suggested a role of 14–21 Hz oscillations in ‘top-down’ predictive processing during sentence comprehension. Our study extends this idea by suggesting that prior experience with the English language affects the grouping of phonemes into words through oscillations in the similar range (12–18 Hz). Speakers who have developed the same ‘prior model’ of their native language over time probably reach an unambiguous match between sensory evidence and lexical prediction at roughly the same time in the sequence of phonemes for each word. By contrast, non-native speakers probably have very different models of English and the matching of phoneme sequences with lexical predictions might happen at different times within a word, if at all. This might give rise to nesting of phoneme-rate dynamics for the native speakers but not for the non-native group. However, the lack of significant differences in the correlation coefficients put limitations to the interpretations of group differences. Considering that, we also think that these results may provide useful insights for other studies investigating the relationship between higher-frequency dynamics and word segmentation/perception.

### 2.5.5 Relationships to Low-frequency Dynamics and Prosody

We have shown that prior knowledge of the language modulates the low-frequency phase dynamics of cortical responses while responding to word boundaries, and higher-frequency dynamics in the phoneme range. However, knowledge of a language also facil-

itates the hierarchical building of distinct linguistic structures (such as syllables to words, words to phrases and phrases to sentences) and the integration of syntactic-semantic information during speech comprehension [108]. These linguistic structures occupy longer time scales than phonemes or syllables. Therefore, the cortical tracking of these linguistic structures through slow frequency oscillations in native speakers of English would also show the engagement of widely distributed cortical networks required for these combinatory operations during speech comprehension [34]. It is also possible that native English listeners might use prosodic cues from English language for computing and binding syntactic structures and for extracting semantic information during speech processing [4][24]. On the other hand, non-native English listeners whose native languages might have different prosody than English language fail to process these cues from English as native English speakers do.

### 2.5.6 Limitations of the study

The results of this chapter should be interpreted in the light of some limitations. The first limitation refers to the caveats of selecting a subset of electrodes, frequencies, and time windows for statistical inference in multidimensional EEG data. We visually inspected our data and chose a time-window for measuring the differences in cortical responses for word boundaries between natives and non-natives. This approach has commonly been used in the EEG research community, however, our results are subjected to biases and possible misinterpretation despite being consistent with previous findings. The reason is that the selection of time-window was made based on the same data as used for further analysis and particularly where the effect is largest. This can lead to distorted descriptive statistics as well as more false-positives [75][77]. However, these issues with the selective analysis can be handled by choosing electrodes, time-windows and frequencies *a-priori* and by using methods that can ensure the independence of statistical results from selection criteria [93][122]. Second, we only collected data from a small subset of the populations of interest (15 natives

and 13 non-natives). Due to limited access of students, particularly of non-native English speakers who could fit in the inclusion criteria, the sample size in this study was small. While we have reported possible estimates of effect (mean, standard deviation, t- and F-statistics, p-values, and effect size), the interpretation of our results is likely to be affected by the small sample size of our study[84]. For example, we showed a lack of significant group differences in the tracking of speech acoustics in the theta band, although the trend was in the direction of greater tracking by native speakers. It is possible that there is an effect of prior experience with the language, but that the result has missed the significance threshold because of the wider confidence interval produced by the small sample size.

# Chapter 3

# Investigation of cortical tracking of speech features in non-native listeners- A longitudinal study

The present chapter attempts to extend the hypothesis that prior experience of a language contributes in mechanisms that give rise to speech tracking responses. Specifically, this chapter investigates the longitudinal changes in cortical responses tracking speech features beyond an amplitude envelope in non-native speakers with time. Using a recently developed regression-based analysis, namely multivariate temporal response function (mTRF), we compared scalp-recorded EEG responses from native and non-native English speakers to acoustically identical sentences. We predicted that changes in the familiarity with language in non-native speakers with time would reflect as changes in EEG responses. Our results showed a robust tracking of speech envelope in the theta band regardless of prior experience with the language. Importantly, we did observe significantly lower cortical tracking of phonemes and speech envelope in the non-native group as compared to the native group in the delta band, but not in the theta band. However, there were no longitudinal effects of prior linguistic experience on phoneme-tracking and envelope-tracking responses in non-natives. The results overall suggested that speech tracking responses are not only resulted from bottom-up acoustical processing of speech input but are also modulated by top-down mechanisms related to prior experience with a language.

## 3.1 Introduction

Speech perception in the human brain is a complicated process. It involves not only lower-level processing of spectrotemporal auditory features in speech, but also relies on higher-level linguistic mechanisms that interact with acoustic input to the cortex. For example, listeners more easily recognize noisy speech streams in a familiar language than in an unfamiliar language. It is possible that familiarity with that language continuously provides feedback about upcoming events in the speech and that aids in deciphering what is being said. Although it is clear that prior experience with a language enables a mechanism of top-down processing, it is unclear what time scale this works on. On one hand the mechanism might work on a long time scale by improving the listener's understanding of the "gist" or big-picture meaning of a sentence. On the other hand, it might be that top-down mechanisms work at the time scale of much shorter speech elements such as phonemes or syllables. Interestingly, these speech events are rhythmic and occur at distinct time scales in speech. For example, syllables occur roughly at the rate of 4-8 Hz while bigger linguistic units such as words and phrases occur at the rate of 1-4 Hz [56]. Several neurophysiological investigations in past two decades have shown that low-frequency brain oscillations track different acoustic and linguistic events in speech[1][90][82][33][60][30][72][126]. It is therefore been suggested that speech tracking in distinct frequency bands, measured by electroencephalography (EEG), magnetoencephalography (MEG) or electrocorticography (ECoG), might reflect distinct computational mechanisms that form the basis for successful speech perception.

However, it is still controversial whether this speech tracking only reflects general mechanisms related to auditory processing or rather can be modulated by mechanisms related to speech and language processing at higher cognitive levels [80][115][103]. Some studies assumed that speech tracking responses are solely driven by acoustic properties of speech. They showed robust speech tracking responses even when the speech input was made unintelligible by means that disrupted higher-order linguistic operations [67][92][136].

On the contrary, other studies found that speech tracking was reduced in situations when the speech input was not as intelligible as the normal speech [33][110][38][92][129]. These observations pointed that speech tracking responses not only follow acoustics onsets in speech, but might also represent top-down processes that transform bottom-up acoustic information into different linguistic representations.

One possible indication of top-down influences is that the perception of degraded speech depends upon the understanding of language-specific contents such as lexicon, syntax and semantics of speech. For example, sentences that are spectrally-degraded by noise but retain higher semantic predictability are perceived better then spectrally-degraded sentences with lower level of semantic predictability. Consistent with this, a recent EEG study revealed that the tracking of speech was enhanced for words that were semantically closer to their context [16]. The familiarity with preceding semantic context enhanced the predictability of upcoming words and overall improved speech perception. Similarly, the perception of degraded speech improves when it is primed with the original, non-degraded version of that speech sample. Importantly, this perceptual improvement was found to be associated with increased functional connectivity between higher-order brain areas [104] and enhanced speech-evoked responses [125]. However, one inherent methodological challenge in assessing whether linguistic mechanisms modulate speech tracking responses is the way the speech input is manipulated. The manipulation in the physical properties of the speech input, either by adding noise or by vocoding alters not only lower-level acoustics but also higher-level linguistic processing [26][110]. Therefore, it becomes difficult to dissociate speech tracking responses due to changes in speech acoustics from speech tracking responses due to changes in the linguistic properties of the speech input.

In the following study, we present an approach that allowed us to separate out linguistics-driven modulations on speech tracking responses from acoustics-driven modulations. We overcame the above-mentioned methodological limitation by using only linguistic manipulation and ruling out acoustic manipulation. We did so by presenting physically identical

speech sentences to two groups of speakers who had different levels of understanding of English language. Therefore, the differences in speech tracking responses between both groups could only reflect the differences in higher-level linguistic processing that link to listener's understanding of English language. Some previous studies have attempted to segregate the aspects of linguistic processing in speech tracking responses from the aspects of acoustic processing using the similar paradigm [112][34] [137]. These studies manipulated the clarity of speech signal by adding different levels of noise (related to differences in the acoustics of the signal) and presented it in an unfamiliar language (related to differences in the accessing of language-related contents). They found that tracking of speech in the theta band was regulated by the acoustics of speech signal regardless of the familiarity with that language. However, speech tracking responses in the delta band was reported to be associated with phrasal and sentential structure of the speech [34], and therefore was different for native and non-native language [44].

In the chapter-2, we studied brain dynamics in native and non-native English listeners. Our results showed that non-native English speakers who did not have a prior language model similar to native English speakers, exhibited different speech-tracking responses measured by EEG. In the present chapter, we further extended this idea that top-down mechanisms related to the familiarity with a language modulate speech-tracking responses. Since this hypothesis predicts that changes in familiarity over time should be accompanied by changes in EEG responses, we took the approach of designing a longitudinal study of non-native speakers participating in an intensive English immersion program. We specifically aimed to estimate longitudinal changes in listeners' ability to process English sentences as their prior experience changed from that of a relatively naive non-native speaker to that of someone familiar with the language. To test this, we recruited non-native English speakers enrolled in a year-long English learning program and with the goal of following them at regular intervals of three to four months. At each follow-up, we recorded high-density EEG signals from non-native speakers while they performed a listening task in

which they typed the sentences they heard during each trial. We compared these responses to a control group of native English speakers. Importantly, we provided the same acoustic input to both groups so that the observed differences could be due to top-down linguistic processing of the speech, rather than bottom-up processing of acoustic features.

We particularly designed this experiment to explore the role of linguistic modulations in the tracking of phonemes. This was mainly inspired by our observations in Chapter-2 that suggested a difference between native and non-native listeners in brain dynamics at the time scale of phoneme representations. The difference in phoneme-level dynamics was measured by calculating the strength of the relationship between phoneme-level dynamics and perception of words using Pearson correlation in both groups. We found the significant positive correlation between brain dynamics in the 12-18 Hz frequency band (corresponding to the rate of phonemes in speech) and perception of words in natives listeners but not in non-native listeners. We therefore followed this up using a more sophisticated method namely multivariate temporal response function (mTRF) in this chapter to quantify speech tracking responses for phonemes. This method first estimates a function to predict cortical responses for the speech stimulus and then assesses speech tracking by matching the predicted cortical response with the actual response. A better matching indicates better speech tracking. This approach has previously been used in several studies investigating the cortical tracking of various speech features such as speech envelope, phonemes and phonetic features [30][60][29].

We predicted that if prior linguistic models provide top-down influences, then non-native speakers of English will exhibit weaker speech-tracking responses for phonemes and phonetic features than their native counterparts at early longitudinal time points. However, if these listeners are immersed in an English-speaking environment for a prolonged time, then their phoneme-tracking responses should become more native-like with time. We also predicted that both native and non-native speakers would be able to track speech envelope in the theta band (at the scale of syllables) in speech, based on the previous results

showing speech tracking responses in the theta band for speech in an unfamiliar language [34][137]. However, we expected to see group-differences in the delta band as linguistic proficiency has been shown to modulate the envelope-tracking responses in the delta band [44]. Our EEG results were mostly consistent with these predictions and showed significantly lower cortical tracking of phonemes in the non-native group, especially in the delta band. However, we did not see the longitudinal pattern of phoneme-tracking responses in non-natives that we expected to find. Additionally, we, as per our predictions, did observe a robust tracking of speech envelope in the theta band regardless of prior experience with the language.

## 3.2 Methods and Materials

### 3.2.1 Participants

A total of 22 native English speakers (6 males & 3 left-handed) and 21 non-native English speakers (2 males & all right-handed) participated in this study. Native English speakers (*mean age in years (SD): 20.41 (1.74), range = 19–25 years*) were recruited through an undergraduate course in the University of Lethbridge.

The group of non-native English speakers (*mean age in years (SD): 20.38 (2.16), range = 19–26 years*) consisted of 11 native Japanese speakers, 8 native Portuguese speakers, one native Thai and one native Korean speaker. All these non-native English speakers arrived Canada an average of 4.48 weeks (*SD: 1.83, range: 2–7 weeks*) prior to the study and had been taking a one-year course namely 'English for academic purposes' in the University of Lethbridge. Before that, they had been exposed to formal English in educational curriculum in their respective countries at an average age of 12.81 years (*SD: 2.44, range: 7–17 years*) but they reported that they rarely used English for regular conversations. They also reported having spent no more than a few days in any English-speaking country. The linguistic competence was assessed based on self-ratings on a ten point scale (*1=least proficient, 10=most proficient*) for speaking, listening, reading and writing skills in their native and non-native

Table 3.1: Mean Linguistics Proficiency of Non-native Participants (SD in parentheses).

| | Native language | English phase-1 | English phase-2 | English phase-3 |
|---|---|---|---|---|
| Speaking | 9.52 (0.81) | 4.48 (1.63) | 6.72 (1.18) | 8.60 (0.55) |
| Listening | 9.76 (0.54) | 5.10 (1.51) | 7.50 (1.10) | 8.80 (0.45) |
| Reading | 9.33 (0.73) | 5.86 (1.85) | 7.50 (1.58) | 9.40 (0.55) |
| Writing | 9.14 (0.57) | 4.90 (1.45) | 6.89 (1.18) | 8.60 (0.55) |

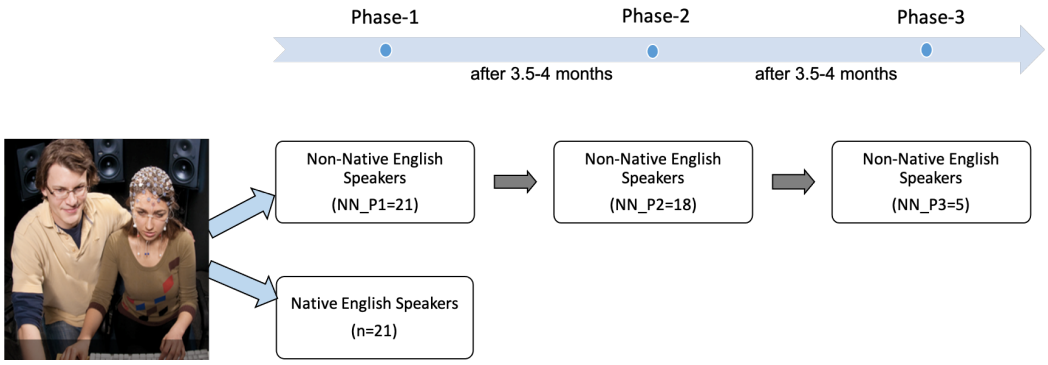


Figure 3.1: **Experimental paradigm**. Speech sentences were presented from a central loudspeaker with a central visual fixation cross on a screen. Participants were prompted to type the words that they heard into a text box after each sentence. Non-native speakers were tested in three phases 3.5-4 months apart.

language (see Table 3.1). Also, they reported to have engaged themselves in different activities (reading content, watching TV shows or movies) in English more frequently than in their own native language.

Because of the longitudinal nature of this study, these non-native participants were tested throughout a year in three phases separated by 3–4 months. However, attrition became a substantial problem in this study. Several participants could not continue to participate in later phases, often because they left the country. Consequently, we missed some data points, especially in the third phase. Therefore, 21 non-native speakers for phase-1, 18 for phase-2 and 5 for phase-3 took part in the final analysis (Figure 3.1). In a repeated-measure ANOVA design, the analysis would have done on only 5 subjects following listwise deletion method for handling missing data. Additionally, data from one native English speaking participant was excluded due to extremely noisy EEG signals.

