2018

Stimulus and cognitive factors in cortical entrainment to speech

Hambrook, Dillon A.

Lethbridge, Alta. : Universtiy of Lethbridge, Department of Neuroscience

STIMULUS AND COGNITIVE FACTORS IN CORTICAL ENTRAINMENT TO SPEECH

DILLON A. HAMBROOK
Master of Science, University of Lethbridge, 2014

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfilment of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

Department of Neuroscience
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

STIMULUS AND COGNITVE FACTORS IN CORTICAL ENTRAINMENT TO SPEECH

DILLON A. HAMBROOK

Date of defence: August 21, 2018

Dr. Matthew S. Tata Supervisor	Associate Professor	Ph.D.
Dr. Artur Luczak Thesis Examination Committee Member	Professor	Ph.D.
Dr. David Euston Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. Georg Boenn Thesis Examination Committee Member	Assistant Professor	Ph.D.
Dr. Kyle Mathewson External Examiner University of Alberta Edmonton, AB	Assistant Professor	Ph.D.
Dr. Claudia Gonzalez Chair, Thesis Examination Committee	Associate Professor	Ph.D.

Abstract

Understanding speech is a difficult computational problem yet the human brain does it with ease. Entrainment of oscillatory neural activity to acoustic features of speech is an example of dynamic coupling between cortical activity and sensory inputs. The phenomenon may be a bottom-up, sensory-driven neurophysiological mechanism that supports speech processing. However, cognitive top-down factors such as linguistic knowledge and attentional focus affect speech perception, especially in challenging real-world environments. It is unclear how these top-down influences affect cortical entrainment to speech. We used electroencephalography to measure cortical entrainment to speech under conditions of acoustic and cognitive interference. By manipulating the bottom-up, sensory features in the acoustic scene we found evidence of top-down influences of attentional selection and linguistic processing on speech-entrained activity.

Table of Contents

List of Figures

List of Abbreviations

ANOVA – Analysis of variance
DISS – Global topographic dissimilarity
ECoG – Electrocorticography
EEG – Electroencephalography
FDR – False discovery rate
HG – Heshcl's gyrus
MEG – Magnetoencephalography
RII – Ratio of intrusion to insertion errors
RMS – Root-mean square
(m)TRF – (multivariate) Temporal response function

1 Introduction

Understanding speech is a difficult computational feat, yet the human brain possesses an uncanny ability to extract meaning from the complex acoustic signal that makes up spoken language. Even more remarkable is that the ability to comprehend speech is surprisingly robust: people can pick out one voice from among many and understand what is being said and can understand speech even despite other loud noises in the environment. The faculty for understanding speech is so predominant that people perceive and understand speech that has been interrupted by silences – an experience anyone with poor mobile phone reception can attest to. Even if the loss of signal leads to the complete removal of speech information the brain is somehow able to restore the perception of speech and essentially make something coherent out of literally nothing. In this thesis, we explore the neural mechanisms that allow the brain to understand speech in challenging situations: In Chapter 2 we consider the role that alignment between temporal modulations in the physical speech signal and oscillatory electrical activity in the brain may play in maintaining attention to one talker among many others. In Chapter 3 we examine the role cortical entrainment to speech plays in segregating behaviorally relevant speech from background noise. In Chapter 4 we explore how speech tracking may influence neural mechanisms responsible for repairing the percept of interrupted speech.

1.1 What is speech to a brain?

From a physical standpoint speech is a dynamic, complex acoustic signal. Speech has a complex spectrum; the speech signal consists of energy at a number of frequencies

and the distribution between frequencies changes from moment to moment. Somehow, the brain can process slight differences in the spectrotemporal properties of the speech signal in order to extract meaningful information.

Spoken language can be broken down along multiple hierarchically organized levels. Phonemes represent the atomic level of speech sounds in that they represent the smallest level at which different sound patterns can change the semantic meaning of a word or utterance. The syllable, consisting of a single vowel phoneme with or without surrounding consonant phonemes, represents the next unit of speech sound. Words are formed by one or more syllable-units, and utterances are made up of one or more words. These hierarchical levels correspond to how speech is synthesized into language in the mind of the listener. Thus, speech consists of a series of phonemes, organized into syllables, which are organized into words, which make up an utterance, which carries some sort of meaningful message.

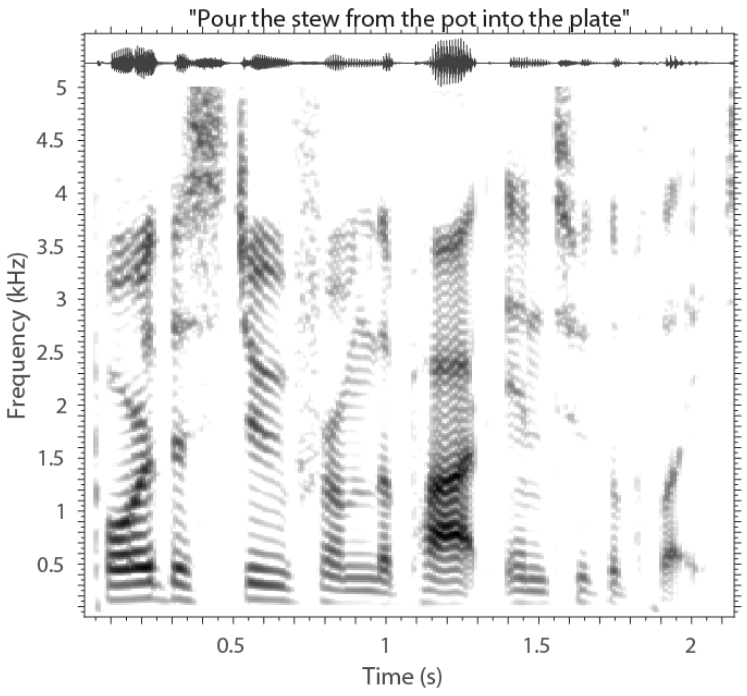

Figure 1.1: Speech waveform (top) and spectrogram for the sentence "pour the stew from the pot into the plate." The acoustic envelope reflects the energy dynamics of the speech signal.

The basic computational problem of understanding speech is fundamentally one of segmentation. The acoustic signal that arrives at the ear does not come with instructions or obvious markers of the boundaries between phonemes, syllables, words, or even utterances. Consider the utterance, "pour the stew from the pot into the plate," (Figure 1.1) shown as a waveform - the acoustic signal generated from a speaker - and as a spectrogram which is analogous to the signal as it is broken down at the cochlea in the early auditory system of the brain. There are not obvious and consistent gaps between each phoneme. So how then, does the brain solve this segmentation problem?

An important feature of speech is that it possesses some degree of temporal regularity. If we re-examine the spectrogram in Figure 1.1 we can see distinct and regular bursts of energy occurring at a rate of roughly 5 Hz. These low-frequency fluctuations in broadband energy are the acoustic or temporal envelope of speech. Acoustic envelope modulations at a rate of between 3-7 Hz seem to be a general feature of speech, corresponding to the syllabic rate across languages (Pellegrino, Coupé, & Marsico, 2011) and was potentially preceded by the development of communicative facial gestures in non-human primates (Ghazanfar, Morrill, & Kayser, 2013) since it also closely matches human mouth movement rates while speaking (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Speech also contains modulations at lower frequencies (1-2 Hz) which correspond to prosodic contours and higher frequencies (30-50 Hz) corresponding to phonemes which typically last 20-40 ms in duration (Arnal & Giraud, 2012; Ghitza & Greenberg, 2009; Poeppel, 2003).

Recent electrophysiological results have sparked increased interest in what kind of potential computational role the acoustic envelope may play in facilitating speech comprehension. Cortical entrainment of auditory cortical activity to the temporal envelope of speech was first demonstrated in magnetoencephalography (MEG) (Ahissar et al., 2001; Luo & Poeppel, 2007), and subsequently observed in electroencephalography (EEG) (Aiken & Picton, 2008) and electrocorticography (ECoG) (Nourski et al., 2009). These studies found that the phase of oscillatory neural activity, at the same modulation rate as the envelope, tracked the acoustic envelope of speech. Later studies also found that, in addition to phase entrainment of low-frequency (<8 Hz) activity, modulations in

power at higher frequencies in the gamma-band range (60-170 Hz) were also related to modulations in the speech envelope (Morillon, Liégeois-Chauvel, Arnal, Bénar, & Giraud, 2012; Pasley et al., 2012). The underlying neural mechanisms and functional roles of the speech envelope tracking response remains controversial. Neural entrainment to the acoustic envelope has been observed for non-speech sounds and unintelligible speech (Lalor, Power, Reilly, & Foxe, 2009; Luo & Poeppel, 2007; Millman, Prendergast, Hymers, & Green, 2013; Steinschneider, Nourski, & Fishman, 2013; Y. Wang et al., 2012) which suggests that envelope tracking is a general bottom-up stimulus-driven response. However, a number of other studies have found that the envelope tracking of speech is modulated by top-down cognitive functions such as attention and intelligibility (Ding & Simon, 2012a; Hambrook & Tata, 2014; Kerlin, Shahin, & Miller, 2010; Mesgarani & Chang, 2012; Peelle & Davis, 2012; Peelle, Gross, & Davis, 2013; Zion Golumbic, Ding, et al., 2013).