All participants provided information about any neurological or psychiatric disorders and gave written informed consent at the beginning of the study. All reported normal hearing and normal or corrected-to- normal vision. The native group received credits for an academic course, whereas the non-native group received a gift-card worth 20 Canadian dollars for their participation. The study was approved by the Human Subjects Ethics Committee of the University of Lethbridge and done in accordance with the Declaration of Helsinki.

### 3.2.2 Stimuli and Experimental task

Sixty speech stimuli were constructed by concatenating sentences taken from TIMIT Acoustic-Phonetic Continuous Speech Corpus [49]. All chosen sentences were recorded at 16 KHz frequency by male speakers from two dialect regions across the United States and varied 3–4 s in length. All sentences were normalized to the equal root mean square (RMS) to avoid variations in loudness among speakers. For each speech stimulus, two sentences were concatenated to create a 5.2–8.4 s long unique speech sample. The order of speech stimuli was randomized for each participant and each speech stimulus was presented only once.

We used an Apple iMac with a firewire audio interface (M-Audio Firewire 410) to present stimuli in free field and send triggers to EGI Net Station data acquisition software. Participants sat one metre from a studio-grade audio monitor (Mackie HR624 MK-2) located on the auditory front midline in a sound attenuated room. It was ensured that EEG recording was free of any speaker or other electromagnetic interference to avoid artifacts. Participants held a keyboard on a table close to them, which they used to report behavioral responses. A customized MATLAB script (The MathWorks Inc., Natick, MA, USA) containing Psychophysics Toolbox functions [13] controlled the presentation of stimuli. All participants completed two blocks of 30 trials each, with a break between blocks. In each trial, participant listened to a speech stimulus while focusing eyes on a '+' sign displaying in the centre of a computer monitor directly in front of them. Participants were tasked to

type all the words they heard in the preceding trial within 30 s and press "Enter" key to start another trial. All participants were given proper instructions and practice stimuli to familiarize with the task before the experiment began. In all testing phases, the same task and stimuli but with random order were presented. To avoid any memory confound, a gap of four months was ensured between any two longitudinal testing phases.

### 3.2.3 EEG data acquisition and preprocessing

High-density EEG was recorded throughout the experiment with 128 Ag/AgCl electrodes in an elastic net (Electrical Geodesics Inc., Eugene, OR, USA), fitted on the participants' head. The brain electrical activity over scalp was recorded at 500 Hz sampling rate and electrode impedances were kept below 100 $K\Omega$. The EEG data was first pre-processed with Brain Electrical Source Analysis (BESA; Megis Software 5.3, Grafelfing, Germany). The data was first band-pass filtered at 0.5 to 30 Hz to avoid unwanted low- and high-frequency fluctuations. Artifacts due to eye movements and blinks were unavoidable in the data because of the lengthy trials. These were subsequently corrected from the data using ocular artifact-correction algorithm [69] and implemented by the BESA analysis software. Bad channels were visually identified and then were replaced using spline interpolation among neighboring electrodes after all artifact corrections. In the next steps, the data was re-referenced to an average reference and transferred to MATLAB (MATLAB version 9.1.0; The Mathworks Inc., 2016, Natick, MA, USA) for subsequent analysis using customized codes and EEGLAB functions. EEG data corresponding to each trial was band-pass filtered between 1 to 20 Hz using FIR filter and then epoched into a 10.4 s long segment beginning 700 ms before the onset of each trial. We kept all trials for both participants groups as we did not observe substantial artifacts in any of our trials.

## 3.3 Data analysis

### 3.3.1 Behavioral data analysis

We were interested in two aspects of speech perception: the ability to solve the segmentation problem, and the ability to accurately perceive words. Thus we measured listener's responses in the behavioral task in two ways: the ability to segment word boundaries (referred as 'total written words'), and the accuracy to correctly identify and report the words (referred as 'correctly written words'). To assess word boundary segmentation: we calculated the fraction of words they reported out of total words presented in each sentence, regardless of their correctness. We measured perceptual accuracy by computing the hit-rate as the percentage of correctly written words out of total actual words in the speech stimulus. We also tracked the changes in the performance of non-native participants as a function of phase over the longitudinal study.

### 3.3.2 EEG data analysis

We restricted all our EEG analysis to a cluster of 14 electrodes covering the frontocentral mid-line scalp. We a priori chose this cluster based on previous work in our lab [61] [60] which prominently showed speech-related electrical activity at these sites. The criteria for selecting particular frequency bands and temporal lags for each EEG analysis is further explained in respective data analysis and result sections below.

#### 3.3.2.1 Cross-correlational analysis of Acoustic Envelope

To measure how the brain tracks acoustic envelope modulations related to syllabic rate, we followed the same cross-correlation procedure explained in the chapter-2, and used in previous studies from our lab [61][63][62]. First, the speech envelope for a speech stimulus was computed by taking the absolute value of the Hilbert transform of the speech signal, which was low pass filtered at 20 Hz using a zero-phase finite impulse response filter, and then down-sampled this low-frequency envelope at 500 Hz to match the sample rate of

EEG. We then calculated the first derivative of the envelope, half-wave rectified and normalized it in such a way that the sum of signal across whole epoch equals 1.0 as in [65]. We zero-padded shorter trials up to the max length of the signal (8.4 s) to keep the trials the same duration. The first derivative of the envelope of each speech stimulus was then cross-correlated with the corresponding EEG signal at each channel. Before that, the initial 500 ms of both signals were discarded to prevent interference from non-specific brain responses due to stimulus onsets. The resultant cross-correlation function reflects the brain electrical responses that are phase-locked to the dynamics of acoustic envelope. This speech-tracking signal was normalized by the activity in the pre-stimulus [-700 0] ms lags averaged across trials, separately at each channel. Finally, they were averaged over all chosen frontocentral electrodes and then across all participants in groups and longitudinal phases to obtain "grand mean cross-correlation" functions for native and non-native speakers, and phases (in the case of non-native speakers).

To analyze the speech-tracked activity in frequency domain, we used time-frequency decomposition for particular lags [-300 700] ms. Finally, these time-frequency representations were trial-averaged, normalized by the power in the [-300 -100] ms lag segment and grand-averaged over participants within a group. The statistical significance of observed speech-tracked activity at particular lags was tested by comparing with the baseline activity during [-300 -100] ms lags in the particular frequency band.

#### 3.3.2.2 mTRF analysis

We further aimed to more broadly investigate the brain dynamics while perceiving various speech features such as speech envelope, phonemes, and phonetic features. For that, we used a recent method [23][31] that uses a regression technique to calculate a multivariate Temporal Response Function (mTRF). This response function acts as a filter that linearly maps speech features onto the EEG. So if a speech feature (input) passes through this filter at particular channel, it might turn speech features into EEG response at that channel

(Figure 3.2 A). The analysis had three main steps: first, calculate three different (speech envelope, phonemes and phonetic features) representations of the speech signal; second, train and evaluate the TRF and other required parameters for each speech representation; and third, predict the EEG based on the estimated TRF model (Figure 3.2 B). This analysis attempts to capture how the brain predicts neural responses to particular speech features, which will depend upon the listener's understanding of the speech input. In our analysis, we primarily used mTRF models trained on native speakers' data and then test the prediction that the non-native speakers would be poorly predicted by the model in the first phase but better predicted in later phases. The accuracy of the prediction was measured using Pearson's correlation between actual EEG signal and predicted EEG signal.The higher accuracy, therefore, not only reflects the better encoding of speech features in the EEG, but it might also mean more consistent encoding across subjects that leads to a better model. It is noteworthy here that we analyzed speech tracking responses in two distinct frequency bands (theta and delta) assuming their different functional roles [37]. Therefore, we band-pass filtered EEG responses into two frequency bands (theta: 4-8 Hz and delta:1-4 Hz) before estimating TRFs for above-mentioned speech features.

Similarly to the cross-correlation analysis described above, the envelope dynamics (Env model) was represented by the first derivative of the acoustic envelope of a speech signal which was calculated as described in the previous "cross-correlation analysis" section. The phoneme representation (Phn model) was computed by taking phonemic transcriptions of speech sentences provided by TIMIT database. The starting and ending time points of each phonemes of a speech stimulus were extracted and converted into a multivariate time series of 0s and 1s (1 for each phoneme). To obtain the phonetic feature representation (Fea model), the phoneme representation was further mapped into a space of 18 features [95][30], based on the manner of articulation (plosive, fricative, nasal, liquid, and glide), the place of articulation (bilabial, labio-dental, lingua-dental, lingua-alveolar, lingua-palatal, lingua-velar, and glottal), the voicing of a consonant (voiced and voiceless), and the back-

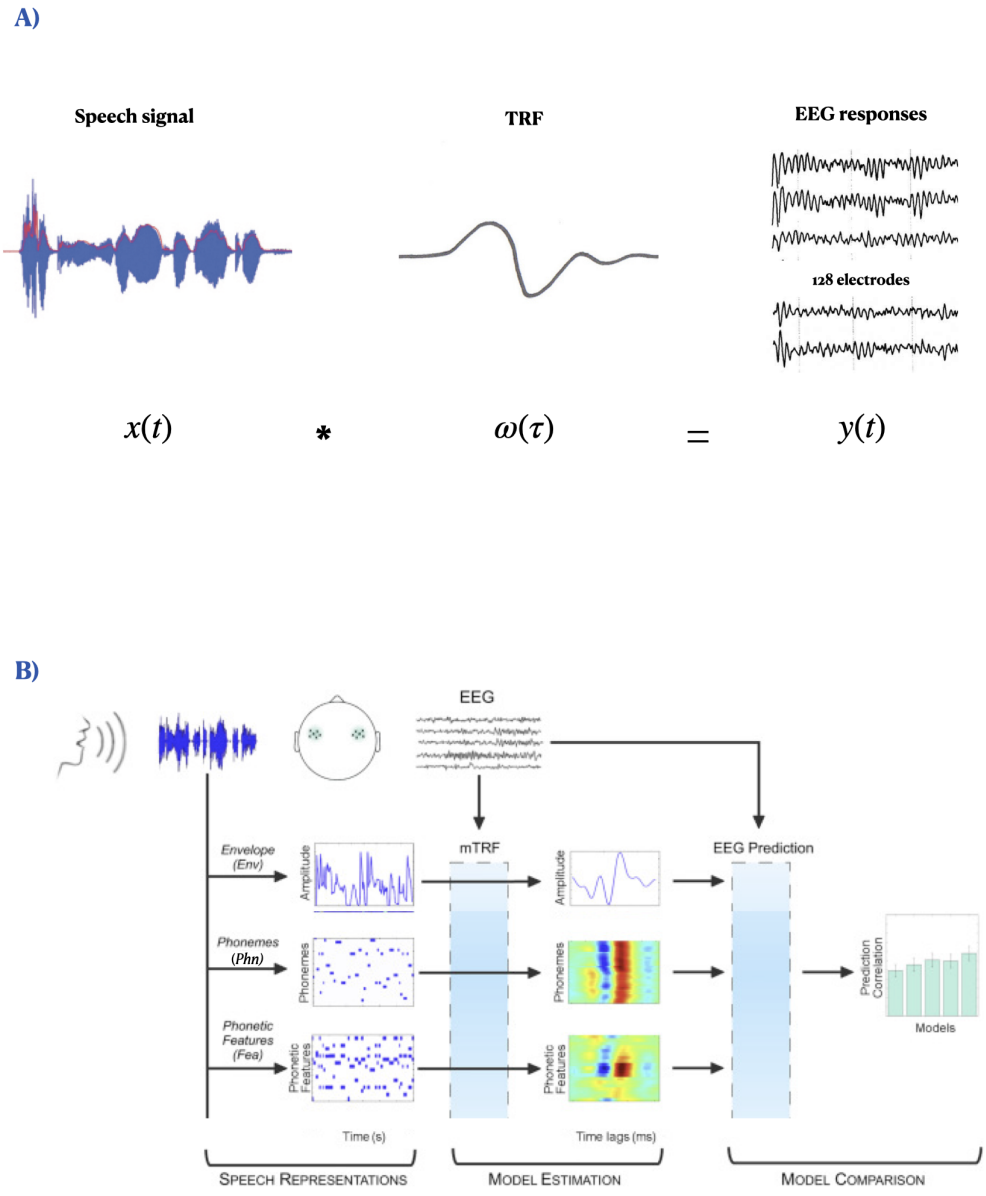


Figure 3.2: **Schematic representation of multivariate Temporal Response Function (mTRF) toolbox.** A) The technique is based on system identification technique which determines the temporal response function (TRF or ω(τ)) based on ridge regression that linearly maps speech input (x(t)) onto EEG responses (y(t)). The example of TRF is given for speech envelope at one channel. B) The analysis first trains and evaluates mTRF for a speech feature based on EEG responses which further predicts EEG responses at chosen channels. Then, the efficiency of the mTRF is measured by calculating Pearson's correlation between actual and predicted EEG. The figure is adapted from a paper by Di Liberto and colleagues (2015) [31] and described in details by Crosse and colleagues (2016) [23].

ness of a vowel (front, central, and back). As a result, a multivariate time series composed of 18 phonetic features reflecting specific articulatory properties of phonetic content in the speech was obtained.

The mapping of speech representations on the EEG can be described by the temporal response function (TRF). Therefore, a separate TRF was estimated for each speech representation using generic forward-model approach [30][60]. In this approach, we first downsampled both EEG data and speech representations at 100 Hz and converted to z-scores to reduce computational load. We then modeled mTRFs for each subject using leave-one-out cross-validation approach which trained the mTRF on 59 'training' trials of a given speech feature-EEG dataset and fit it to predict EEG for the remaining 'test' trial. The efficiency of this mTRF model was assessed by the correlation between actual and predicted EEG for the given 'test' trial. The cross-validation process was rotated 60 times such that each trial was considered as 'test' trial. This process was repeated 25 times for a range of logarithmically spaced values from $[10^1, 10^6]$ to find the appropriate value of ridge parameter, $\lambda$ for better speech feature-EEG mapping [23]. For the optimization purpose, we trained mTRFs on speech-relevant time lags ranging from -200 ms to 500 ms. The value of ridge parameter may vary across-group, therefore we normalized mTRF models for each subject by subtracting the mean pre-stimulus baseline and then dividing by the standard deviation across the above-mentioned window of chosen lags. We then selected a subset of 14 frontocentral electrodes with highest correlation coefficients for further analysis. We trained subject-specific mTRFs for natives and non-natives separately but we used subject-specific mTRFs from native participants only for further steps.

Ultimately we aimed to test the prediction that the EEG dynamics of non-native speakers is initially unlike that of native speakers, but grows more similar with immersion and experience with the spoken language (i.e. across longitudinal phases). Thus, we created a 'native mTRF model' by averaging optimized subject-specific mTRFs from native participants only. We reasoned that the ability of this native-speaker mTRF model to predict

non-native speaker's EEG provides a metric for the similarity (or dissimilarity) between native and non-native responses to speech dynamics. We thus used this generic model to predict EEG signal in both native and non-native participants. The reported prediction accuracy here was computed by taking the mean over trials, channels and subjects in a group. This cross-group EEG prediction approach was used to highlight the differences between natives and non-natives at each phases.

### 3.3.3 Statistical Analysis

All statistical tests were done using SPSS and R packages. The assumptions of normality were verified before conducting the statistical tests using both the visual inspection (histogram and Q-Q plots) and the Shapiro-Wilk's test. During statistical analyses, the measures of perceptual performance, theta-band speech-tracked activity and EEG prediction accuracies based on mTRF models were analysed with linear mixed effect models (LMEs) using lmerTest package [78] in R [114].