1.2 The potential functions of cortical speech tracking

In the following section we will discuss possible functional roles of neural entrainment to the acoustic envelope of speech. There are a number of hypothesized cognitive and computational functions that speech tracking may fulfill. Some functions (syllabic parsing) are definitively speech specific, while others (attentional selection, auditory scene analysis) are more general and may reflect adaptations to a broad range of pseudo-rhythmic acoustic stimuli. The hypothesized functions of cortical entrainment to the acoustic envelope share common proposed mechanisms based on two important observations: First, neural excitability is modulated by oscillatory phase (Volgushev,

Chistiakova, & Singer, 1998). Second, the momentary strength of connections between neural ensembles is modulated by their relative phase relationship at any given moment (Fries, 2005, 2015). Taken together these observations describe a potential mechanism by which entrainment to an external rhythm (e.g. the speech envelope) can determine the sensitivity of auditory areas and *ad hoc* cortical networks that support speech processing in various ways.

1.2.1 Syllabic parsing

The correspondence between the syllabic rate, which is roughly 5 Hz, and theta-band (4-8 Hz) oscillatory activity entrained to the acoustic envelope has led to the suggestion that entrainment may reflect an active parsing mechanism that is responsible for segmenting the continuous acoustic signal into syllabic and phonemic units. The boundaries between syllables are relatively well encoded by the acoustic envelope of speech (Ghitza, 2013; Greenberg, 1996; Stevens, 2002). It has been hypothesized that by entraining oscillatory activity to the syllable rate as it is encoded by the envelope, the brain creates "windows" of enhanced sensitivity in order to optimally process the spectrotemporal features that distinguish phonemes (Giraud & Poeppel, 2012). The TEMPO model (Ghitza, 2011) describes a more formal connection between the envelope and oscillatory activity: Theta oscillations entrain to the envelope acting both as a master-clock in the oscillator array and modulating the beta and gamma oscillations (at frequencies 4x and 10x the theta frequency respectively) which correspond to dyadic groupings of phonemes and the rapid spectrotemporal modulations within phonemes. This system of cascaded oscillators parses the acoustic stream into linguistic "chunks".

These “chucks” are decoded by template-matching the syllable and phoneme level information chunks to internal models of the temporal and spectrotemporal features of syllables and phonemes respectively. According to this hypothesis envelope entrainment reflects actively segmenting and decoding speech. The hypothesized role of speech envelope tracking is supported by behavioral studies that found that speech intelligibility (Ghitza & Greenberg, 2009) and envelope tracking (Kayser, Ince, Gross, & Kayser, 2015) is reduced in response to irregular speech rates produced by manipulating the length of pauses between syllables or words.

1.2.2 Auditory scene analysis

Adding complexity to the problem of understanding speech is the fact that we rarely hear a single voice in clear detail and isolated from other competing sounds. Therefore, the neural mechanisms for understanding speech must contain or interact with mechanisms for isolating the target speech from the acoustic mixture. Isolating one sound or set of sounds from a mixture is commonly referred to as auditory scene analysis and functions by grouping sounds into “streams” based on features including frequency, pitch, timbre, timing, location, and applied contextual cues (Bregman, 1990). Traditional neurological hypotheses of auditory scene analysis maintain that sound segregation is achieved by differential responses in spatially well-separated auditory neuron populations tuned to the acoustic features that support the formation of distinct streams (Bee & Klump, 2005; Fishman, Arezzo, & Steinschneider, 2004; Fishman, Reser, Arezzo, & Steinschneider, 2001; Micheyl, Tian, Carlyon, & Rauschecker, 2005; Pressnitzer, Sayles, Micheyl, & Winter, 2008). While this theory is convincing for

streaming based on spectral (frequency, pitch, timbre) similarity because there is well defined tonotopy throughout the auditory system, and streaming based on spatial similarity because acoustic space is encoded by topographically sensitive neuron populations in primary auditory cortex (Middlebrooks, Dykes, & Merzenich, 1980; Mrsic-Flogel, King, & Schnupp, 2005), it cannot account for streaming based on the relative timing of sounds; for example, it fails to predict that simultaneously presented tones that are well separated in frequency will be perceived as a single stream (Elhilali, Ma, Micheyl, Oxenham, & Shamma, 2009). Shamma et al. (2011) have suggested that temporal coherence, both between components of an acoustic stream and the activity of neural populations encoding that component, may serve to bind components of a stream together. In this temporal coherence model of scene analysis, selective attention acts both to enhance the representation of salient acoustic features (Fritz, Elhilali, David, & Shamma, 2007) and modulates the timing of responses to maintain coherence among neural ensembles representing the target stream (Elhilali, Xiang, Shamma, & Simon, 2009).

1.2.3 Attentional selection

Maintaining the representation of a single stream within the brain is known as selective auditory attention and it provides a systematic enhancement of the representation of the selected stream within the brain (Fritz et al., 2007; Kaya & Elhilali, 2017). Entrainment of low-frequency oscillatory activity has been suggested as a potential neurophysiological mechanism for enhancing the cortical representation of rhythmic stimulus both between (Lakatos, Karmos, Mehta, Ulbert, & Schroeder, 2008;

Schroeder & Lakatos, 2009) and within (Lakatos et al., 2013; Zion Golumbic, Ding, et al., 2013) stimulus modalities.

A two-talker paradigm (Cherry, 1953), in which two streams of speech are presented simultaneously while listeners are instructed to focus attention on a target stream while ignoring the other, competing stream is commonly used to study attention to speech. The addition of distractor speech to the acoustic scene reduces speech intelligibility in a complex manner depending on several factors related to the two speech signals including their spectral similarity, temporal correlation, and spatial proximity (Bronkhorst, 2015). Many electrophysiological studies using the two-talker paradigm have found that tracking of attended speech streams is more robust than tracking of simultaneously presented unattended speech (Ding, Chatterjee, & Simon, 2014; Ding & Simon, 2012b, 2012a; Hambrook & Tata, 2014; Horton, D'Zmura, & Srinivasan, 2013; Kerlin et al., 2010; Kong, Mullangi, & Ding, 2014; Mesgarani & Chang, 2012; Power, Foxe, Forde, Reilly, & Lalor, 2012; Rimmele, Zion Golumbic, Schröger, & Poeppel, 2015; Zion Golumbic, Ding, et al., 2013). Enhanced tracking of the attended speech stream is associated with enhanced perceptual awareness of the target speech (Hambrook & Tata, 2014; Mesgarani & Chang, 2012).

Some part of this attentional effect may be explained by a general attentional enhancement of auditory features represented in sensory cortex; however, selective speech tracking responses have also been observed in areas outside of sensory cortex but only in response to the attended speech stream (Zion Golumbic, Ding, et al., 2013). This

result suggests that one function of attentional modulation of cortical speech tracking is to provide "temporal binding" between neural ensembles in auditory cortex and higher-order areas responsible for transforming sound into speech in the brain. The theory of *communication through coherence* (Fries, 2005, 2015; Womelsdorf & Everling, 2015) suggests that by linking the phase of oscillatory activity in one brain area to the phase of oscillatory activity in another communication between the two areas becomes more effective and selective. The *selective entrainment* hypothesis (Schroeder & Lakatos, 2009; Zion Golumbic, Cogan, Schroeder, & Poeppel, 2013) proposes that attention phase-locks oscillatory activity in higher-order speech-specific brain areas to oscillatory activity in the auditory cortex, effectively selecting the acoustic signal that is being tracked by the auditory cortex.