The first research question assessed whether the degree of linguistic experience affects the cortical tracking of different speech features. While addressing this question using linear mixed effect analysis, we put the measures of speech perception or of cortical tracking of speech features as dependent variables and *Phase* as an independent fixed-effect variable. We also included random intercepts for participants indicating variances among participants in terms of their proficiency in English in the beginning of the study.

Our next research question evaluated how these above-mentioned measures change across phases in non-natives. To examine the longitudinal changes, we constructed a mixed LME and included *Phase* as a fixed-effect term in the model to show the main effect of phase. As random effects, we included random intercepts for non-native participants (indicating variances in non-natives' English proficiency when the study started), and by-participant random slopes for the effect of phase (showing that the effect of phase is not the same for all non-native participants). All models were fit using the Maximum Likelihood

criterion. A Satterthwaite adjustment was used for computing the degree of freedom for the reported statistics.

It is notable that we chose LME over traditional repeated-measures ANOVA (rm-ANOVA) to assess longitudinal changes. The reason is that the LME model has the flexibility to handle issues such as attrition bias (loss of data over time) or power reduction due to missing data points from non-native participants in later phases while considering variabilities within and across participants. This is because the LME model considers the "missingness" in data as *missing at random* (MAR), which assumes that missingness does depend upon the information available on the observed variables but does not depend upon the information that has not been observed [89]. Therefore, data points from an individual participant can be used that were observed during initial phases, despite their missing data in the later phases. The LME model fits the regression line through the observed data points and would suggest that missing data would approximately follow the same regression line. On the other hand, a rm-ANOVA model assumes that data points are *missing completely at random* (MCAR). That means that the values of dependent variable are missing completely by chance and are independent of the values of other observed variables [89]. Therefore, observed data points will also be excluded from the analysis as missing data points can be thought of as random values among all of the values. Though linear mixed effects models might be computationally challenging and may be restrictive while making assumptions about distribution of either observations or errors in the model, they offer powerful tools to analyze longitudinal data.

## 3.4 Results

### 3.4.1 Behavioral Results

We analyzed behavioral performance on two perceptual measures: first, the ability to segment words from the acoustic stream regardless of their correctness ('total written words') and second, the ability to correctly identify and recall words ('correctly written

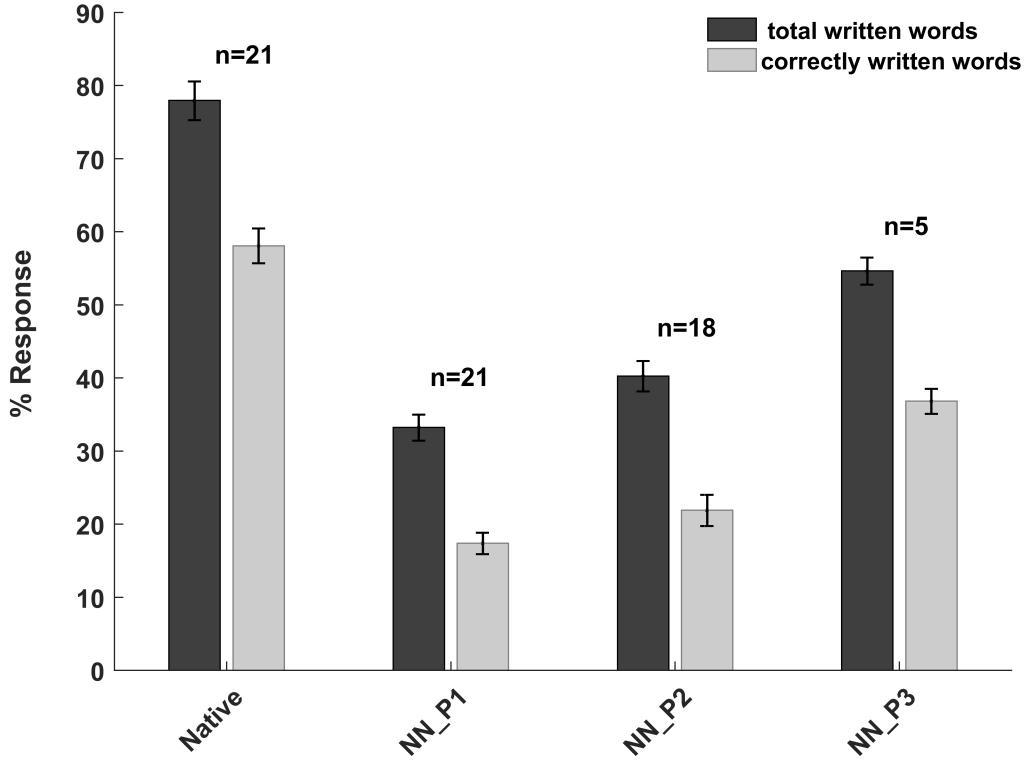


Figure 3.3: **Behavioral responses in native and non-native speakers for all three phases.** Mean percentage of total written words, regardless of their accuracy, divided by total words presented in the speech stimulus (in black) and mean percentage of correctly written words divided by total words presented in the speech stimulus (in gray). For both measures, native speakers responded better than non-native speakers in any phase ($p < 0.0001$). Also, there was a gradual improvement in non-natives as the function of phase ($p < 0.0001$). Error bars indicate standard error of the mean.

words'). The behavioral results indicate the differences in the linguistic profile of both groups. The percentage response (averaged across subjects within a group) in the Figure 3.3 showed that native English speakers wrote significantly more words than non-native English speakers (difference between Natives and NN_P1: $\hat{\beta} = -44.714, t = -14.568, p < 10^{-16}$; difference between Natives and NN_P2: $\hat{\beta} = -38.338, t = -12.434, p = 10^{-16}$;difference between Natives and NN_P3: $\hat{\beta} = -32.613, t = -9.976, p = 10^{-14}$). The Figure 3.3 also demonstrated that native speakers reported significantly more correct words than their non-native counterparts (difference between Natives and NN_P1: $\hat{\beta} = -40.680, t = -14.48, p < 10^{-16}$; difference between Natives and NN_P2: $\hat{\beta} = -36.219, t = -12.84, p < 10^{-16}$; difference between Natives and NN_P3: $\hat{\beta} = -31.583, t = -10.62, p = 10^{-15}$). Moreover, the percentage of correctly written words was significantly lower than the percentage of total written words within a group (paired t-test, $p < 10^{-11}$ for Natives, NN_P1, NN_P2 and $p < 10^{-2}$ for NN_P3). That suggests that both native and non-native speakers also reported words that were not present in the sentence they heard.

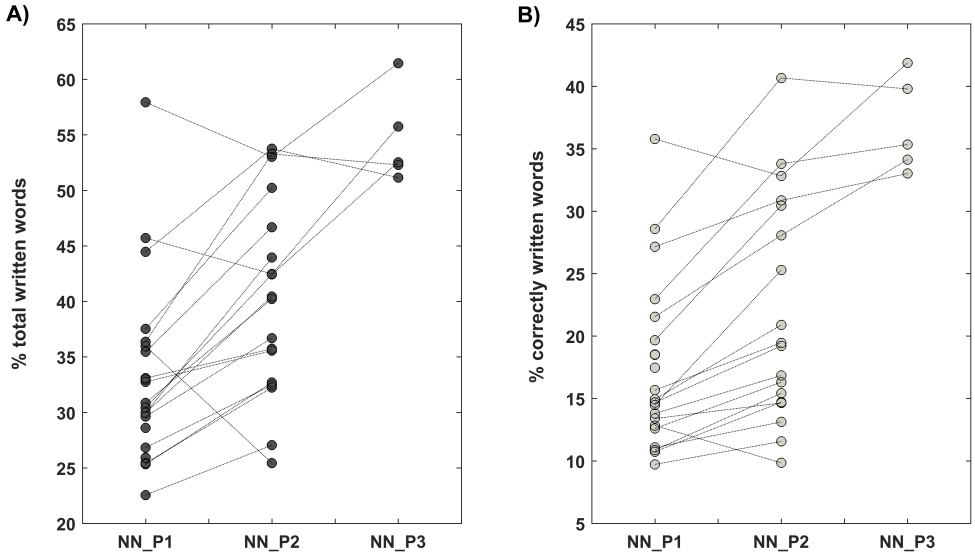


Figure 3.4: **Changes in behavioral performance of non-native speakers as a function of phase.** A) Changes in mean percentage of totally written words and B) Changes in mean percentage of correctly written words.

We also noticed the pattern of gradual improvement in performances of the non-native

group on both behavioural measures of speech perception (Figure 3.3 and 3.4). To measure the changes in non-natives as a function of longitudinal phase, we constructed separate LME models for 'total written words' and 'correctly written words' as dependent variables. The independent fixed-effect variable was *Phase* and the random factor was each individual non-native participant. For 'total written words' model, we found a significant main effect of *Phase* ($\hat{\beta} = 6.404, t = 5.323, p = 10^{-5}$). Non-natives reported 6.47 percent more 'total words' in the phase-2 and 12.49 percent more 'total words' than they did in the phase-1 (difference between NN_P2 and NN_P1:$\hat{\beta} = 6.465, t = 5.176, p = 10^{-5}$; difference between NN_P3 and NN_P1:$\hat{\beta} = 12.487, t = 4.505, p = 0.00016$). Similarly, we noticed a significant main effect of *Phase* for 'correctly written words' model ($\hat{\beta} = 4.021, t = 5.828, p = 10^{-5}$) and significant phase-dependent improvements in non-natives (difference between NN_P2 and NN_P1:$\hat{\beta} = 4.486, t = 5.688, p = 10^{-5}$; difference between NN_P3 and NN_P1:$\hat{\beta} = 5.993, t = 3.523, p = 0.0026$).

The random effects output for 'total written words' model showed that the estimated variance in non-native participants' intercepts was 91.08, the estimated variance in random slopes was 20.11 and the estimated residual variance was 247.86. The estimated random effects for 'correctly written words' model were 35.91 for between-subjects variance in intercept, 4.86 for between-subjects variance in slopes and 160.89 for residual variability. These values suggest that despite a substantial amount of between-subjects variability, most of the non-native participants improved gradually as indicated by the overall positive slope for both models (Figure 3.4).

### 3.4.2 EEG Results

We measured the cortical tracking of the acoustic envelope, time-aligned sequence of phonemes, and time-aligned sequence of phonetic features through low-frequency (1-20 Hz) EEG oscillations. To model this relationship between speech features and neural data, we used cross-correlation analysis (mainly for the speech envelope) and a regression-

based mTRF technique for the speech envelope, and for categorical representations of both phonemes and phonetic-features (see section 3.3.2.2. The mTRF method employs a cross-validation approach to train mTRF models using native participants' data and uses these models to predict EEG data in natives as well as in non-natives. The prediction of EEG (measured by Pearson's correlation 'r') was derived for both delta (1-4 Hz) and theta (4-8 Hz) bands as speech tracking has been suggested to have distinct functional role in these frequency bands [37]. The effect of prior linguistic experience on the cortical tracking of various speech features was assessed by comparing native to non-native speakers at all three phases using linear mixed effects models. The EEG prediction measures was the dependent variable and phase was a fixed-effect factor. The variabilities among participants were included as random effects. The summary statistics of EEG prediction accuracies in both the delta and theta bands for native and non-native speakers are listed in Table 3.2 and depicted in Figure 3.5. As expected, we found that EEG prediction accuracies were significantly better than chance for all models in the native group (in the delta band: $p < 0.0001$,in the theta band: $p < 0.0001$, one-sample t-test), in the NN_P1 group (in the delta band: $p < 0.0001$,in the theta band: $p < 0.0001$, one-sample t-test) and in the NN_P2 group (in the delta band: $p < 0.0001$,in the theta band: $p < 0.0001$, one-sample t-test). However, the prediction accuracies in non-natives at phase-3 did not show such significance ($p < 0.05$), except for the Phn model in the delta band (one-sample t-test, $p = 0.013$) and for the Fea model in the theta band (one-sample t-test, $p = 0.038$).

#### 3.4.2.1 Prior Linguistic experience affects cortical tracking of speech features in the delta band

In line with our hypothesis that prior linguistic experience provide top-down influences on speech tracking responses, we found differences among native and non-native speakers' neural responses (Figure 3.5). The linear mixed-effects analysis of EEG prediction accuracies using Phn model revealed that the model predicted EEG significantly better in natives

Table 3.2: Mean and standard deviation (in parentheses) of EEG prediction accuracies for envelope, phoneme and phonetic-feature model in native and non-native participants.

| | **Env model** | | **Phn model** | | **Fea model** | |
|---|---|---|---|---|---|---|
| | Delta | Theta | Delta | Theta | Delta | Theta |
| Natives | 0.0291 | 0.0408 | 0.0279 | 0.0220 | 0.0205 | 0.0277 |
| | (0.013) | (0.020) | (0.012) | (0.012) | (0.009) | (0.013) |
| NN_P1 | 0.0212 | 0.0379 | 0.0165 | 0.0187 | 0.0166 | 0.0212 |
| | (0.014) | (0.020) | (0.014) | (0.016) | (0.011) | (0.014) |
| NN_P2 | 0.0193 | 0.0370 | 0.0128 | 0.0212 | 0.0160 | 0.0223 |
| | (0.013) | (0.019) | (0.011) | (0.008) | (0.011) | (0.014) |
| NN_P3 | 0.0128 | 0.0272 | 0.0147 | 0.0136 | 0.0133 | 0.0240 |
| | (0.014) | (0.034) | (0.008) | (0.014) | (0.016) | (0.018) |

than non-natives (difference between Natives and NN_P1: $\hat{\beta} = -0.0114, t = -3.011, p = 0.004$; difference between Natives and NN_P2: $\hat{\beta} = -0.0147, t = -3.750, p < 0.0001$; difference between Natives and NN_P3: $\hat{\beta} = -0.0119, t = -2.031, p = 0.044$). Interestingly, no such differences were observed for EEG prediction in the theta band (difference between Natives and NN_P1: $\hat{\beta} = -0.0033, t = -0.851, p = 0.398$; difference between Natives and NN_P2: $\hat{\beta} = -0.0012, t = -0.291, p < 0.772$; difference between Natives and NN_P3: $\hat{\beta} = -0.0055, t = -0.931, p = 0.353$). To our surprise, linguistic experience did not seem to modulate the EEG-measured cortical tracking of phonetic-features (Figure 3.5, Fea model). Since Fea model basically represents a linear mapping of phonemes into a lower dimensional space of acoustic and articulatory features, both these models can be considered highly related. We therefore expected to see similarity between mapping of phonemes onto EEG data and mapping of phonetic features onto EEG data as also shown in a previous study [31]. However, we did not find any group differences in EEG prediction using Fea model in the delta band as we noticed in the Phn model (difference between Natives and NN_P1: $\hat{\beta} = -0.0039, t = -1.149, p = 0.254$; difference between Natives and NN_P2: $\hat{\beta} = -0.0044, t = -1.234, p < 0.220$; difference between Natives and NN_P3: $\hat{\beta} = -0.005, t = -0.947, p = 0.345$). Similarly, cortical tracking of phonetic features in

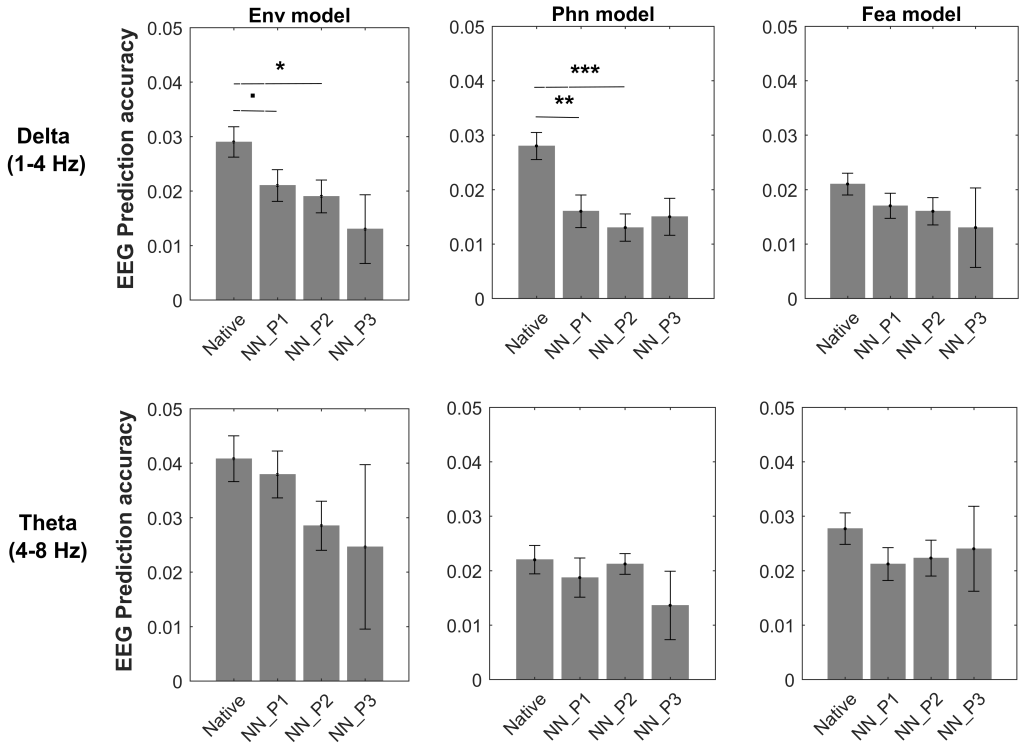


Figure 3.5: **EEG prediction using models of various speech features.** (left to right) Grand-mean prediction accuracy (Pearson's r) averaged over frontocentral electrodes, trials, and participants within each group using envelope model (Env), the phoneme model (Phn) and the phonetic-feature model (Fea). Each model was trained on EEG data from native speakers in the delta-band (top panel) and the theta-band (bottom panel) separately and was used to predict EEG in native and non-native speakers in phase-1 (NN_P1), phase-2 (NN_P2) and phase-3 (NN_P3). Significance level: $< 0.0001(***), < 0.001(**), < 0.05(*), 0.055(\cdot)$. Error bars represent standard error of the mean.

the theta band was not seen to be modulated by linguistic proficiency (difference between Natives and NN_P1: $\hat{\beta} = -0.0065, t = -1.537, p = 0.129$; difference between Natives and NN_P2: $\hat{\beta} = -0.0055, t = -1.249, p < 0.216$; difference between Natives and NN_P3: $\hat{\beta} = -0.0025, t = -0.376, p = 0.707$).

Next, our results suggest that both natives and non-natives (mainly at phase-1 and phase-2) tracked the speech envelope better than chance. However, the influences of top-down linguistic mechanisms on the tracking of speech acoustics was evident only in the delta band (Env model, Figure 3.5). The analysis of linear mixed effects model with prediction accuracies using Env model revealed the differences between natives and non-natives in the delta

band (difference between Natives and NN_P1: $\hat{\beta} = -0.0079, t = -1.958, p = 0.055$; difference between Natives and NN_P2: $\hat{\beta} = -0.01, t = -2.405, p < 0.019$; difference between Natives and NN_P3: $\hat{\beta} = -0.0122, t = -2.075, p = 0.039$).

The tracking of speech envelope in the theta-band however, did not seem to be affected by prior language experience (difference between Natives and NN_P1: $\hat{\beta} = -0.0029, t = -0.487, p = 0.628$; difference between Natives and NN_P2: $\hat{\beta} = -0.0037, t = -0.601, p < 0.550$; difference between Natives and NN_P3: $\hat{\beta} = -0.007, t = -0.837, p = 0.404$). The finding was also corroborated by our cross-correlation analysis that showed the cross-correlation of the first derivative of the speech envelope with corresponding EEG data (Figure 3.6A). The figure shows envelope-tracked cortical activity measured by EEG in both native and non-native English speakers. We found the similar positive-negative-positive pattern of speech-tracked activity between [0 300] ms lags in both groups as we observed in the chapter-3. The time-frequency decomposition of the cross-correlation function showed that cortical activity tracking the temporal dynamics of speech increased as compared to the baseline [-300 -100] ms lags (Figure 3.6B). We specifically investigated speech-tracked power within the 3-6 Hz frequency band as it has been consistently shown to be involved in the neural tracking of speech acoustics [91] and roughly aligns with the syllabic rate of our speech stimuli[53]. We found that the speech-tracked EEG activity in this frequency band significantly increased at 0-300 ms lags as compared to the baseline activity ( $p<0.01$, one-tailed paired t-test, FDR-corrected) in natives and non-native speakers in both phase-1 and phase-2 (Figure 3.7). The linear-mixed effects analysis revealed that non-natives tracked the theta-band envelope comparable to native speakers (difference between Natives and NN_P1: $\hat{\beta} = -0.0047, t = -0.232, p = 0.817$; difference between Natives and NN_P2: $\hat{\beta} = -0.0183, t = -0.866, p < 0.387$; difference between Natives and NN_P3: $\hat{\beta} = -0.0379, t = -1.159, p = 0.247$). Together, all these results suggest that the delta-band, but not theta-band EEG measures of cortical tracking of various speech representations are sensitive to the effect of prior linguistic experience.

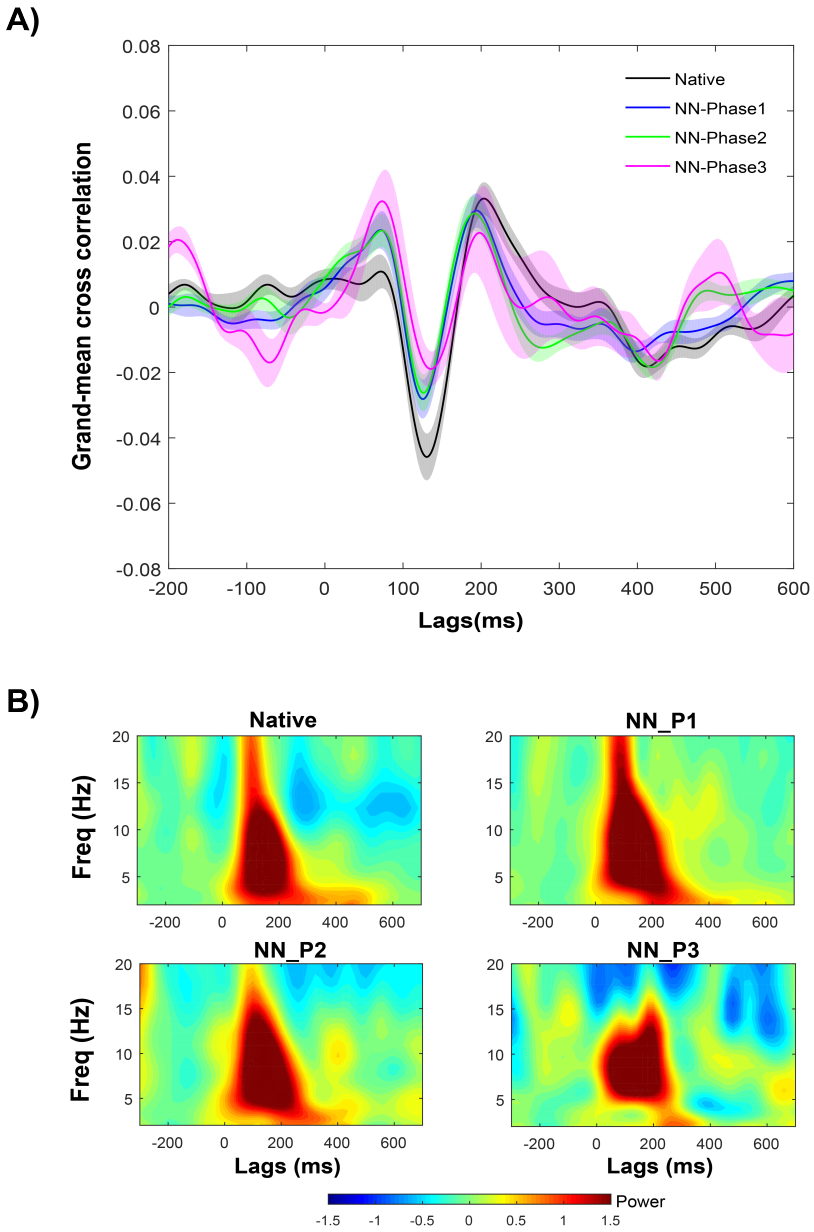


Figure 3.6: **Grand-mean cross-correlation function for native and for non-native English speakers in all three phases.** A) The first derivative of speech stimulus was cross-correlated with corresponding EEG and averaged over frontocentral electrodes, trials, and participants within each group. Shaded area indicates standard error of the mean. B) Time-frequency representations of the speech x EEG cross-correlation function.

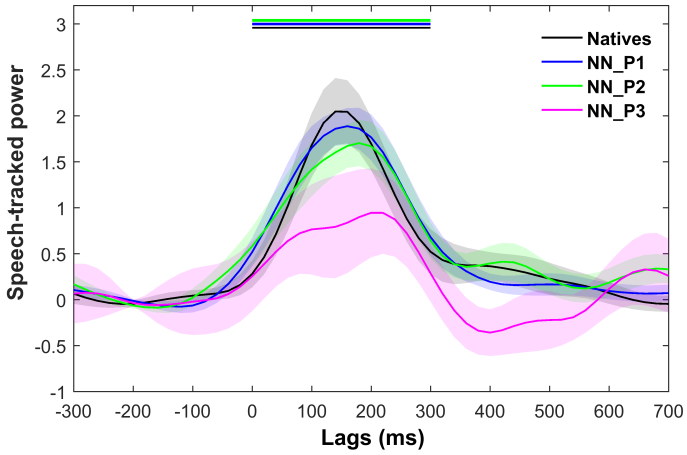


Figure 3.7: **Speech-tracked power in the theta (3.2-6 Hz) band.** Time-domain representation of speech-tracked power in 3.2–6 Hz band shows the highest peak between [0 300 ms] lags in both native and non-native speakers. Horizontal bars represent that speech-tracked theta power was significantly greater during [0-300] ms lags than the baseline [-300 -100] ms lags (One-tailed t-test, $p<0.01$, FDR corrected). Shaded area indicates standard error of the mean.

#### 3.4.2.2 Time-dependent changes in cortical tracking of speech features

An important goal to design this longitudinal study was to examine whether cortical processing of speech changes with time in non-native speakers after immersing in an English environment. To test that, we performed regression analysis on cortical measures of speech tracking in non-natives. Specifically, we constructed a separate linear mixed-effects model for each speech feature and for each frequency band, with EEG prediction accuracy as the dependent variable and phase as the fixed-effects variable. We also accounted for random effects as in random intercepts for individual non-native participants and by-participants random slopes for the effect of phase. The results of linear mixed-effects analysis are illustrated in Table 3.3. We expected that the EEG tracking of speech would change in non-natives with longitudinal phase (especially in the delta band) as the listener's linguistic model of English changed from unfamiliar to more native-like. However, the results in Table 3.3 shows no significant effect of longitudinal phase for any speech representation, something that ran in contrast to our prediction. Moreover, the overall slope ($\hat{\beta}$) indicating the effect of phase on EEG prediction accuracies was found negative for the

Table 3.3: Results of linear mixed-effects model on the effect of longitudinal phase in EEG prediction accuracies for non-native participants. Separate mixed-effects models were constructed for speech envelope (Env), phoneme (Phn) and phonetic-feature (Fea) mTRF models to examine the longitudinal changes in cortical tracking of corresponding speech representation in non-natives. For each LME model, EEG prediction accuracy was considered as the dependent variable and longitudinal phase as the fixed effect variable. Non-native participants were included as random effects to allow intercepts for participants as well as by-participant random slopes for the effect of longitudinal phase.

| | | **Delta band** | | | | **Theta band** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\beta}$ | SE($\hat{\beta}$) | $t$ | Sig.($p$) | $\hat{\beta}$ | SE($\hat{\beta}$) | $t$ | Sig.($p$) |
| Env | Intercept | 0.0232 | 0.0043 | 5.374 | 1.2e-6 | 0.0391 | 0.0056 | 6.963 | 1.3e-11 |
| | Phase | -0.0021 | 0.0023 | -0.915 | 0.361 | -0.0011 | 0.0038 | -0.290 | 0.774 |
| Phn | Intercept | 0.0176 | 0.0050 | 3.505 | 0.0019 | 0.0206 | 0.0061 | 3.377 | 0.0028 |
| | Phase | -0.0016 | 0.0024 | -0.648 | 0.5199 | -0.0009 | 0.0031 | -0.291 | 0.7739 |
| Fea | Intercept | 0.0016 | 0.0004 | 4.114 | 4.2e-5 | 0.0019 | 0.0005 | 4.084 | 5.5e-5 |
| | Phase | 2e-4 | 2.5e-3 | 0.081 | 0.936 | 1.6e-3 | 3e-3 | 0.543 | 0.592 |

Env and Phn model in the delta-band. It suggests that the delta-band cortical tracking of speech acoustics and phonemes does not improve in non-natives even after becoming more familiar with English language. However, the variability in the effect of longitudinal phase among subjects was higher in the Phn model than in the Env model (variance (Phase) for $Phnmodel = 1.8e-05; Envmodel = 5.5e-08$). That is to say, the effect of longitudinal phase in non-natives looked more consistent while tracking to the acoustic envelope than tracking the phonemes in speech. Individually, all non-native participants had negative slope for the Env model, whereas 9/21 participants had positive slope for the Phn model. Interestingly, we noticed an overall positive slope for the Fea model but found no significant differences among all three phases. Altogether, the analysis of our EEG data does not indicate any significant longitudinal changes in non-native speakers. Several caveats should be considered when interpreting this null result, and these are discussed in the following sections.

## 3.5 Discussion

The present study investigated the contributions of language-related mechanisms to the cortical tracking of acoustic and phonetic speech features by comparing neural responses from listeners having different levels of linguistic proficiency and familiarity. We recorded scalp EEG activity from native and non-native speakers of English while they listened to sentences in English, and we analyzed these signals using the multivariate temporal response function (mTRF) method based on ridge regression and EEG prediction [31][23]. The approach used in the current study allowed us to model how EEG relates to different hierarchical stages of speech processing such as encoding of acoustic and phonetic features. The present longitudinal research had two main objectives: first, whether linguistic proficiency modulates the cortical tracking of speech features (such as acoustic envelope, phonemes and phonetic features) while perceiving speech in real-time. If so, this would be reflected in lower EEG prediction accuracies of the mTRF models for non-native, relative to native English speakers. Alternatively, if speech-related EEG signals are unrelated to language proficiency (i.e. they are stimulus-driven or "bottom up"), then mTRF models should interchangeably predict native and non-native responses. Second, whether the prolonged exposure to the language improves the cortical tracking of speech features in non-native English speakers as a function of duration of immersion. To examine this particular question, we followed our non-native English speakers at three time points throughout a year with a gap of 3-4 months between sessions, and used linear mixed effects models to examine developmental trajectories of speech tracking responses for given speech features over the course of the study. Our EEG findings did provide evidence to support our first prediction: mTRF models trained on native speakers were poor models for predicting delta-band EEG of non-native listeners. Our second prediction, however, was not supported by our EEG results as we did not find any longitudinal phase related changes for EEG predictability measures in non-native speakers.