1.3 Entrainment to other speech features

While the discussion of neural entrainment to speech has thus far focused on the acoustic envelope as the entraining speech feature, both because it is easily computed and because its modulation rate matches the frequency of easily measurable neural oscillations, there is substantial evidence that suggests entrainment is driven by other acoustic and linguistic features, and not the acoustic envelope *per se*. In one EEG study Obleser et al. (2012) found that comparable phase-tracking occurred for both amplitude modulated complex tones and frequency modulated complex tones which had a constant amplitude (and therefore a flat acoustic envelope). An MEG study by Doelling et al. (2014) used a noise vocoding scheme to generate speech samples containing envelope information based on the broadband acoustic envelope, the acoustic envelope

within discrete frequency bands, an artificial envelope consisting of peaks of uniform height and shape, and an acoustic envelope without modulations between 2-9 Hz. They found that both the intelligibility of the synthesized speech and the neural tracking of speech was most sensitive to manipulation of the acoustic envelope within discrete frequency bands, suggesting that the acoustic speech tracking response actually reflects sensitivity to temporal modulations within frequency bands rather than across all frequencies. This notion is confirmed by experiments that effectively eliminate broadband envelope fluctuations in the acoustic scene by presenting carefully modulated noise concurrent with speech; despite the removal of the broadband acoustic envelope cue there is a robust speech-tracking response when actively (See Chapter 3) and passively (Zoefel & VanRullen, 2016) listening to speech.

Many studies have also reported neural entrainment to linguistic features of speech. Studies have identified cortical entrainment responses that reflect the encoding of phonetic articulatory features in both ECoG (Mesgarani, Cheung, Johnson, & Chang, 2014) and EEG (Di Liberto, O'Sullivan, & Lalor, 2015; Di Liberto, Crosse, & Lalor, 2018; Di Liberto, Lalor, & Millman, 2018). Kayser et al. (2015) found that disrupting the regular rate of speech reduced pre-frontal delta-band activity phase-locked to the speech envelope while the evoked responses to acoustic transients were maintained, suggesting that low-frequency phase-locking responses cannot be explained solely by evoked responses to acoustics; the low-frequency speech tracking response must reflect some degree of neural entrainment of oscillatory activity. A number of interesting results have emerged based on hierarchically constructed isochronous speech synthesis techniques in

which syllables occur isochronously at a frequency of 4 Hz while words, phrases, and sentences constructed from those syllables occur at distinct (lower) frequencies. Crucially, for speech constructed in this manner, the acoustic envelope only provides cues regarding syllable boundaries – tracking of higher-order structures reflects entrainment based on the abstract linguistic connections between syllables and not acoustic features. Using this kind of stimulus while recording ECoG, Ding et al. (2016) found evidence of systematic neural entrainment to higher-order features (i.e. words, phrases, and sentences), as well as entrainment to syllabic features that was not associated with non-speech acoustic stimuli. Importantly, entrainment to higher-order features was dependent on listeners understanding the presented speech: English speakers did not entrain to Chinese words or phrases, and Chinese speakers did not show entrainment to English words or phrases. A subsequent EEG study by Makov et al. (2017) replicated the finding that the intelligibility of speech was crucial to tracking higher-order features and found that these higher-order structures were not tracked in sleeping listeners. Taken together these results suggest that neural entrainment to speech is not limited to theta-band tracking of the acoustic envelope, but rather reflects entrainment to phrasal/prosodic structures as well.

1.4 Intelligibility or acoustics?

The relationship between neural entrainment to speech acoustic features and speech intelligibility has been persistent. The finding that neural entrainment to a speech signal modulates the intelligibility of that speech signal would indicate that entrainment to speech plays a mechanistic role supporting speech comprehension. In fact, a number

of studies have found that delta- and theta-band speech tracking is enhanced for intelligible versus unintelligible speech (Di Liberto, Lalor, et al., 2018; Doelling et al., 2014; Gross et al., 2013; Park, Ince, Schyns, Thut, & Gross, 2015; Peelle et al., 2013), native vs foreign-language speech (Pérez, Carreiras, Gillon Dowens, & Duñabeitia, 2015), and comprehended vs misunderstood speech (Hambrook & Tata, 2014; Mesgarani & Chang, 2012; Steinmetzger & Rosen, 2017). However, several studies have failed to replicate the apparent connection between speech tracking and intelligibility and thus must be reckoned with. Howard and Poeppel (2010) found no differences in speech-locked theta-band activity for normal versus time-reversed speech, despite the time-reversed speech being entirely unintelligible; however, we note that their behavioral task involved matching two consecutively presented speech samples – a task that does not require explicit linguistic processing. Similarly, Pena and Melloni (2012) found low-frequency speech tracking activity did not differ for native versus foreign language speech; yet, we note again that their behavioral task did not require explicit linguistic analysis as it involved matching a brief sample probe, drawn from the pool of speech stimuli, to the previously presented speech sample. Millman et al. (2015) used three brief speech samples rendered unintelligible by processing the speech stimuli using a tone-vocoder with only 3 frequency channels which was rendered "intelligible" through a perceptual training process in which a degraded speech stimulus was presented in sequence with the unprocessed speech stimulus until listeners indicated that they now found the degraded stimulus intelligible. They found no difference between the pre- and post-training speech tracking response, even though post-training the vocoded speech was

rated as intelligible by the listeners. Once again, we note that the purported improvement in intelligibility may be explained by simply perceptually mapping the degraded, vocoded speech to the intact speech rather than a perceptual restoration of the degraded speech itself; the mechanism implied by intelligibility improvement in the former case would not explicitly require linguistic processing. Studies using a similar paradigm, in which the perception of speech degraded by vocoding is restored through presentation of the un-vocoded speech, have found that "priming" degraded speech in this manner enhances tracking of phonetic features in the primed, degraded speech (Di Liberto, Crosse, et al., 2018; Di Liberto & Lalor, 2016; Di Liberto, Lalor, et al., 2018). Finally, Zoefel and VanRullen (2016) reported similar low-frequency tracking of normal and time-reversed speech; their behavioral task involved detection of a tone-pip embedded in the speech signal which, again, does not explicitly require linguistic processing. Given the number of studies that have found a connection between speech intelligibility and the neural entrainment to speech, and the apparent commonality between studies that have failed to replicate this effect, we suggest that entrainment to speech is related to its intelligibility through focused top-down mechanisms that are brought to bear only when the speech is task-relevant and understandable as speech. In our view the speech tracking response reflects the combination of stimulus-driven, bottom-up activity related to the features of the acoustic signal itself, and top-down modulatory activity mediated by cognitive processes including: task demands, prior knowledge, and contextual factors.

1.5 Novel thesis contributions: Top-down factors affecting speech tracking

In this thesis we explore the influence of top-down factors affecting cortical speech tracking responses to speech signals in complex acoustic scenes. Several ECoG studies have shown that activity in higher-order brain areas modulated speech-related auditory activity based on attention (Zion Golumbic, Ding, et al., 2013) and speech intelligibility (Ding, Melloni, et al., 2016; Leonard, Baud, Sjerps, & Chang, 2016), which suggests that neural entrainment to speech in auditory cortex is subject to top-down modulation by non-auditory areas. In Chapter 2 we consider cortical responses to target and distractor speech streams in a multi-talker environment. While previous studies have described an enhancement of neural entrainment to attended versus ignored speech in two-talker paradigms, we aimed to expand on that result by testing the effect of increasing the set-size of distractors in the acoustic scene. We also consider a possible mechanism of distraction in which to-be-ignored speech signals intrude on perception due to their actively being tracked in place of the to-be-attended speech. In Chapter 3 we question the role of the broadband acoustic envelope as a key feature of speech that enables entrainment by embedding the speech signal in a background of carefully modulated noise to eliminate amplitude fluctuations in the acoustic scene and we describe a novel component of the speech-tracking response related to segregating speech from background noise. In Chapter 4 we consider entrainment to acoustic and phonetic features during interrupted speech and examine the relationship between neural speech tracking and perceptual restoration of noise-interrupted speech.

2 The Effects of Distractor Set-size on Neural Tracking of Attended Speech

2.1 Introduction

The perception of natural speech in real-world environments requires the auditory system to extract a complex, dynamic acoustic signal from a complex background. A typical acoustic scene contains a mixture of sounds emitted from any number of sources, yet the human auditory system is able to routinely isolate a single voice from the mixture and extract meaningful information from it. This phenomenon, and the associated computational challenges, are commonly referred to as the "cocktail party problem" (Cherry, 1953). Despite more than half a century of dedicated study of this problem, the neural mechanisms that enable the human brain to solve the cocktail party problem and understand speech in challenging acoustic environments have not been fully elucidated. Recent work on selective attention has begun to elucidate the importance of the low-frequency dynamics that are inherent to speech stimuli.

Selective attention can enhance perception and memory of a single attended voice, even in environments with competing sound sources (Broadbent, 1952; Treisman, 1964). Relative differences in loudness, spectral distinctiveness, spatial separation, and similarity between temporal envelopes are known to influence the discriminability of target speech in environments with two competing speakers (Arbogast, Mason, & Kidd, 2002; Bronkhorst, 2015; Brungart, 2001), while adding more distractors to the scene can also impair perception of the target stream (Brungart, Simpson, Ericson, & Scott, 2001; Ericson, Brungart, & Brian, 2004).

Since speech allows us to communicate in noisy environments, a selective attention mechanism is fundamentally important to the perception of speech. There is limited processing of unattended speech (Cherry, 1953; Holender, 1986; Lachter, Forster, & Ruthruff, 2004; Treisman, 1964), however the exact limits remain a matter of some controversy (cf. Aydelott, Jamaluddin, & Nixon Pearce, 2015; Rivenez, Guillaume, Bourgeon, & Darwin, 2008). Neurobiologically, attention may enhance speech comprehension by increasing the brain's sensitivity to physical features related to the attended speech stream, while decreasing sensitivity to the features of competing sounds (Kaya & Elhilali, 2017; Knudsen, 2007; Lakatos et al., 2013) and by strengthening the relative connection amongst language processing neural networks (Giraud & Poeppel, 2012; Hickok & Poeppel, 2007; Morillon et al., 2012; Vander Ghinst et al., 2016).

Two inter-related neural mechanisms have recently been proposed to explain how the brain solves the cocktail party problem. These are based on two important neurophysiological results: First, neural sensitivity is modulated by subthreshold, low-frequency oscillations of the membrane potential (Fries, 2005; Volgushev et al., 1998). Second, that the phase of oscillations in auditory cortex tracks low-frequency amplitude modulations in speech signals (Abrams, Nicol, Zecker, & Kraus, 2008; Ahissar et al., 2001; Hertrich, Dietrich, Trouvain, Moos, & Ackermann, 2012; Luo & Poeppel, 2007) (low-frequency phase tracking). Thus, the selective entrainment hypothesis (Schroeder & Lakatos, 2009; Zion Golumbic, Poeppel, & Schroeder, 2012) proposes that the phase-tracking of low-frequency modulations of a speech signal by neuroelectric oscillatory activity increases cortical sensitivity to the target acoustic stream. By extension, selective

entrainment also reduces sensitivity to the spectrotemporal features of distractor streams to which neuroelectric oscillations are not entrained. A second, segmentation focused hypothesis suggests that low-frequency modulations enhance the segmentation of the continuous acoustic speech signal into discrete syllables, which are subsequently analyzed by the brain for their linguistic content (Ghitza, 2011; Ghitza & Greenberg, 2009; Greenberg, 1996). These hypotheses are not mutually exclusive, rather they are linked by a common proposed mechanism and taken together they suggest an explanation for how failures of attention impair speech processing: failure to entrain to a target speech stream entails the dysfunction of an entrainment-based segmentation mechanism.

There is a growing body of literature that has studied the neural phase-tracking of speech in the presence of competing sounds. Evidence from scalp-recorded EEG (Hambrook & Tata, 2014; Horton et al., 2013; Kerlin et al., 2010; Kong et al., 2014; Power et al., 2012), MEG (Ding et al., 2014; Ding & Simon, 2012b; Rimmele et al., 2015; Zion Golumbic, Cogan, et al., 2013), and intracranial recordings (Mesgarani & Chang, 2012; Zion Golumbic, Ding, et al., 2013) have all shown that attention modulates the neural phase-tracking of speech signals and that such tracking is associated with enhanced perception of the target speech stream. However, while these attentional studies frequently evoke the cocktail party problem, they use simple acoustic scenes consisting of a single target speech stream competing with a single distractor source.

The current study seeks to extend those neurophysiological results to more complex acoustic scenes containing more than two simultaneous talkers and elucidate the neural mechanisms of speech-on-speech interference described by previous psychophysical studies. A number of psychophysical studies have investigated the effect of adding multiple talkers to an acoustic scene (Brungart et al., 2001; Humes, Kidd, & Fogerty, 2017; Miller, 1947; Simpson & Cooke, 2005) and found that speech perception is systematically impaired as the number of talkers in a scene increases from two to eight, however as these were purely behavioral studies they shed little light on the neural mechanisms responsible for the reduced performance.

If the neural tracking of speech dynamics is a mechanism for implementing selective attention, then we should expect perceptual performance and the speech-locked phase-tracking of the EEG signal to vary together as more distractors are added to the scene. The present study investigated two additional questions about phase-tracking of low-frequency speech dynamics: first, we used both natural and vocoded speech – a processed version of speech in which acoustic energy is filtered into well-defined, non-overlapping frequency bands - to consider whether phase-tracking is a within-band mechanism of selection. Second, we measured whether distractor streams are phase-tracked on trials in which a distractor is perceived instead of the target. In this way we tested the hypothesis that transient phase-tracking of a distractor is an active mechanism of distraction.

2.2 Methods

2.2.1 Participants

31 undergraduates from the University of Lethbridge were recruited and participated for course credit: 17 participants (mean age: 21.9 years; 7 females; 5 left-handed) heard natural speech stimuli while 14 participated (mean age: 21.6 years; 8 females; 0 left-handed) in a version of the experiment in which the speech stimuli were first vocoded (see details below). Participants provided informed written consent. Procedures were in accordance with the Declaration of Helsinki and were approved by the University of Lethbridge Human Subjects Review Committee. Participants were neurologically normal and reported normal hearing.

2.2.2 Stimuli and task

All stimuli were presented in free field by an Apple Mac Pro with a firewire audio interface (M-Audio Firewire 410). Participants sat in the center of an array of near-field studio monitors (Mackie HR624 MK-2) arranged in a circle. A target speech stream was presented from a speaker directly in front of the participant. Distractor speech streams were presented from two, four, or six speakers in symmetric locations around the circular array (Figure 2.1). Speech stream presentation was controlled by a program custom coded using Apple Computer's Core Audio framework (Mac OS 10.6).

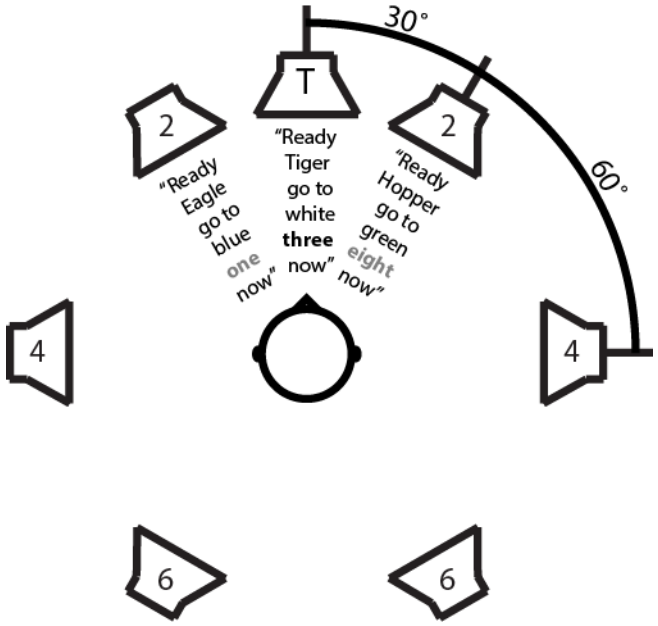

Figure 2.1: Schematic of the speaker array. Target speech streams (labelled "T") were presented simultaneously with two (from speakers labeled "2"), four (from speakers labeled "2" and "4"), or six (from all numbered speakers) distractor speech streams while listeners monitored the target stream for number keywords.

Each speech stream consisted of the concatenation of eight sentences, spoken by the same speaker, from the Coordinate Response Measure (CRM) Corpus (Bolia, Nelson, Ericson, & Simpson, 2000). The CRM corpus consists of predictably structured sentences of the format: "Ready <call sign> go to <color> <number> now," spoken by four male and four female speakers. On each trial listeners were simultaneously played one target speech stream and up to six distractor speech streams, each spoken by a unique speaker. Each block contained twelve 15.5 second stimuli which were divided into pseudo-randomly ordered sub-blocks of four stimuli at each distractor set size.