### 3.5.1 Language-related effects are specific to delta frequency

One of the main results of this study is that language-related effects on the cortical tracking of speech features emerged only in the delta-band but not in theta band of EEG (see Figure 3.5). This aligns with a current viewpoint that the speech tracking in the theta-band may primarily reflect the acoustical processing of speech, while delta-band may reflect the more abstract linguistic processing and predictive modulations of auditory processing [80]. In particular, speech tracking responses in the delta-band have been shown to be associated with the processing of higher-level linguistic units such as syntactic structure of speech [34] as well as semantic dissimilarity between successive words [16]. Moreover, speech tracking quantified by mutual information between MEG times series and speech envelope was higher in the delta band for correctly comprehended sentences than incorrectly comprehended sentences [72]. Presumably, the processing of semantic and syntactic aspects of a language are crucial for speech comprehension, but are difficult to be perceived by a non-native listener. Failure to capture those subtleties of the language is reflected in a difference between native and non-native listeners in the delta band of the EEG.

### 3.5.2 Linguistic influences on cortical tracking of phonemes

Another important finding of this study was the group differences in cortical responses while tracking phonemes presented in our speech stimuli. When the model featuring phonemic representation of sentences was trained on data from native speakers, the difference in the prediction of delta-band EEG was noticed for natives and non-natives (Top panel, Phn model in Figure 3.5). The results indicate that neural responses in the native English speakers related to the categorical encoding of phonemes in speech do not overlap with neural responses of the non-native speakers while listening to the identical stimulus. This observation fits with the view that listener's experience with a given language modulates their ability to categorize and discriminate phonemes in that language [25]. We had a mix of native-Portuguese and native-Japanese speakers who have differential repertoire of

phonemes than English and might still not be perceptually attuned to specific phones in English language. Therefore, it is likely that non-native speakers' lesser familiarity with the English language limits their sensitivity to the categorical representations of phonemes/-phonetic features, which in turn leads to the poor EEG prediction using the native phoneme model. Our finding corroborates an electrophysiological study ([14]) demonstrating the decoding of English phonemes at the single-trial level using multivariate pattern classification. The results of that study showed that the performance of a classifier trained on native speakers' data was good for native speakers but decreased for non-native speakers. The alignment between our results and those by Brandmayer and colleagues provides better ecological validity in the real world, as we used naturalistic and continuous speech sentences rather than the simple, discrete CV syllables used by Brandmayer and colleagues.

### 3.5.3 Cortical tracking of speech envelope and language experience

Our results also indicate that non-native listeners successfully tracked the speech envelope in both delta and theta band despite their sparser experience with English language (see Env model in 3.5 and 3.6. However, the modulatory effects of linguistic proficiency on the encoding of the speech envelope were found only in the delta band. One interpretation is that these frequency bands entail different functional roles in speech perception [37]. This result is consistent with similar recent studies that have shown differences in brain dynamics while tracking the acoustic envelope in native and non-native languages. These neurophysiological studies have either compared participant's envelope-tracking responses while listening to both native and foreign language [44], or compared the envelope-tracking responses in native and non-native speakers of a particular language [137]. The absence of group-differences in theta-band cortical measures of envelope tracking also shows consistency with previous results that observed robust theta-band tracking of the speech envelope even when the linguistic structure of the speech was compromised [67][136], when participants listened to sentences made of pseudo-words [92], and when participants listened

to speech stimuli in the language they were not proficient in [44]. In chapter-2, we also observed a lack of a significant main effect of language experience on theta-band envelope tracking responses, although theta responses looked qualitatively stronger in natives. Therefore, we pursued the detailed analysis of EEG activity following the acoustic envelope of speech in this chapter. The possible explanation for these above-mentioned findings is that envelope tracking responses in the theta band reflect lower-level speech processing at auditory cortical areas instead of higher-level speech processing at language-related cortical regions. The activation patterns of cortical areas within the frontal lobe measured by functional magnetic resonance imaging (fMRI), have been shown to be spatially different for languages other than the native language [76]. Whereas similar cortical areas within the temporal lobe have been activated for native and other acquired languages [76]. This might be the reason our non-native speakers were able to follow energy fluctuations in the speech successfully even though they poorly comprehended English sentences. If theta-band envelope tracking responses reflect the processing of language-specific content then we (and others) would see differences between groups. We also noticed similar results in the Chapter-2, confirming that language proficiency acquired through experience with the language does not substantially interact with the processing of acoustic cues in the speech in the theta-band. Apparently, the auditory brain can track the envelope fluctuations of any appropriately dynamic stimulus, even if it is speech of an unfamiliar language. If language familiarity does modulate theta-band envelope tracking, it is subtle and possibly due to non-linguistic factors such as auditory selective attention.

Overall, our findings in this chapter-3 may explain the similarities between our results and a previous EEG study [31] showing weaker speech tracking for time-reversed speech relative to forward speech using the similar analysis. Presenting speech reverse in time would have disrupted the linguistic structure of the speech but kept the long-term amplitude modulation spectrum intact. As a result, higher-level mechanisms related to the processing of linguistic content could not contribute as much to the tracking of time-reversed speech

as they would to the tracking of forward speech, as evidenced by variations in EEG predictability indices. Similar to the case of time-reversed speech, it is possible that top-down predictions, which come with linguistic proficiency, can not modulate bottom-up processing of acoustic and phonetic information in listeners who do not have familiarity with a language. Thus, this might be the reason why our non-native speakers show lower cortical measures of EEG prediction than natives speakers in the delta band (3.5). In this speculative view, the common aspect between time-reversed speech and speech in an unfamiliar language is that the listener's brain is unable to make predictions about upcoming speech features.

### 3.5.4 EEG encoding of different speech features within a group

The relative performance of different models (envelope, phoneme and phonetic-features model) within native speakers showed that the envelope model outperformed other models (see EEG prediction accuracies in the theta-band in Figure 3.5), something that we did not anticipate. We expected to see higher EEG prediction accuracies for the phoneme/phonetic-features models as they contain more speech-specific information and have more free dimensions. Furthermore, this finding apparently contrasts with the results of other studies showing that low-frequency EEG activity indexes the processing of phoneme/phonetic-features better than the processing of energy fluctuations in speech [31][29]. One possible reason may be the limited amount of total speech data (10 min) used in the current study as compared to the total of 20 min or more analyzed in previously-mentioned studies. As a result, it is likely that some uncommon phonemes occurred infrequently in our presentation set and thus did not contribute sufficiently to create a good model fit. Interestingly, this pattern of prediction accuracies of different models in the native group (higher prediction accuracy using the Env model than the Phn/Fea model) looked qualitatively similar to the non-native group, indicating (along with better-than-chance behavioural performance) that non-natives did not find English sentences completely incomprehensible. They could ex-

tract phonemes and their characteristics to some extent probably because of their weaker yet substantial exposure of English during formal education.

### 3.5.5 Challenges with the longitudinal study

Unlike the perceptual performance, the effect of longitudinal phase, for non-native speakers, on the cortical tracking of various speech features does not reach statistical significance in the present study (Table 3.3). One crucial consideration while discussing this null result is the longitudinal design of the study. Usually, the intensive collection of data over prolonged periods of time such as months is beneficial to strengthen the statistical power of the study and to provide more precise results. However, there were several inherent challenges associated with the process of designing, implementing, and interpreting the results of the current longitudinal study. The major challenge while conducting the current longitudinal research was to deal with missing data due to study attrition; not all participants of the study continued till the end [81][89][117]. A majority of our participants in this study left the country and moved to a different geographical location - particularly between longitudinal phases 2 and 3. That restricted our ability to follow them up remotely and to perform an experimental procedure that is highly static in nature such as collecting EEG data. Alternatively, some participants may have found the experimental procedure very complex and dull during the first few testing phases and therefore dropped out in later phases. Missing data can decrease the statistical power of the study due to the reduced sample size and increase bias. However, we tried to handle the problem of missing data by examining the mechanisms behind "missingness" and using the statistical models accordingly (as explained in the section 3.3.3). For instance, in this longitudinal study, non-native participants were tested at three phases during a year but not all participants continued for the entire length of the study. If participants' drop-out in the later phases was related to some random reasons such as budget-cut or breakdown of the infrastructure, the probability of being missing would be the same for all participants. We would then randomly choose

participants for next testing phases from participants initially tested. In that case, this missingness would have been considered *missing completely at random* (MCAR). We would use rm-ANOVA in that case and would need to exclude all data points from the participants that did not complete the study. Instead, participants in our study mainly dropped out because they completed their courses earlier and did not need to stay in the ESL program. Thus, the missingness of participant's data can be considered as *missing at random* (MAR) because the probability for values of dependent variables to be missing in the later phases depends upon the observed information at longitudinal phase 1 and/or 2 and does not depend upon the unobserved data in later phases. Therefore, we used LME models that allowed us to use all data points that were collected from the participants regardless of their presence in the later longitudinal phases of the study. Despite these advantages, the profound loss of data made LME analysis unlikely to provide significant insights into the main question of how the learning the foreign language gradually changes the neural encoding of distinct speech features in non-native speakers.

Another challenge was related to the demands associated with a longitudinal study using the EEG methodology. These demands necessarily cause within-subject variance that is confounded with the longitudinal phase: As with other longitudinal designs, it costs time as well as an infrastructure to sustain and work robustly for the whole duration of the study [101]. We made our best effort to use identical and consistent EEG recording equipment and experimental materials across all longitudinal testing phases. We used the same EEG sensor nets and Electrical Geodesics amplifier, and applied the nets in the same way. However, EEG nets became older overtime, which could have changed the signal to noise ratio across testing phases. Moreover, there was an inevitable variation in the placement of EEG nets that could have possibly affect the results overtime. Placing and EEG sensor net on a participant, then re-placing that net on the same participant but 3-4 months later exposes the design to tremendous variability. Altogether, the combination of considerable variability due to EEG nets and data attrition put severe limitations on detailed investigation and

interpretation of our EEG results. In response to this attrition and in the face of null results, we had planned to enroll another cohort of foreign students in a continuation of this study, however the global pandemic beginning in early 2020 made this impossible.

### 3.5.6 Effects of age on language learning

Another potential factor that could be considered while interpreting the absence of longitudinal changes in non-native speakers is the age of second-language acquisition. It is well documented that children show advantages in learning languages other than their native language over people who start learning in adolescence [71]. The diminishing ability to learn faster or effectively in adults has often been described through a theoretical framework called ”critical period hypothesis”. Considering this, it is possible that learning some specific aspects of English in our non-native participants were affected by the late onset of English exposure as they began acquiring English later in their life at the age between 7-17 years[64] that reflected as no improvement in speech-tracking responses.

# Chapter 4

# Conclusions

The functional roles of neural oscillations during speech processing have been an issue of great interest in the past decade. Recent research has shown that low-frequency brain oscillations track speech dynamics at various scales (termed as 'speech tracking'). Most researchers have primarily analyzed brain activity that seems to be tracking acoustics and pre-lexical information such as syllables and phonemes in speech. However, cortical tracking responses to linguistic information that does not have corresponding physical boundaries in speech are not yet understood properly. Does the brain merely follow energy fluctuations present in the acoustic stimulus during speech tracking, or does speech tracking reflect other levels of processing underlying speech perception beyond the bottom-up processing of acoustical features, or possibly both? Furthermore, do speech-tracking responses in various frequency bands track distinct aspects of speech processing? This thesis tested the theory that linguistic aspects, learned through prior language experience, are part of the mechanisms that give rise to the phenomenon of EEG speech tracking. Based on our findings we conclude that speech tracking responses are not only resulted from bottom-up acoustical processing of speech input but are also modulated by top-down mechanisms enabled by a deep familiarity with a language. Specifically, speech tracking in the theta band ($\sim$3-8 Hz) primarily reflects the processing of acoustic and phonological aspects of speech. By contrast, oscillations in the delta band ($\sim$1-4 Hz) are sensitive to the higher-level aspects that are associated with the linguistic experience. The results of this thesis contribute to current literature concerning the interpretation of cortical oscillations as biological signatures

for speech perception.

One of the main strengths of this thesis is the methodology we used to address whether speech tracking responses are related to linguistic processing. Unlike most previous studies in the field, our approach was to vary speech intelligibility by altering the prior linguistic experience rather than by adjusting the physical features of speech. That is, we took a between-groups approach to the design. This approach prevents the confounds that could happen if variations in the acoustic properties of speech are not controlled for. For instance, modifications of the acoustical features of speech stimuli might make them incomprehensible, but then the changes in the speech tracking responses might be mainly attributed to the disruptions in the spectrotemporal characteristics of the speech signal instead of the processing of linguistic structure of the speech input. Therefore we believe that comparing listeners with different experience with the testing language while keeping the acoustic structure of the speech intact can be a better approach for determining language-related contributions in speech tracking responses.

One of the most compelling hallmarks of familiarity with a spoken language is the ability to parse the acoustic stream into words. Finding perceptual boundaries between words, even when they are not accompanied by discontinuities in the speech envelope, is a challenging computational problem known as the Segmentation Problem. Anyone with experience trying to learn a second language (especially in adult life) will be familiar with the frustrating sense that the unfamiliar speech sounds like a meaningless stream of sounds, rather than discrete words. In Chapter-2 we addressed the speech segmentation problem by examining speech-related brain responses that rely on prior experience of the language at various steps of speech processing. In particular, we presented native and non-native speakers of the English language with identical acoustic stimuli and compared the dynamics of brain electrical activity in response to the boundaries of language-specific structural units, particularly words. Doing so we ensured that the observed differences in neural correlates of speech segmentation between both groups would be explained by the differences

in their ability to segment speech into linguistic units, rather than differences in the physical properties of speech sentences. In a simple listening task, participants listened to English sentences and responded by writing words they could recall immediately after listening to them. In this study, each sentence was presented four times in sequence. Thus, we modified speech intelligibility by considering the prior experience with language as well as by priming the speech sentence with its previous presentation. We tested two hypotheses: First, the brain's ability to identify and parse linguistic units in speech is dependent upon a deep understanding of the language. We reasoned therefore, that native and non-native speakers would show different phase dynamics underlying the tracking of linguistic units in speech, particularly words. As we predicted, we clearly found group-differences in the phase dynamics of EEG responses to both word boundaries. Moreover, our results provided suggestions regarding the group-differences in the relation between the phase dynamics at the time scale of phonemes (12-18 Hz) and the perception of words. These results indicate that the prior experience with the language influences speech tracking responses: native English speakers showed different phase dynamics of neural oscillatory responses to lexical and sentential information relative to non-native speakers. Second, cortical tracking responses at the syllabic time scale mainly reflect the bottom-up processing of acoustical dynamics of speech. We assessed EEG-measured cortical responses to speech acoustics of both group of listeners. Our results validate this hypothesis by showing that both our native and non-native speakers similarly tracked the speech envelope in the theta band, in agreement with recent studies [44][34]. If language-related mechanisms acquired through familiarity with the language modulate the bottom-up acoustic or syllabic-level processing and contribute to the tracking of the speech envelope, then we would expect to see different envelope tracking responses in the theta-band between language groups. Although, theta-band tracking was somewhat less robust among the non-native speakers, this difference was not statistically significant. Given that our study enrolled a relatively large sample and used natural stimuli similar to other prior studies, we conclude that there is no prominent

modulation of theta-band tracking by language familiarity (and subsequent intelligibility). Overall, based on these results presented in Chapter-2 we can conclude that cortical tracking responses also represent top-down linguistic information beyond the bottom-up sensory information in a continuous speech signal.