Listeners were tasked with reporting the number word ("one", "two", etc.) spoken from the target stream at the center speaker by pressing the corresponding number key on a keyboard in front of them. Participants responded as quickly and as accurately as possible. Unique number words occurred in all streams in close temporal proximity; the standard deviation from the mean latency of number word onsets across all speech streams on a given trial was 55 ms. Trials for which participants reported the number from the target stream were considered *correct*; responses in which listeners reported the number from a distractor stream were labeled *intrusion* errors; responses in which listeners reported a number that was not present in any stream were labeled *insertion* errors. Thus, intrusion and insertion errors differed in the likely source of the error: Intrusion errors are so-called because words from a distractor stream seem to have intruded on the successful perception of the target stream, while insertion errors occur when listener's perceptual mechanism has inserted an unheard word into the scene. The presumptive causes of these two types of errors are fundamentally different: intrusion errors occur when information from a distractor stream interferes with the perception of the target, while insertion errors most likely occurred when the listener lacked information completely and was forced to guess from among the limited pool of possible number words.

Because the number of distractor streams varied between distraction conditions while the possible pool of numbers in the auditory scene was always eight (i.e. "one" "two" through "eight") the relative distribution of intrusion and insertion errors one would expect by chance differs between conditions, making a direct comparison of error

rates between conditions difficult to interpret. To address this difficulty, we instead consider the log-transformed ratio of intrusion errors to insertion errors (RII), normalized by the ratio predicted by chance based on the number of distractors in the scene. The interpretation of the RII is straightforward: values greater than zero indicate that listeners are more likely commit intrusion errors than insertion errors while values less than zero indicate that listeners were more likely to commit insertion errors relative to intrusion errors. Differences between distraction conditions can be meaningfully compared because the differences in the distribution of errors we would expect by chance have been normalized.

As the number of distractor streams increases, the total level of acoustic energy in the scene increases as well, which could potentially mask the target simply due to interference in the auditory periphery (i.e. energetic masking). To address this potential confound, the stimuli for one experimental group of 14 participants was vocoded to produce intelligible but spectrally non-overlapping speech signals (Arbogast et al., 2002; Brungart, Simpson, Darwin, Arbogast, & Kidd, 2005; Dorman, Loizou, & Rainey, 1997; Ihlefeld & Shinn-Cunningham, 2008; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Each speech signal was bandpass filtered into 16 fixed-frequency bands of 1/3 octave width, with center frequencies distributed on a logarithmic scale from 175 Hz to 5.6 kHz every 1/3 octave. The envelope of each frequency band was extracted using the Hilbert transform and that envelope was multiplied by a pure tone carrier at the center frequency of that band. For each trial, target stimuli were constructed by randomly selecting and summing four low-frequency (175-882 Hz) bands and four high-frequency

(1.1 - 5.6 kHz) bands; distractor stimuli were constructed by combining the eight remaining bands not used in the construction of the target. This resulted in minimal spectral overlap between the target and distractors, minimizing interference at the level of the basilar membrane.

2.2.3 EEG analysis

EEG was recorded with 128 Ag/Ag-Cl electrodes in an elastic net (Electrical Geodesics Inc., Eugene, OR, USA). Scalp voltages were recorded at a 500 Hz sampling rate and impedances were maintained under 100 kΩ. Data were first analyzed using the BESA software package (Megis Software 5.3, Grafelfing, Germany). Data were visually inspected for bad channels and the signal from a small number of electrodes (10 or fewer) was replaced with an interpolated signal. Because each trial was 15.5 seconds long, eye movement artifacts occurred in a majority of trials, therefore eye movement artifacts were corrected using an adaptive artifact correction algorithm (Ille, Berg, & Scherg, 2002). Data were interpolated to an 81-channel 10-10 montage and further analyzed in MATLAB (MATLAB version 7.10.0; The Mathworks Inc., 2010, Natick, MA, USA) using custom scripts and EEGLAB functions (Delorme & Makeig, 2004).

To isolate EEG activity phase-locked to each of the unique competing speech streams, the first derivative of the acoustic envelope for each stream was calculated and cross-correlated with the EEG. This acoustic envelope for each speech stream was calculated by taking the absolute value of the Hilbert transform and low-pass filtering the resulting waveform with a cut-off at 25 Hz. The acoustic envelope was then down-

sampled to match the sample rate of the EEG data. The first-derivative of the resulting signal was calculated, half-wave rectified, and normalized such the sum of the signal across the whole epoch equaled 1 (Hambrook & Tata, 2014; Hertrich et al., 2012). Thus, we obtain a signal that captures transient energy increases, an aspect of acoustic stimuli to which the auditory system is known to be tuned (Fishbach, Nelken, & Yeshurun, 2001; Howard & Poeppel, 2010). These speech envelopes were then cross-correlated with each channel of the time-aligned EEG data to arrive at a cross-correlation function that reflects activity phase-locked to the acoustic dynamics of each particular speech stream. Peri-target epochs were defined as [-1000 1000] ms for the acoustic signal and [-1700 2300] ms for the recorded EEG data; a longer epoch was used for the EEG data to remove the need to pad the data with zeros or normalize the cross-correlation function at extreme lags. Trials were separated based on task performance relative to each target as past studies have shown minimal tracking of the target stream on error trials (Hambrook & Tata, 2014; Mesgarani & Chang, 2012).

To determine the frequency content of the observed phase-locked activity, wavelet decomposition was performed on the cross-correlation function for the interval of cross-correlation lags [-200 800] ms. Evoked power was calculated as the power in the trial-averaged cross-correlation function, normalized by the mean evoked power across the whole epoch. For all distractor set sizes the power from all two, four, or six distractor streams was averaged before comparison with power phase-locked to the target stream.

2.3 Results

2.3.1 Correct responses

Listener's ability to identify number words from the target stream was impaired as the number of distractors in the auditory scene increased (Figure 2.2A). A 2x3 mixed ANOVA with stimulus type (natural, vocoded) as a between-subject factor and distractor number (two, four, six) as a within-subject factor revealed a significant main effect of distractor number on correct response rate ($F(2,58)=373.3$, $p<0.001$, $\eta^2=0.93$) as well as an interaction between the distractor number and stimulus vocoding ($F(2,58)=37.86$, $p<0.001$, $\eta^2=0.57$). Analysis of the simple main effects identified significant effects of distraction for both natural ($F(2,28)=244.5$, $p<0.001$ Benjamini-Hochberg adjusted, $\eta^2=0.95$) and vocoded ($F(2,28)=55.44$, $p<0.001$ Benjamini-Hochberg adjusted, $\eta^2=0.80$). There was also a simple main effect of stimulus type for two distractors ($F(1,29)=8.02$, $p=0.017$ Benjamini-Hochberg adjusted, $\eta^2=0.22$), but there was not a significant effect for four or six distractors ($F(1,29)<1.34$, $p>0.512$ Benjamini-Hochberg adjusted, $\eta^2<0.04$), suggesting that the vocoding process impaired baseline intelligibility of the target, but did not result in systematically different distraction at higher numbers of distractors. Increased distractor set-size impaired performance for both natural and processed speech stimuli; crucially, the spectral separation between the target and distractor streams only somewhat mitigated the effect of distraction indicating that the target and distractor streams are primarily interfering after they pass through the auditory periphery.

2.3.2 Analysis of errors

Intrusion errors, that is reporting a number word from a distractor stream, increased with increasing set size of distractors (Figure 2.2B & 2.2C). A 2x3 mixed ANOVA with RII as the measurement variable, stimulus type (natural, vocoded) as a between-subject factor and distractor set size (two, four, six) as a within-subject factor revealed a significant main effect of distractor set size ($F(1.66,48.10)=47.83$, $p<0.001$, $\eta^2=0.62$, Greenhouse-Geisser corrected) and stimulus type ($F(1,29)=10.50$, $p=0.003$, $\eta^2=0.27$), as well as a significant interaction between distractor set size and stimulus type ($F(1.66,48.10)=6.09$, $p=0.007$, $\eta^2=0.17$). Analysis of the simple main effects identified significant effects of distractor set size for both natural ($F(2,28)=52.22$, $p<0.001$ Benjamini-Hochberg adjusted, $\eta^2=0.79$) and vocoded ($F(2,28)=9.00$, $p=0.001$ Benjamini-Hochberg adjusted, $\eta^2=0.39$) stimuli. There was not a significant simple main effect of stimulus type for two distractors ($F(1,29)=0.063$, $p=0.80$ Benjamini-Hochberg adjusted , $\eta^2=0.002$), however there were significant effects at four distractors ($F(1,29)=6.07$, $p=0.020$ Benjamini-Hochberg adjusted, $\eta^2=0.17$) and six distractors ($F(1,29)=11.32$, $p=0.002$ Benjamini-Hochberg adjusted, $\eta^2=0.28$) suggesting that natural, unfiltered distractors are more likely to "intrude" on perception than distractors that do not spectrally overlap with the target stream. It is also possible that the perception of individual distractors in the vocoded version of the experiment is impaired due to their perfect spectral overlap with all the other distractors.

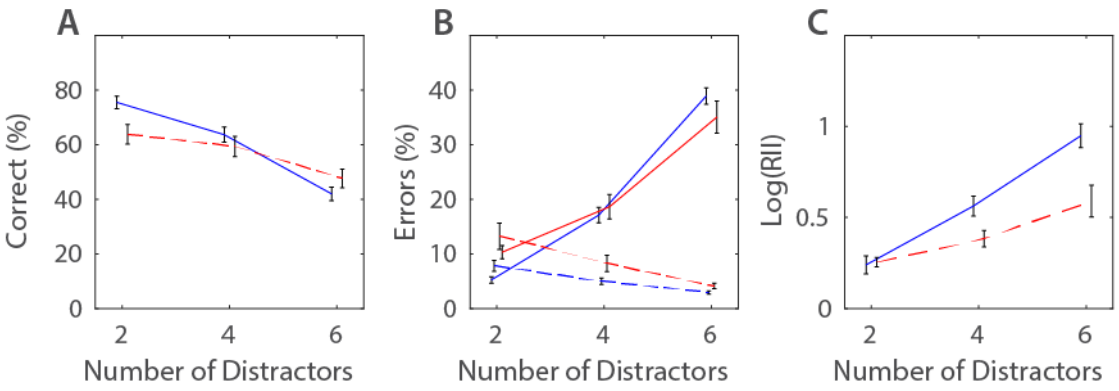

Figure 2.2: Behavioral results for the listening task using natural (blue) and vocoded (red) speech samples. (A) Mean correct response rate plotted as a function of the number of distractors in the acoustic scene. (B) Mean intrusion error rate (solid line) and insertion error rate (dashed line) plotted as a function of the number of distractors in the scene. (C) Log-transformed ratio of chance-normalized intrusion errors to insertion errors. Error bars indicate standard error of the mean.

2.3.3 EEG results

We used a wavelet time-frequency decomposition to explore the time-frequency content of the cross-correlation function for target and distractor speech streams (Figure 2.3). Previous studies strongly suggested that EEG signals maximally phase-locked to attended speech within the theta band (i.e. 4-8 Hz). For the target stream we observed a peak in phase-locked theta-band power at a lag of approximately 100 ms for all distractor set sizes (Figure 2.4). We performed a 2x3x2 mixed ANOVA in which phase-locked theta-band power from [40 160] ms lag was the measurement variable; stimulus type was a between-subject factor (natural, vocoded); distractor number (two, four, six), and attention (attended stream, distractor stream) were within-subject factors (Figure 2.5A). This analysis identified significant main effects of distractor number

($F(2,58)=10.00$, $p<0.001$, $\eta^2=0.26$) and attention ($F(1,29)=19.58$, $p<0.001$, $\eta^2=0.40$), as well as an interaction that trended towards significance between attention and distractor number ($F(2,58)=2.84$, $p=0.067$, $\eta^2=0.09$). There was not a significant effect of stimulus type ($F(1,29)=0.21$, $p=0.65$, $\eta^2=0.007$), and there were no significant interactions between stimulus type and distractor number ($F(2,58)=0.87$, $p=0.42$, $\eta^2=0.029$) nor between stimulus type and attention ($F(1,29)=0.034$, $p=0.86$, $\eta^2=0.001$).

While similar experiments have previously found that tracking of attended speech is reduced in epochs surrounding errors of perception, they were not designed to interrogate how tracking of *distractors* is affected during task errors in which the distractor is perceived as the target. One potential mechanism of distraction is that distractor streams momentarily co-opt the neural dynamics that should track the target speech. In this case we would predict that the epoch around "successful" intruding distractors would be tracked more than other distractors that were not perceived. We would further predict that the epoch around a perceived distractor on intrusion errors would be tracked similarly to the epoch around perceived targets on correct trials. The first prediction was tested using a 2x2 mixed ANOVA in which phase-locked theta band power within a [40 160] ms lag was the measurement variable, stimulus type as a between-subject factor (natural, vocoded), and perception of the distractor stream as a within-subject factor (intruding distractor, rejected distractor). This analysis revealed no significant effects of perception ($F(1,91)=0.54$, $p=0.46$, $\eta^2=0.006$) or stimulus-type ($F(1,91)=0.56$, $p=0.46$, $\eta^2=0.006$); successfully intruding distractor streams are not preferentially tracked by the EEG (See Figure 2.5B). A second 2x2 mixed ANOVA with

phase-locked power as the measurement variable, stimulus type as a between-subject factor (natural, vocoded), and attention (attended, ignored) as a within-subject factor. This analysis revealed a significant effect of attention on phase-tracking ($F(1,91)=46.54$, $p<0.001$, $\eta^2=0.34$) with no significant effect of stimulus type ($F(1,91)=0.99$, $p=0.32$, $\eta^2=0.011$); ignored speech that intruded onto perception was not tracked as well as successfully perceived attended speech (Figure 2.5B). Taken together, these results suggest that active but transient phase-tracking of a distractor stream is not the mechanism by which distracting speech intrudes into perception.

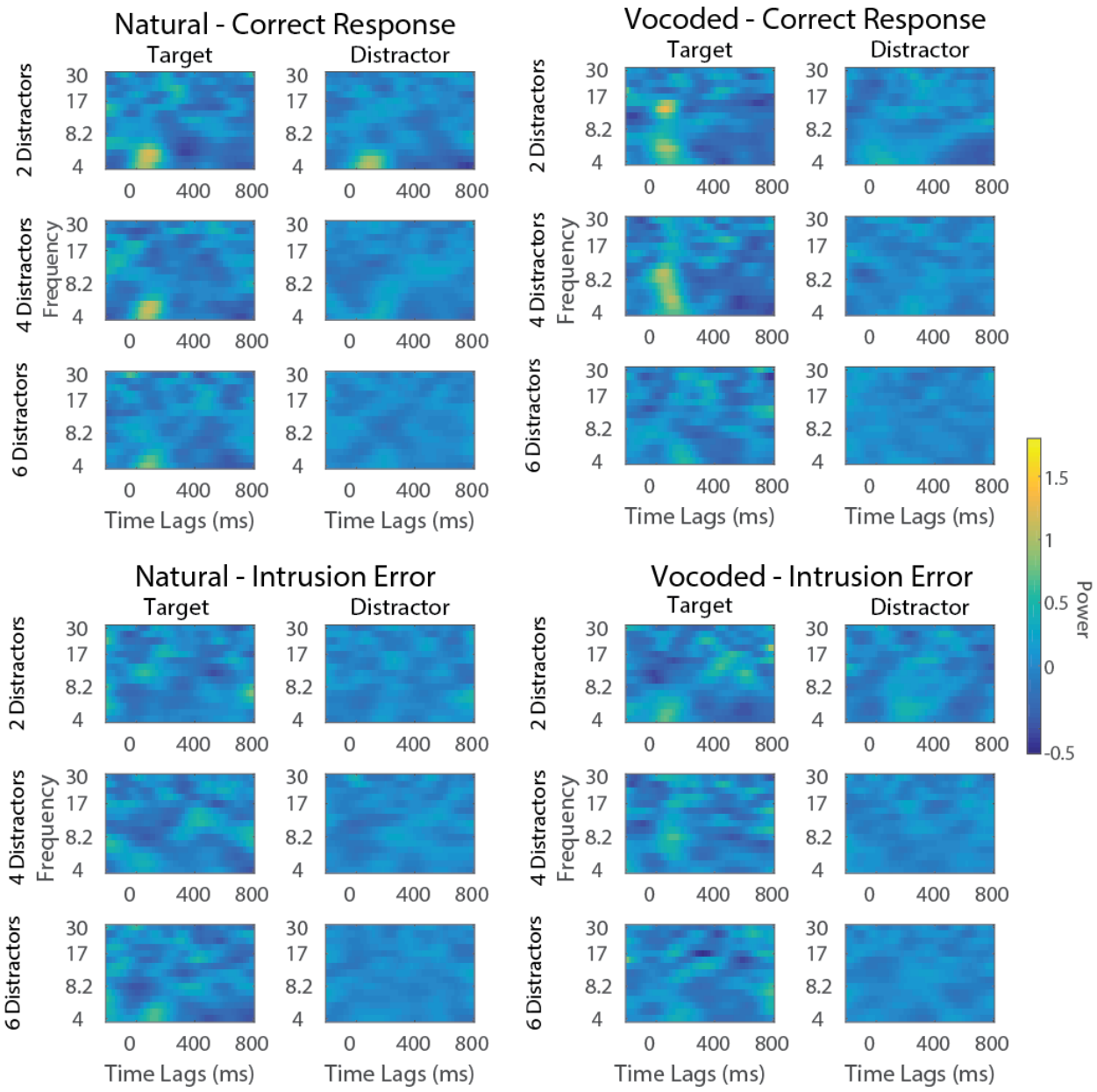

Figure 2.3: Time-frequency representation of cross-correlation function. Phase-locked power in the speech-EEG cross-correlation function for natural (left column) and vocoded (right column) speech for target and distractor speech streams split into correct responses (top row) and intrusion errors (bottom row).