In Chapter-3 we expand the idea that top-down mechanisms related to prior experience with a language modulate speech tracking responses in two ways. First, this chapter was mainly inspired by the results observed in Chapter-2 that suggest that brain activity temporally correlating with slow rhythmic fluctuations in speech at theta ( 3-8 Hz) rate are mainly driven by general processing of the acoustic features of the speech, but not by linguistic processing in speech. In this view, listeners can track the speech envelope of the language they are not well versed with. However, does this mean that prior experience with the language has no role whatsoever when neural oscillations follow amplitude modulations in speech? Or, is it possible that neural oscillations could represent the modulations related to complex linguistic mechanisms on sound processing at time scales other than theta band? The findings of Chapter-2 provide cues about speech dynamics outside of the theta band of the acoustic envelope: 1) the phase dynamics of EEG responses to word-boundaries looked different among both groups of listeners 2) the within-sentence phase coherence (WSPC) in the range of 12-18 Hz showed interesting patterns and hinted at the nesting of faster phoneme-level bursts within slower periodicity of words. Previous studies used a more sophisticated analytical approach based on regression (termed as multivariate temporal response function (mTRF))[23][29] to measure encoding of speech in distinct bands. Therefore, we used mTRF approach to measure cortical tracking of not only the acoustic envelope but also of phonemic/phonetic features of speech in two distinct frequency bands, theta (1-4 Hz) and delta (1-4 Hz), and compared them between native and non-native speakers of English. We predicted that changing the prior experience with the language would influence the stimulus-driven aspects related to the processing of sensory information in the delta band, and not in the theta band. Our results supported this prediction as native and

non-native speakers of English showed similar cortical responses reflecting the encoding of envelope and phonemes in the theta band but revealed differences in the delta band, aligning with current literature [34][72][96]. Oscillatory activity in the delta band has been found to be involved in the chunking of syllables into abstract syntactic structures such as words and phrases [10][11][50]. Therefore, when native speakers tracked the speech envelope through delta oscillations better than non-native speakers, we speculate that it points to an underlying process that unifies syllables into word-like building blocks, which is less likely to happen without prior experience with the language. It is also possible that the observed group differences in the delta band during the tracking of phonemes might be consequences of other cognitive mechanisms that influence the generation and expectation of upcoming words as the sequence of phonemes is unfolded. Nonetheless, this result reinforces our interpretation of the results presented in Chapter-2 that prior experience with the spoken language may modulate the correlation between phoneme-rate dynamics (12–18 Hz) and perceptual segmentation of words.

Second, Chapter-3 took the additional approach of employing a longitudinal design to test the hypothesis that changes in the familiarity with the foreign language due to prolonged immersion in that foreign language environment should be accompanied by changes in the cortical tracking of speech in that language. Therefore, we predicted that speech tracking responses in non-natives would look more native-like as they gained more familiarity with the language. This prediction was not supported by our electrophysiological data as there were no changes in the cortical tracking of various speech features as a function of the longitudinal phase. We attribute our null results to several constraints that are associated with designing and executing a longitudinal experiment - notably data attrition and within-subject variability over repeated sessions. We recommend future studies utilizing a similar experimental paradigm yet minimising data attrition in order to get a better understanding of long-term effects of language learning on electrophysiological measures of speech tracking.

These studies demonstrate that there are top-down language-related influences on brain

oscillatory activity in response to speech, and that cortical tracking of clear speech is driven not only by acoustic boundaries in a speech signal but also by linguistic mechanisms, putatively learned through prior experience of the language. Our results help in advancing the current theories of speech perception that have attempted to place the role of cortical tracking in speech processing. Most of those theories have suggested that brain oscillatory activity tracks the rhythmic acoustic cues in the speech that occur roughly at the syllabic rate and enables speech perception by segmenting continuous speech into syllable-size chunks [52][56]. Specifically, these theories conceptualize theta-band cortical tracking of speech amplitude modulations as a measure or correlate of speech perception. In strict accordance to these theories, non-native English speakers in our studies would have shown little speech tracking in the theta-band as they comprehended speech poorer than native English speakers. Instead, both native and non-native speakers showed comparable tracking of the speech envelope in the theta-band. Thus, our results challenge a strict interpretation of these theories that might link envelope tracking to speech perception, and rather support the idea that the tracking of speech envelope by theta-band dynamics might not be sufficient for speech perception. Clearly other higher-level processes are involved. In particular, these theories of speech perception seem to be conservative as they overly focus on acoustic processing of an auditory input by theta-band oscillations as a mechanism of speech perception, and ignore other top-down computational aspects working on other time scales and are crucial for speech perception. The present work support a different view in which both acoustic and linguistic events in the speech stream drive EEG dynamics. For example, linguistic events might follow from predictive coding mechanisms and might reflect the synthesis of predictions based on accumulated linguistic knowledge with bottom-up sensory processing of speech input. In line with this proposal, listeners who have insufficient experience with a language would show reduced neural oscillatory activity during the computation of abstract linguistic structures such as words, syntax, sentences and so on, and in time-scales that are sensitive to the generative and predictive functions during abstract linguistic processing.

This view is well-aligned with previous work showing that neural tracking of speech determines the words we hear [79] and also represents the categorical perception of phonemes [29].

However, it is difficult to know the precise neuroanatomical sources of the observed cortical tracking responses in our studies. In the literature, it has been proposed that the functional engagement of the underlying oscillatory network leverages the representation stages of cognitive abstraction [97][45]: lower stages of abstraction (e.g. acoustic cues and sub-syllabic level) might employ sensory and association cortices; higher abstraction stages such as lexical-, syntactic- and semantic-level might extend cortical networks beyond sensory cortices. Based on this idea, we may speculate that the reduction in speech tracking responses in non-native speakers results from the lesser or probably negligible involvement of larger oscillatory networks extended to frontal and posterior cortices. In other words, non-familiar speech simply doesn't activate as much cortex as familiar speech, because it doesn't engage the complete set of language-processing mechanisms. Evidence supporting this proposal is observed in neural oscillations accompanying the categorization of phonemic features in superior temporal gyrus [95]. The authors reported increased synchronization of delta-band oscillations in frontal cortices while listening to clear speech as compared to spectrally-rotated speech for which linguistic inferences and predictions are disrupted [99]. Furthermore, stronger cortical tracking of phrases and words was found in the left premotor cortex and in left middle temporal cortex, respectively, for correctly comprehended speech [72]. Indeed, all these studies offer valuable hypothesis for evaluating top-down language-related effects of cortical speech tracking. Since our studies were not designed to answer questions about anatomical localization, we are not able to bring our data to bear on this hypothesis. Future studies are certainly needed that could use the superior spatial localization capabilities of intracranial recording or MEG and advance techniques of EEG source imaging for determining the causal relationship among the neural generators underlying higher-level abstract and generative linguistic aspects, and bottom-up

sensory aspects of cortical tracking responses that have been described in this thesis.

It would be worth to explore whether poor cortical tracking 'causes' poor performance on perceptual measures of speech perception, or poor cortical tracking of speech is 'caused by' weaker perceptual engagement of non-natives due to the unfamiliarity with the language, or if both poor comprehension and poor tracking are reflective of the same (failed) perception mechanisms. These possibilities need to be verified in future work with the same experimental design in order to quantify if frontal cortices causally impact auditory cortical areas, similar to what has been done to determine the effect of focal attention on the cortical tracking of speech envelope [86]. Nevertheless, the correspondence between our behavioural results and EEG prediction measures regarding group-differences certainly suggests the association of speech tracking measures with speech intelligibility [110][38].

While the studies described in this thesis yield important insights, there are certain constraints that naturally influence the outcomes and conclusions of the thesis. The most challenging aspect of our studies was regarding the recruitment of participants. The details of their selection criteria along with their demographic and linguistic profiles have already been described in section 2.2.1 and 3.2.1. We recruited native and non-native speakers of the English language for our experiments who were available on our university campus. Therefore, like many studies of special populations, we had a relatively small pool of non-native English speakers who could fit in the inclusion criteria. Furthermore, many of our non-native participants described in Chapter-3 dropped out of the study between longitudinal phase 2 and 3, simply because they left the country. We had planned to recruit another cohort of international students in a continuation of our longitudinal study to overcome the limits imposed by participant attrition, but the global pandemic that began in early 2020 rendered this impossible. Due to one or both of these reasons, our studies remained limited to smaller sample size. While we have reported possible estimates of effect (mean, standard deviation, t- and F- statistics, p-values, and effect size), the interpretation of the results in this thesis, particularly in Chapter-2, is likely to be affected by small sample sizes

of our studies [19][84]. Additionally, there could be a potential impact of heterogeneity of our non-native group due to handedness (as there were a few left-handed participants, mainly in Chapter-2) and their diverse linguistic background. This might cause more variance among participants which could under-powered the results, especially in Chapter-3. Our participants in the non-native group were mainly native speakers of Japanese. Thus, the interpretation of the results of this thesis should be limited to the groups examined in this thesis. Though our findings show consistency with other neurophysiological studies that have shown differential brain dynamics among Chinese and English speakers [34], Dutch and English speakers [44], and Cantonese and Mandarin speakers [137], we suggest that readers should keep this limitation in mind when interpreting the results of the thesis. It is probable that novel relationships between brain dynamics and acoustic and linguistic features remain to be discovered when different linguistic populations are considered in future research.

It is also worth to notice that we here did not consider the role of other processes related to working memory, attention, auditory-motor skills and task difficulty while interpreting our results. These processes are likely to interact with speech perception mechanisms in our speech segmentation task. For instance, it is possible that non-native listeners were not paying as much attention towards listening speech in an unfamiliar language as native listeners were paying, or there could be a variability in typing speed of participants among both groups. However, it is difficult to disentangle other cognitive, perceptual and motor processing that could confound linguistic processing in the tasks used in this thesis. We recommend to future researchers designing behavioural tasks that could shed more lights on how multiple cognitive processes influence each other. Nonetheless, this thesis can act as a reference for future cross-linguistic research that intends to explore comparable objectives using diverse cohorts of participants.

# References


[1] E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. M. Merzenich. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci U S A*, 2001.

[2] S. J. Aiken and T. W. Picton. Human cortical responses to the speech envelope. *Ear Hear*, 29(2):139–57, 2008.

[3] Motoko Akakura. Evaluating the effectiveness of explicit instruction on implicit and explicit l2 knowledge. *Language Teaching Research*, 16(1):9–37, 2011.

[4] Evelien Akker and Anne Cutler. Prosodic cues to semantic structure in native and nonnative listening. *Bilingualism: Language and Cognition*, 6:81–96, 2003.

[5] L. H. Arnal and A. L. Giraud. Cortical oscillations and sensory predictions. *Trends Cogn Sci*, 16(7):390–8, 2012.

[6] Luc H. Arnal, Keith B. Doelling, and David Poeppel. Delta–beta coupled oscillations underlie temporal prediction accuracy. *Cerebral Cortex*, 25(9):3077–3085, 2014.

[7] Luc H. Arnal, Valentin Wyart, and Anne-Lise Giraud. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, 14(6):797–801, 2011.

[8] Lucas S. Baltzell, Ramesh Srinivasan, and Virginia M. Richards. The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. *Journal of neurophysiology*, 118(6):3144–3151, 2017.

[9] Elisabeth Beyersmann, Deirdre Bolger, Chotiga Pattamadilok, Boris New, Jonathan Grainger, and Johannes C. Ziegler. Morphological processing without semantics: An erp study with spoken words. *Cortex*, 116:55–73, 2019.

[10] Corinna E. Bonhage, Lars Meyer, Thomas Gruber, Angela D. Friederici, and Jutta L. Mueller. Oscillatory eeg dynamics underlying automatic chunking during sentence processing. *NeuroImage*, 152:647–657, 2017.

[11] Victor J. Boucher, Annie C. Gilbert, and Boutheina Jemel. The role of low-frequency neural oscillations in speech processing: Revisiting delta entrainment. *Journal of Cognitive Neuroscience*, 31(8):1205–1215, 2019.

[12] D. H. Brainard. The psychophysics toolbox. *Spatial Vision*, 10(4):433–436, 1997.

[13] D. H. Brainard. The psychophysics toolbox. *Spat Vis*, 10(4):433–6, 1997.

[14] A. Brandmeyer, J. D. Farquhar, J. M. McQueen, and P. W. Desain. Decoding speech perception by native and non-native speakers using single-trial electrophysiological data. *PLoS One*, 8(7):e68261, 2013.

[15] M. R. Brent. Speech segmentation and word discovery: a computational perspective. *Trends Cogn Sci*, 3(8):294–301, 1999.

[16] M. P. Broderick, A. J. Anderson, and E. C. Lalor. Semantic context enhances the early auditory encoding of natural speech. *J Neurosci*, 39(38):7564–7575, 2019.

[17] Marco Buiatti, Marcela Peña, and Ghislaine Dehaene-Lambertz. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *NeuroImage*, 44(2):509–519, 2009.

[18] N. A. Busch, J. Dubois, and R. VanRullen. The phase of ongoing eeg oscillations predicts visual perception. *J Neurosci*, 29(24):7869–76, 2009.

[19] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5):365–376, 2013.

[20] Chandramouli Chandrasekaran, Andrea Trubanova, Sébastien Stillittano, Alice Caplier, and Asif A. Ghazanfar. The natural statistics of audiovisual speech. *PLoS computational biology*, 5(7):e1000436–e1000436, 2009.

[21] R. A. Cole, J. Jakimik, and W. E. Cooper. Segmenting speech into words. *J Acoust Soc Am*, 67(4):1323–32, 1980.

[22] M. J. Crosse, J. S. Butler, and E. C. Lalor. Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J Neurosci*, 35(42):14195–204, 2015.

[23] Michael J. Crosse, Giovanni M. Di Liberto, Adam Bednar, and Edmund C. Lalor. The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, 10:604, 2016.

[24] A. Cutler, D. Dahan, and W. van Donselaar. Prosody in the comprehension of spoken language: a literature review. *Lang Speech*, 40 ( Pt 2):141–201, 1997.

[25] A. Cutler, A. Weber, R. Smits, and N. Cooper. Patterns of english phoneme confusions by native and non-native listeners. *J Acoust Soc Am*, 116(6):3668–78, 2004.

[26] M. H. Davis, M. A. Ford, F. Kherif, and I. S. Johnsrude. Does semantic context benefit speech understanding through "top-down" processes? evidence from time-resolved sparse fmri. *J Cogn Neurosci*, 23(12):3914–32, 2011.

[27] M. H. Davis, I. S. Johnsrude, A. Hervais-Adelman, K. Taylor, and C. McGettigan. Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *J Exp Psychol Gen*, 134(2):222–41, 2005.

[28] A. Delorme and S. Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *J Neurosci Methods*, 134(1):9–21, 2004.

[29] G. M. Di Liberto, M. J. Crosse, and E. C. Lalor. Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech. *eNeuro*, 5(2), 2018.

[30] G. M. Di Liberto and E. C. Lalor. Isolating neural indices of continuous speech processing at the phonetic level. *Adv Exp Med Biol*, 894:337–45, 2016.

[31] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr Biol*, 25(19):2457–65, 2015.

[32] Giovanni M. Di Liberto, Edmund C. Lalor, and Rebecca E. Millman. Causal cortical dynamics of a predictive enhancement of speech intelligibility. *NeuroImage*, 166:247–258, 2018.