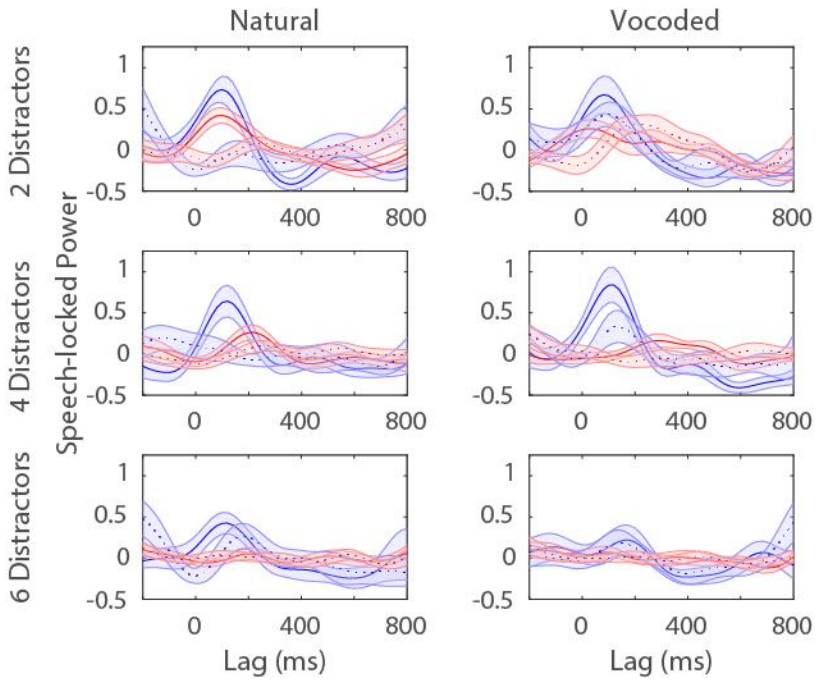

Figure 2.4: Phase-locked theta-band (4-8 Hz) power in the speech-EEG cross-correlation function for natural (left) and vocoded (right), target (blue) and ignored (red) speech streams surrounding correct responses (solid lines) and intrusion errors (dashed lines). Light shaded outline indicates standard error or the mean.

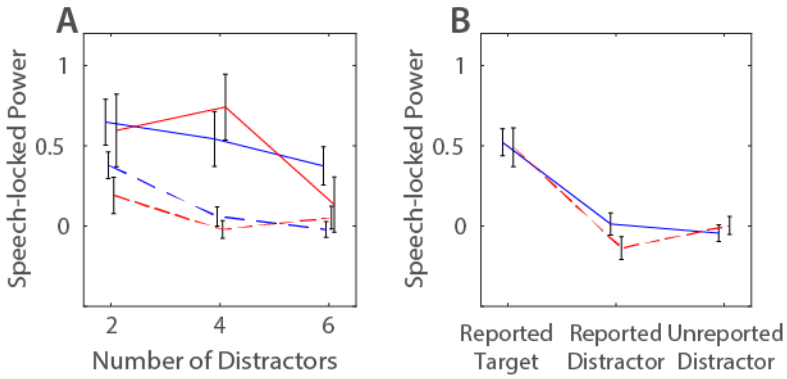

Figure 2.5: (A) Effect of attention on phase-locked theta power. Phase-locked theta-band (4-8 Hz) power in the speech-EEG cross-correlation function for the interval [40 160] ms lag for natural (blue) and vocoded (red) speech streams surrounding correct responses. Responses to attended speech are plotted with solid lines, the mean response to distractor speech is plotted with dashed lines. (B) Phase-locked theta power as a function of perception and attention. Phase-locked theta power for the interval [40 160] ms lag for natural (blue) and vocoded (red) speech streams. The speech tracking signal is strongest to attended and successfully perceived speech (Reported Target); there is no apparent increase in speech tracking of the successfully intruding distractor (Reported Distractor) relative to successfully ignored distractors (Unreported Distractor). Error bars indicate standard error of the mean.

2.4 Discussion

Our behavioral results show a clear effect of distractor set size on listeners' ability to identify target words in a target speech stream. Participants were significantly more likely to identify words from the target stream when there were fewer distractors, and more likely to make intrusion errors when there were more distractors. Vocoding the target and distractor streams to be spectrally distinct somewhat mitigated the effects of increasing distractor set size. However, even when the target and distractors were spectrally distinct, and energetic interference at the auditory periphery was minimized, increasing the number of distractors still impaired performance. For both natural and vocoded speech, the ratio of intrusion errors to insertion errors increased relative to

chance as distractor set size increased. If the primary effect of distraction was due to increased interference at the auditory periphery then we should expect this ratio to return to chance values as distractors were added to the scene. Taken together, these data strongly suggest that the informational content present in the distractor streams interferes with the representation of the target speech stream at the cortical level.

Our electrophysiological results show increased theta-band EEG power phase-locked to the acoustic dynamics of target speech, compared to distractor speech. This result agrees with previous studies using two-talker auditory scenes, which have found that attention enhances low-frequency activity evoked by continuous speech (Ding & Simon, 2012b; Hambrook & Tata, 2014; Kerlin et al., 2010; Power et al., 2012; Zion Golumbic, Ding, et al., 2013). In the present study, this enhancement was maintained, even in very crowded acoustic scenes with six distractors, suggesting that phase-tracking of the acoustic dynamics of a target represents a generalized mechanism for maintaining the neural representation of that stream.

A study by Rimmele et al. (2015), in which subjects heard simultaneously presented natural and vocoded speech while monitoring one stream for a loudness increase, previously found that the tracking of vocoded speech was not modulated by attention. They suggested that their results indicate that the attentional enhancement of speech tracking depends on the presence of fine-structure in the stimulus and that fine-structure in natural speech is only utilized when the speech is the object of attention. They go on to suggest that processing the temporal fine structure of speech reflects

linguistic processing as both eliminating fine-structure by vocoding (Dorman et al., 1997; Shannon et al., 1995; Sheldon, Pichora-Fuller, & Schneider, 2008; Smith, Delgutte, & Oxenham, 2002) and ignoring speech (Cherry, 1953; Treisman, 1964) impair speech perception. Our current results stand in contrast to their findings. We found that attention enhanced the tracking of successfully perceived vocoded speech, relative to ignored speech, in the absence of fine structure cues. A possible explanation for this discrepancy can be found in the different stimulus processing procedure used by Rimmele et al. (2015). In their experiment they used four vocoder bands spread over the entire range of the human cochlea (80Hz – 20 kHz; center frequencies: 0.292, 1.15, 3.75, 11.7 kHz), while previous studies of the intelligibility of vocoded speech have restricted the frequency range to a maximum cut-off frequency of around 6-8 kHz reflecting the limited vocal range of the male speakers used in those studies (Dorman et al., 1997; Ihlefeld & Shinn-Cunningham, 2008; Shannon et al., 1995; Sheldon et al., 2008; Smith et al., 2002). Thus, their choice of stimulus processing bands resulted in speech of significantly degraded acoustic quality and intelligibility, which suggests listeners could only extract limited linguistic information. It is also worthwhile to note that the task used by Rimmele et al. (2015) simply required monitoring the attended stream for changes in loudness, a task which may benefit from but does not explicitly require linguistic processing. While our results dispute the claim that the modulation of neural tracking of speech requires access to fine structure information, they do support the overall conclusion that attended speech is tracked more effectively due to linguistic processing. Linguistic processing may provide a top-down influence on the tracking of

speech mediated by attention. In the previous study by Rimmele et al. (2015) the reduced linguistic content available to listeners of band-limited vocoded speech and the non-linguistic nature of the behavioral task possibly led to an attenuation of top-down attentional factors that enhance the tracking of speech.