[33] N. Ding, M. Chatterjee, and J. Z. Simon. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage*, 88:41–6, 2014.

[34] N. Ding, L. Melloni, H. Zhang, X. Tian, and D. Poeppel. Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci*, 19(1):158–64, 2016.

[35] N. Ding and J. Z. Simon. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc Natl Acad Sci U S A*, 109(29):11854–9, 2012.

[36] N. Ding and J. Z. Simon. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J Neurophysiol*, 107(1):78–89, 2012.

[37] N. Ding and J. Z. Simon. Cortical entrainment to continuous speech: functional roles and interpretations. *Front Hum Neurosci*, 8:311, 2014.

[38] K. B. Doelling, L. H. Arnal, O. Ghitza, and D. Poeppel. Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage*, 85 Pt 2:761–8, 2014.

[39] Sophie Dufour and Jonathan Grainger. Phoneme-order encoding during spoken word recognition: A priming investigation. *Cognitive Science*, 43(10):e12785, 2019.

[40] Erik Edwards and Edward F. Chang. Syllabic (∼2–5 hz) and fluctuation (∼1–10 hz) ranges in speech and auditory processing. *Hearing Research*, 305:113–134, 2013.

[41] Taffeta M. Elliott and Frédéric E. Theunissen. The modulation transfer function for speech intelligibility. *PLOS Computational Biology*, 5(3):e1000302, 2009.

[42] A. K. Engel, P. Fries, and W. Singer. Dynamic predictions: oscillations and synchrony in top-down processing. *Nat Rev Neurosci*, 2(10):704–16, 2001.

[43] Andreas K. Engel and Pascal Fries. Beta-band oscillations—signalling the status quo? *Current Opinion in Neurobiology*, 20(2):156–165, 2010.

[44] Octave Etard and Tobias Reichenbach. Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *The Journal of Neuroscience*, 39(29):5750, 2019.

[45] A. D. Friederici and W. Singer. Grounding language processing on basic neurophysiological principles. *Trends Cogn Sci*, 19(6):329–38, 2015.

[46] P. Fries. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn Sci*, 9(10):474–80, 2005.

[47] K. Friston. A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci*, 360(1456):815–36, 2005.

[48] P. Gagnepain, R. N. Henson, and M. H. Davis. Temporal predictive codes for spoken words in auditory cortex. *Curr Biol*, 22(7):615–21, 2012.

[49] J. Garofolo, Lori Lamel, W. Fisher, Jonathan Fiscus, D. Pallett, N. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*, 11 1993.

[50] Heidi Getz, Nai Ding, Elissa L. Newport, and David Poeppel. Cortical tracking of constituent structure in language acquisition. *Cognition*, 181:135–140, 2018.

[51] Asif A. Ghazanfar, Ryan J. Morrill, and Christoph Kayser. Monkeys are perceptually tuned to facial expressions that exhibit a theta-like speech rhythm. *Proceedings of the National Academy of Sciences*, 110(5):1959, 2013.

[52] O. Ghitza. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Front Psychol*, 2:130, 2011.

[53] O. Ghitza. On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. *Front Psychol*, 3:238, 2012.

[54] O. Ghitza. The theta-syllable: a unit of speech information defined by cortical function. *Front Psychol*, 4:138, 2013.

[55] O. Ghitza and S. Greenberg. 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):113–26, 2009.

[56] A. L. Giraud and D. Poeppel. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci*, 15(4):511–7, 2012.

[57] Manuel Gomez-Ramirez, Simon P. Kelly, Sophie Molholm, Pejman Sehatpour, Theodore H. Schwartz, and John J. Foxe. Oscillatory sensory selection mechanisms during intersensory attention to rhythmic auditory and visual inputs: A human electrocorticographic investigation. *The Journal of Neuroscience*, 31(50):18556, 2011.

[58] Steven Greenberg, Hannah Carvey, Leah Hitchcock, and Shuangyu Chang. Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, 31(3):465–485, 2003.

[59] Joachim Gross, Nienke Hoogenboom, Gregor Thut, Philippe Schyns, Stefano Panzeri, Pascal Belin, and Simon Garrod. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLOS Biology*, 11(12):e1001752, 2014.

[60] D. A. Hambrook, S. Soni, and M. S. Tata. The effects of periodic interruptions on cortical entrainment to speech. *Neuropsychologia*, 121:58–68, 2018.

[61] D. A. Hambrook and M. S. Tata. Theta-band phase tracking in the two-talker problem. *Brain Lang*, 135:52–6, 2014.

[62] D. A. Hambrook and M. S. Tata. The effects of distractor set-size on neural tracking of attended speech. *Brain Lang*, 190:1–9, 2019.

[63] D. A. Hambrook and M.S. Tata. Cortical entrainment to speech occurs without broadband envelope dynamics. 2018.

[64] Joshua K. Hartshorne, Joshua B. Tenenbaum, and Steven Pinker. A critical period for second language acquisition: Evidence from 2/3 million english speakers. *Cognition*, 177:263–277, 2018.

[65] I. Hertrich, S. Dietrich, J. Trouvain, A. Moos, and H. Ackermann. Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. *Psychophysiology*, 49(3):322–34, 2012.

[66] T. Houtgast and H. J. M. Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, 77(3):1069–1077, 1985.

[67] M. F. Howard and D. Poeppel. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J Neurophysiol*, 104(5):2500–11, 2010.

[68] Ailsa Humphries, Zhe Chen, and Ewald Neumann. Comparing repetition priming effects in words and arithmetic equations: Robust priming regardless of color or response hand change. *Frontiers in Psychology*, 8(2326), 2018.

[69] N. Ille, P. Berg, and M. Scherg. Artifact correction of the ongoing eeg using spatial filters based on artifact and brain signal topographies. *J Clin Neurophysiol*, 19(2):113–24, 2002.

[70] Tania Ionin and Silvina Montrul. The role of l1 transfer in the interpretation of articles with definite plurals in l2 english. *Language Learning*, 60(4):877–925, 2010.

[71] Jacqueline S Johnson and Elissa L Newport. Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive Psychology*, 21(1):60–99, 1989.

[72] A. Keitel, J. Gross, and C. Kayser. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biol*, 16(3):e2004473, 2018.

[73] Anne Keitel, Robin A. A. Ince, Joachim Gross, and Christoph Kayser. Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *NeuroImage*, 147:32–42, 2017.

[74] Jess R. Kerlin, Antoine J. Shahin, and Lee M. Miller. Attentional gain control of ongoing cortical speech representations in a "cocktail party". *The Journal of Neuroscience*, 30(2):620, 2010.

[75] J. M. Kilner. Bias in a common eeg and meg statistical analysis and how to avoid it. *Clinical Neurophysiology*, 124(10):2062–2063, 2013.

[76] Karl H. S. Kim, Norman R. Relkin, Kyoung-Min Lee, and Joy Hirsch. Distinct cortical areas associated with native and second languages. *Nature*, 388(6638):171–174, 1997.

[77] N. Kriegeskorte, W. K. Simmons, P. S. Bellgowan, and C. I. Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*, 12(5):535–40, 2009.

[78] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1 – 26, 2017.

[79] Anne Kösem, Anahita Basirat, Leila Azizi, and Virginie van Wassenhove. High-frequency neural activity predicts word parsing in ambiguous speech streams. *Journal of neurophysiology*, 116(6):2497–2512, 2016.

[80] Anne Kösem and Virginie van Wassenhove. Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. *Language, Cognition and Neuroscience*, 32(5):536–544, 2017.

[81] N. M. Laird. Missing data in longitudinal studies. *Stat Med*, 7(1-2):305–15, 1988.

[82] E. C. Lalor and J. J. Foxe. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *Eur J Neurosci*, 31(1):189–93, 2010.

[83] E. C. Lalor, A. J. Power, R. B. Reilly, and J. J. Foxe. Resolving precise temporal processing properties of the auditory system using continuous stimuli. *J Neurophysiol*, 102(1):349–59, 2009.

[84] Michael J. Larson and Kaylie A. Carbine. Sample size calculations in human electrophysiology (eeg and erp) studies: A systematic review and recommendations for increased rigor. *International Journal of Psychophysiology*, 111:33–41, 2017.

[85] Juyeon Lee, Soyoung Park, and Michael Heinz. Exploring patterns of article use by advanced korean learners of english and spanish. *International Review of Applied Linguistics in Language Teaching*, 56(1):79–101, 2018.

[86] D. Lesenfants and T. Francart. The interplay of top-down focal attention and the cortical tracking of speech. *Scientific Reports*, 10(1):6922, 2020.

[87] Ashley G. Lewis and Marcel Bastiaansen. A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, 68:155–168, 2015.

[88] Ashley G. Lewis, Jan-Mathijs Schoffelen, Herbert Schriefers, and Marcel Bastiaansen. A predictive coding perspective on beta oscillations during sentence-level language comprehension. *Frontiers in Human Neuroscience*, 10(85), 2016.

[89] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[90] H. Luo and D. Poeppel. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6):1001–10, 2007.

[91] H. Luo and D. Poeppel. Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex. *Front Psychol*, 3:170, 2012.

[92] Guangting Mai, James W. Minett, and William S. Y. Wang. Delta, theta, beta, and gamma brain oscillations index levels of auditory sentence processing. *NeuroImage*, 133:516–528, 2016.

[93] Eric Maris and Robert Oostenveld. Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190, 2007.

[94] William D. Marslen-Wilson and Alan Welsh. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10(1):29–63, 1978.

[95] Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F. Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

[96] L. Meyer, M. J. Henry, P. Gaston, N. Schmuck, and A. D. Friederici. Linguistic bias modulates interpretation of speech via neural delta-band oscillations. *Cereb Cortex*, 27(9):4293–4302, 2017.

[97] Lars Meyer, Yue Sun, and Andrea E. Martin. Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, 35(9):1089–1099, 2020.

[98] Rebecca E. Millman, Sam R. Johnson, and Garreth Prendergast. The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. *Journal of Cognitive Neuroscience*, 27(3):533–545, 2015.

[99] N. Molinaro and M. Lizarazu. Delta(but not theta)-band cortical entrainment involves speech-specific processing. *Eur J Neurosci*, 48(7):2642–2650, 2018.

[100] Benjamin Morillon, Catherine Liegeois-Chauvel, Luc Arnal, Christian Bénar, and Anne-Lise Giraud. Asymmetric function of theta and gamma activity in syllable processing: An intra-cortical study. *Frontiers in Psychology*, 3(248), 2012.

[101] Anne B. Newman. An overview of the design, implementation, and analyses of longitudinal studies on aging. *Journal of the American Geriatrics Society*, 58 Suppl 2(Suppl 2):S287–S291, 2010.

[102] Kirill V. Nourski, Richard A. Reale, Hiroyuki Oya, Hiroto Kawasaki, Christopher K. Kovach, Haiming Chen, Matthew A. Howard, and John F. Brugge. Temporal envelope of time-compressed speech represented in the human auditory cortex. *The Journal of Neuroscience*, 29(49):15564–15574, 2009.

[103] J. Obleser and C. Kayser. Neural entrainment and attentional selection in the listening brain. *Trends Cogn Sci*, 23(11):913–926, 2019.

[104] Jonas Obleser, Molly J. Henry, and Peter Lakatos. What do we talk about when we talk about rhythm? *PLOS Biology*, 15(9):e2002794, 2017.

[105] K. Okano, J. Grainger, and P. J. Holcomb. An erp investigation of visual word recognition in syllabary scripts. *Cogn Affect Behav Neurosci*, 13(2):390–404, 2013.

[106] Eleni Orfanidou, William D. Marslen-Wilson, and Matthew H. Davis. Neural response suppression predicts repetition priming of spoken words and pseudowords. *Journal of Cognitive Neuroscience*, 18(8):1237–1252, 2006.

[107] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cereb Cortex*, 25(7):1697–706, 2015.

[108] C. Pallier, A. D. Devauchelle, and S. Dehaene. Cortical representation of the constituent structure of sentences. *Proc Natl Acad Sci U S A*, 108(6):2522–7, 2011.

[109] Brian N. Pasley, Stephen V. David, Nima Mesgarani, Adeen Flinker, Shihab A. Shamma, Nathan E. Crone, Robert T. Knight, and Edward F. Chang. Reconstructing speech from human auditory cortex. *PLOS Biology*, 10(1):e1001251, 2012.

[110] J. E. Peelle, J. Gross, and M. H. Davis. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex*, 23(6):1378–87, 2013.

[111] Maria Pefkou, Luc H. Arnal, Lorenzo Fontolan, and Anne-Lise Giraud. Theta- and beta-band neural activity reflect independent syllable tracking and comprehension of time-compressed speech. *The Journal of Neuroscience*, pages 2882–16, 2017.

[112] Marcela Peña and Lucia Melloni. Brain oscillations during spoken sentence processing. *Journal of Cognitive Neuroscience*, 24(5):1149–1164, 2012.

[113] Alejandro Pérez, Manuel Carreiras, Margaret Gillon Dowens, and Jon Andoni Duñabeitia. Differential oscillatory encoding of foreign speech. *Brain and Language*, 147:51–57, 2015.

[114] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[115] J. M. Rimmele, B. Morillon, D. Poeppel, and L. H. Arnal. Proactive sensing of periodic and aperiodic auditory patterns. *Trends Cogn Sci*, 22(10):870–882, 2018.

[116] Suzanne Romaine. *Variation*, pages 410–435. Blackwell Publishing Ltd, eds. edition, 2003.

[117] DONALD B. RUBIN. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[118] Lisa D. Sanders and Helen J. Neville. An erp study of continuous speech processing: I. segmentation, semantics, and syntax in native speakers. *Cognitive Brain Research*, 15(3):228–240, 2003.

[119] Lisa D. Sanders and Helen J. Neville. An erp study of continuous speech processing: Ii. segmentation, semantics, and syntax in non-native speakers. *Cognitive Brain Research*, 15(3):214–227, 2003.

[120] Lisa D. Sanders, Helen J. Neville, and Marty G. Woldorff. Speech segmentation by native and non-native speakers: the use of lexical, syntactic, and stress-pattern cues. *Journal of speech, language, and hearing research : JSLHR*, 45(3):519–530, 2002.

[121] Lisa D. Sanders, Elissa L. Newport, and Helen J. Neville. Segmenting nonsense: an event-related potential index of perceived onsets in continuous speech. *Nature Neuroscience*, 5(7):700–703, 2002.

[122] Jona Sassenhagen and Dejan Draschkow. Cluster-based permutation tests of meg/eeg data do not establish significance of effect latency or location. *Psychophysiology*, 56(6):e13335, 2019.

[123] C. E. Schroeder and P. Lakatos. Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci*, 32(1):9–18, 2009.

[124] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–4, 1995.

[125] E. Sohoglu, J. E. Peelle, R. P. Carlyon, and M. H. Davis. Predictive top-down integration of prior knowledge during speech perception. *J Neurosci*, 32(25):8443–53, 2012.

[126] S. Soni and M. S. Tata. Brain electrical dynamics in speech segmentation depends upon prior experience with the language. *Brain Lang*, 219:104967, 2021.

[127] Mitchell Steinschneider, Kirill V. Nourski, and Yonatan I. Fishman. Representation of speech in human auditory cortex: is it special? *Hearing research*, 305:57–73, 2013.

[128] Annie Tremblay, Jui Namjoshi, Elsa Spinelli, Mirjam Broersma, Taehong Cho, Sahyang Kim, Maria Teresa Martínez-García, and Katrina Connell. Experience with a second language affects the use of fundamental frequency in speech segmentation. *PLOS ONE*, 12(7):e0181709, 2017.

[129] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart. Speech intelligibility predicted from neural entrainment of the speech envelope. *J Assoc Res Otolaryngol*, 19(2):181–191, 2018.

[130] Y. Wang, N. Ding, N. Ahmar, J. Xiang, D. Poeppel, and J. Z. Simon. Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: Meg evidence. *J Neurophysiol*, 107(8):2033–41, 2012.

[131] Geoffrey F. Woodman. A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics*, 72(8):2031–2046, 2010.

[132] S. Ylinen, M. Huuskonen, K. Mikkola, E. Saure, T. Sinkkonen, and P. Paavilainen. Predictive coding of phonological rules in auditory cortex: A mismatch negativity study. *Brain Lang*, 162:72–80, 2016.

[133] Ling Zhong, Chang Liu, and Sha Tao. Sentence recognition for native and non-native english listeners in quiet and babble: Effects of contextual cues. *The Journal of the Acoustical Society of America*, 145(4):EL297–EL302, 2019.

[134] E. Zion Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *J Neurosci*, 33(4):1417–26, 2013.

[135] E. M. Zion Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon, D. Poeppel, and C. E. Schroeder. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, 77(5):980–91, 2013.

[136] B. Zoefel and R. VanRullen. Eeg oscillations entrain their phase to high-level features of speech sound. *Neuroimage*, 124(Pt A):16–23, 2016.

[137] Jiajie Zou, Jun Feng, Tianyong Xu, Peiqing Jin, Cheng Luo, Jianfeng Zhang, Xunyi Pan, Feiyan Chen, Jing Zheng, and Nai Ding. Auditory and language contributions to neural encoding of speech features in noisy environments. *NeuroImage*, 192:66–75, 2019.

# Appendix A

# Appendix: Supplementary figures

## A.1 Figure Supplement

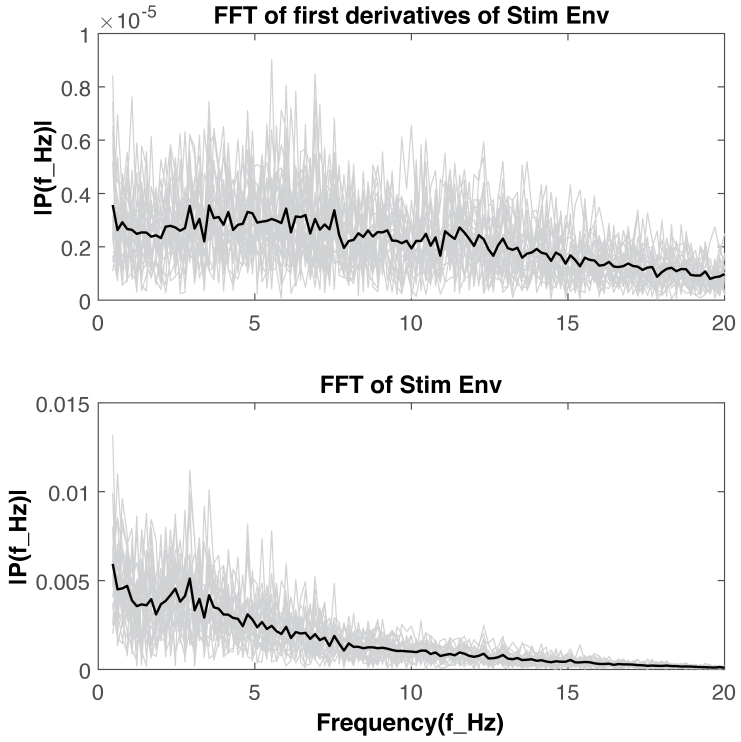


Figure A.1: **Periodograms of first derivatives of stimulus envelopes (top) and stimulus envelopes (bottom)**. Gray color lines represent periodogram of individual stimulus and black line represents the mean periodogram of all stimuli.

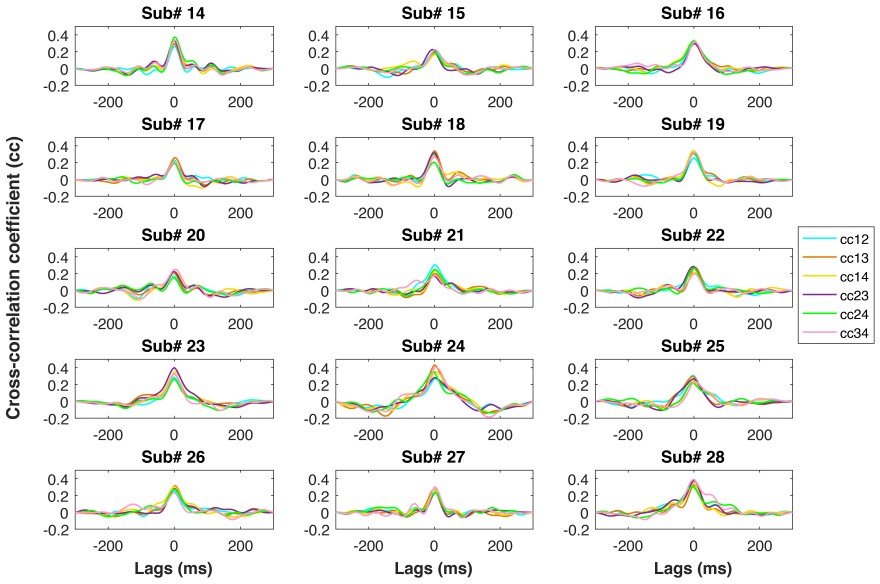


Figure A.2: **Within-subject phase similarity across presentations in native English speakers.** Brain responses to word boundaries were cross-correlated across different pairs of presentations.

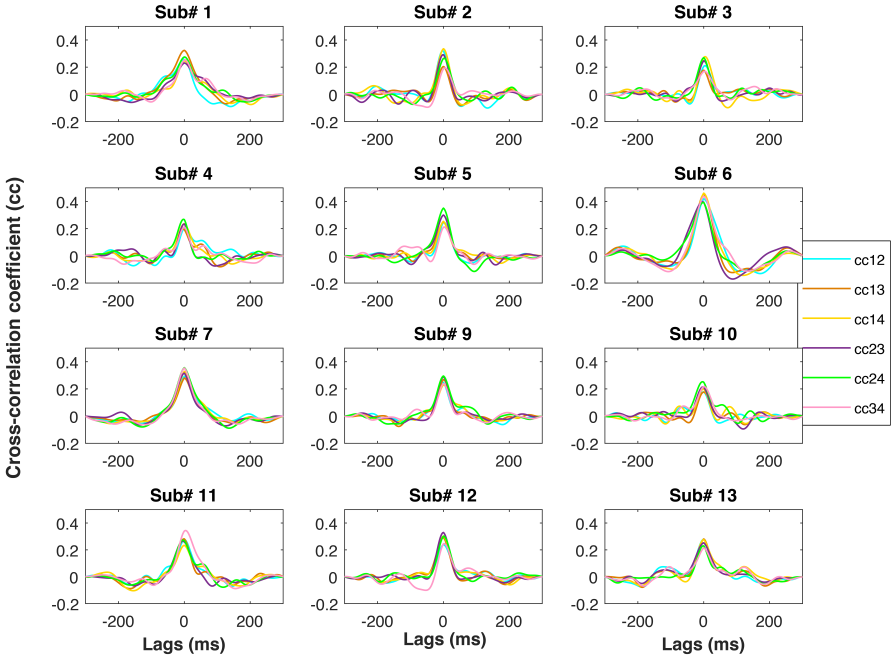


Figure A.3: **Within-subject phase similarity across presentations in non-native English speakers.** Brain responses to word boundaries were cross-correlated across different pairs of presentations.

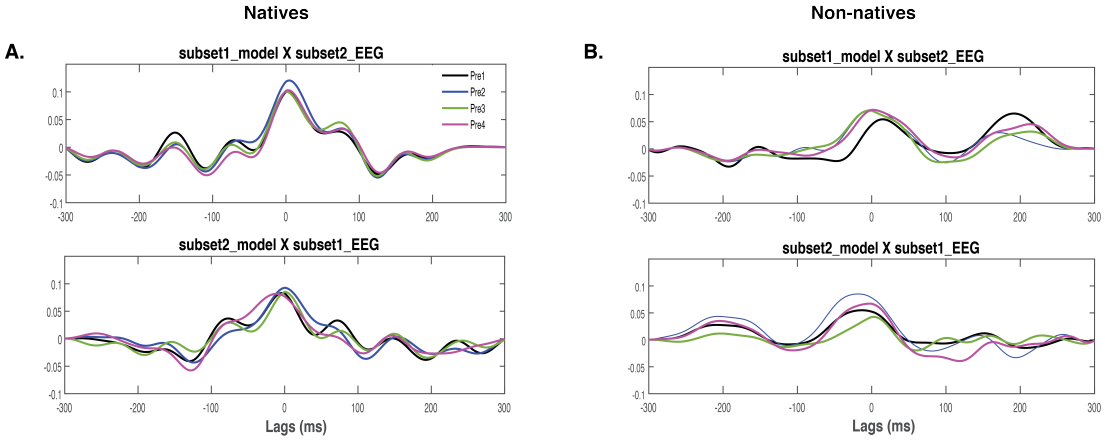


Figure A.4: **Split-group analysis for (A) the native group and (B) the non-native group.** In this analysis, the group was split into two subsets and the model was calculated by averaging responses to word boundaries from subjects in one subset. Then, the model was cross-correlated to each subject's EEG in the other subset.

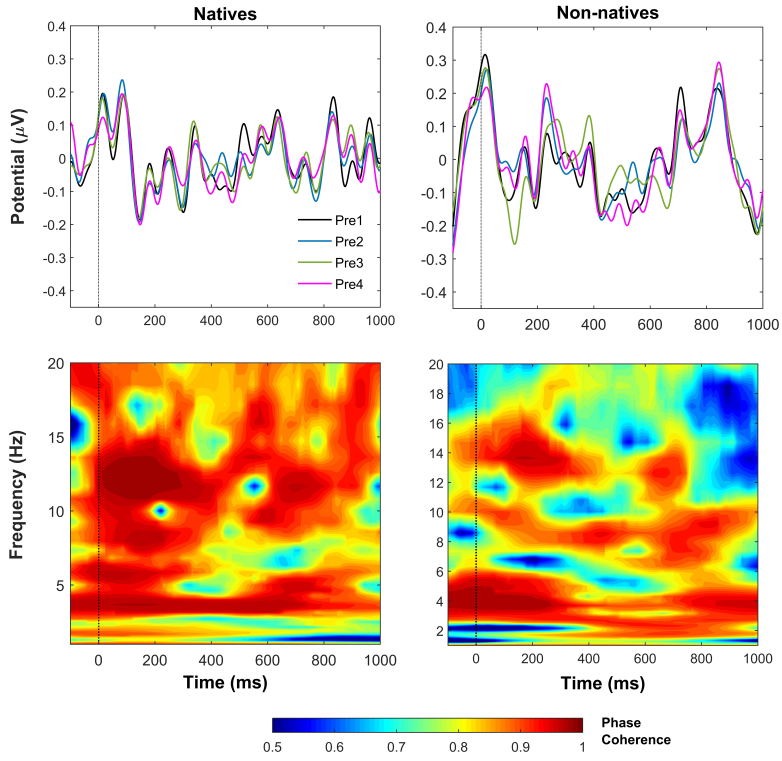


Figure A.5: **Sentence-level dynamics (top) and phase coherence across all presentations (bottom) in native (left) and non-native (right) speakers.**

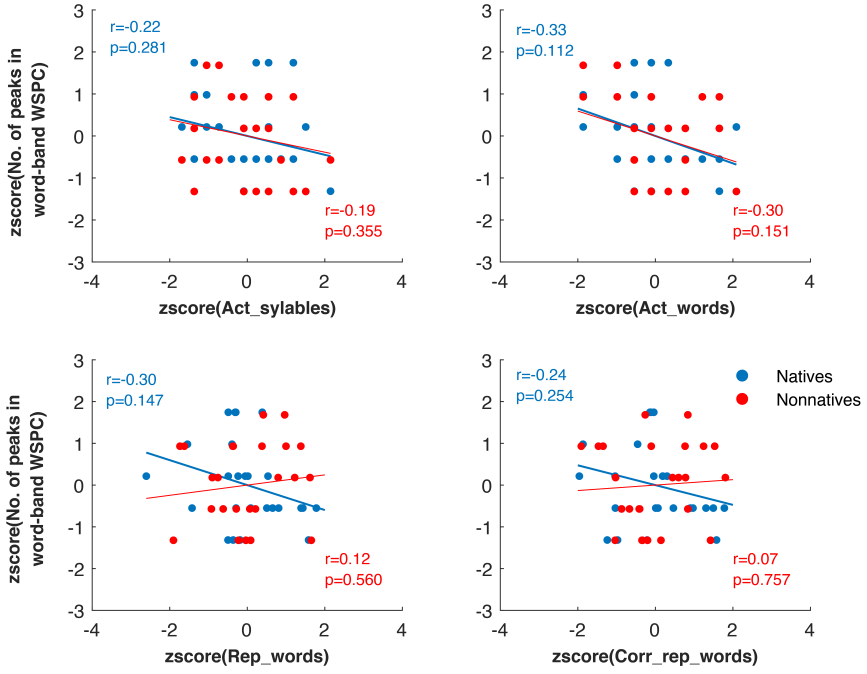


Figure A.6: **The correlation between the number of actual syllables, actual words, reported words, and correctly reported words and the number of peaks in within-sentence phase coherence at word rate (2–5 Hz) in both native (blue) and non-native (red) speakers.**

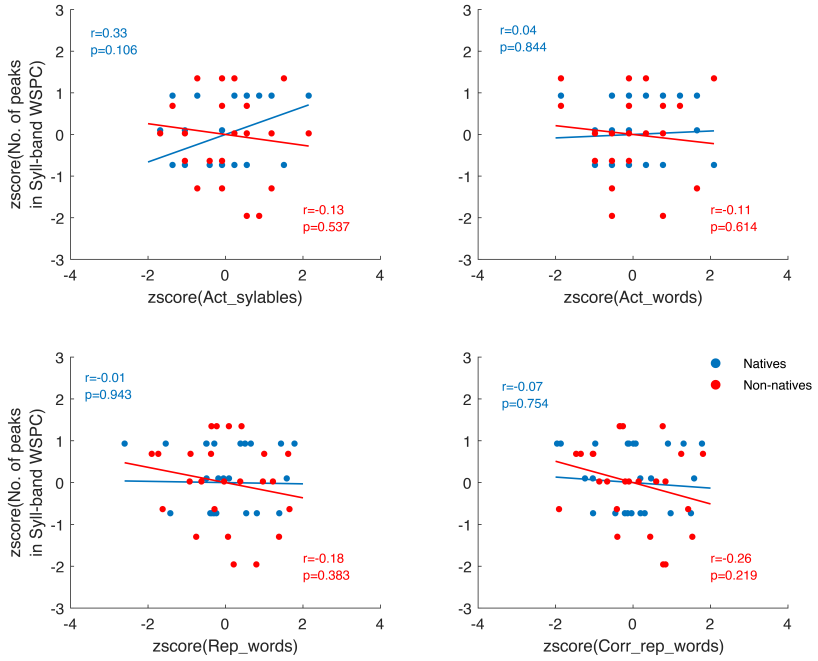


Figure A.7: **The correlation between the number of actual syllables, actual words, reported words, and correctly reported words and the number of peaks in within-sentence phase coherence at syllabic rate (3.2–4.9 Hz) in both native (blue) and non-native (red) speakers.**