Previous psychoacoustic studies have considered the differential effects of energetic and informational interference on the perception of speech. Ihlefeld & Shinn-Cunningham (2008) identified three linked mechanisms affecting speech identification in this kind of task: across-time linkage, short-term segmentation, and selective attention.

Across-time linkage refers to the integration of the features of an acoustic stream across temporal discontinuities like silent gaps or stop consonants and is influenced by stable (in the current experiment) factors including spatial location, pitch, timbre, and overall intensity (Bregman, 1990; Culling & Summerfield, 1995; Darwin, 1997). Task errors due to a failure of temporal integration would result in listeners temporarily monitoring a distractor stream for a task-relevant keyword as if it were the target stream and we would expect them to commit intrusion errors as a result. Errors of this type, due to a failure of temporal integration of the target stream, should be accompanied by erroneous phase-tracking of the reported distractor stream; however, we found no evidence that the successful distractor stream was tracked differently than other distractors, and it certainly was not tracked as if it was the target stream. Thus, failures of across-time linkage are an unlikely mechanism for causing errors in this task.

Short-term segmentation refers to the process by which some portion of the acoustic mixture of the scene is segregated into discrete speech sounds. Segmentation is primarily based on the brain's analysis of the spectrotemporal properties of a sound stream; in particular, low-frequency modulations are believed to provide a basis for dividing an incoming speech signal into syllabic units (Ghitza, 2011; Ghitza & Greenberg, 2009; Greenberg, 1996). Such a mechanism, operating within discrete frequency bands (Doelling et al., 2014), is robust to energetic interference from stationary signals but may be susceptible to interference by competing signals which share similar dynamics to the target signal. Indeed, such an interference effect may be the reason that adding more distractor streams to the scene impairs perception of a target stream, even when distractors are spectrally distinct from the target speech stream as in the vocoded speech group.

Selective attention refers to the ability to selectively tune the brain's sensitivity to a single target stream among a mixture of competing sounds. Selective attention may be directed to a number of acoustic features including spatial location, prosody, pitch, timbre, and speaker identity (Darwin, Brungart, & Simpson, 2003; Darwin & Hukin, 2000; Freyman, Helfer, McCall, & Clifton, 1999; Shinn-Cunningham, Ihlefeld, Satyavarta, & Larson, 2005). Attention can enhance the sensory representation of target features while suppressing the representation of competing signals (Desimone & Duncan, 1995). This effect is appears in our data as increased EEG phase-tracking of speech is evident for attended speech streams but not ignored streams.

The selective entrainment hypothesis (Schroeder & Lakatos, 2009; Zion Golumbic et al., 2012) proposes that phase entrainment of neural oscillations to the temporal dynamics of a behaviorally relevant auditory stream is a mechanism for attentional selection. The hypothesis arises from the fact that neural sensitivity is modulated by the phase of low-frequency oscillations (Engel, Fries, & Singer, 2001; Volgushev et al., 1998). Thus, oscillatory activity forms temporal windows in which post-synaptic cells may be more (or less) sensitive to excitatory input. This principle forms the basis of the theory of communication by coherence (Fries, 2005), which states that communication between neuronal ensembles is optimally efficient when graded potentials in pre- and post-synaptic cells are phase-aligned This phase alignment ensures that synaptic transmission occurs within those windows during which the post-synaptic cell is biased towards depolarization. By modulating the phase of these oscillations relative to a stimulus stream, a selective entrainment mechanism forms a sort of filter – allowing some neural assemblies to ignore inputs from non-selected cells while enhancing sensitivity to selected cells.

Selective attention enables the enhanced representation of a single information source at the cost of impairing the perception of other sources. Selective entrainment may provide a mechanistic explanation for selective attention to temporally predictable auditory streams. By phase-locking auditory neural activity to the dynamics of an attended stream, neurons encoding the relevant features of that stream may be biased to fire more readily. The theory of communication by coherence provides a framework by which attended auditory signals are transmitted to other brain areas responsible for

higher-order cognitive processes including: semantic and grammatical processing, working memory, reward-processing, and response-planning (Ding, Melloni, et al., 2016; Giraud & Poeppel, 2012). Selective entrainment not only biases the brain to respond to the attended stream, it can also selectively block competing signals from those same higher-order cognitive processes, even if they share similar spectral content, by virtue of the periodic nature of neural oscillations. Signals that do not share the same spectrotemporal dynamics of the attended stream will arrive during non-optimal temporal windows and be suppressed. The presence of additional competing speech to the acoustic scene seems to degrade this attentional mechanism as we observed reduced entrainment to the target speech as distractor set-size increased.

We should note that the two proposed mechanistic roles of neural phase-tracking: as a mechanism for selective attention and providing a framework for speech segmentation are not mutually exclusive. Indeed, both make similar predictions about the perceptual consequences of phase-tracking to a target speech stream. The selective entrainment hypothesis suggests that neural tracking of a speech stream enables an enhanced representation of the features of that stream which leads to improved perception of that stream. The theory that oscillatory activity supports the segmentation of speech likewise predicts that phase-tracking of speech enables the parsing of the acoustic signal into meaningful speech sounds. The current experiment provides little insight into a possible dissociation between these two theories; it can only confirm the strong link between successful perception of a speech stream and the brain's tracking of the dynamics of that stream.

Speech-tracking fails in the presence of distracting voices, even when the target speech occupies its own frequency bands as in our vocoded condition. This points to a central, rather than peripheral, mechanism that is vulnerable to interference by the additional load of acoustically dynamic, information-containing speech streams. Within auditory cortex itself competing, spectrally-overlapping streams can degrade the representation of a target stream by introducing increasing spike-activity unrelated to the target and suppressing target-related activity (Narayan et al., 2007). Reducing the spectral overlap between target and competing streams, as in our vocoded speech group, reduces the degree of interference within auditory cortex (Larson, Maddox, Perrone, Sen, & Billimoria, 2012). With these results in mind it appears that the reduction in the phase-tracking response due to increased distractor set-size is driven by interference within association areas, including within the language processing network. The proposed mechanistic explanation of reduced phase-tracking with increased distractor set-size fits well with psycholinguistic results that demonstrate that linguistic interference between competing speech streams is dependent on the intelligibility of the competing speech (Brouwer, Van Engen, Calandruccio, & Bradlow, 2012; Calandruccio, Dhar, & Bradlow, 2010; Van Engen & Bradlow, 2007), which cannot be explained by physical spectrotemporal similarities between the target and distractor speech (Calandruccio, Brouwer, Van Engen, Dhar, & Bradlow, 2013). Our results suggest a complicated interaction between attentional processes, speech intelligibility, and speech-entrained auditory cortical activity. This interaction suggests that language processing areas beyond auditory cortex exert some top-down influence that enhances the entrained

response to the acoustics of task-relevant speech, but that the effectiveness of this top-down modulation is itself modulated by the linguistic representation of the speech within cortex. Thus, this effect may be speech-specific, although other information-dense, dynamic, and salient stimuli such as music might produce similar interference effects.

We found that phase-tracking an attended speech stream is associated with the successful perception of that speech, even in a crowded 'cocktail party'-type environment with as many as seven concurrent speakers. The neural tracking of speech varied with the number of distractors in the acoustic scene, irrespective of the spectral overlap between targets and distractors, suggesting that the addition of more speech sources in the scene interferes with the cortical mechanisms – related to selective attention and or stream segregation – responsible for tracking a target speech signal.

3 Cortical Entrainment to Speech Occurs Without Broadband Envelope Dynamics

3.1 Introduction

Speech is an inherently rhythmic acoustic signal. The amplitude envelope of speech signals is modulated at around 5 Hz, putatively corresponding to the syllable rate and seemingly regardless of language or speech context (Ding, Patel, et al., 2016; Goswami & Leong, 2013). It has been proposed that the brain might leverage the predictability of this rhythmicity to facilitate speech perception by aligning oscillatory neural activity to the speech envelope. This entrainment is proposed to enhance stream segregation (Krishnan, Elhilali, & Shamma, 2014; Shamma et al., 2011), attentional selection (Ding & Simon, 2012b; Hambrook & Tata, 2014; Kerlin et al., 2010; Mesgarani & Chang, 2012; Power et al., 2012), and speech segmentation (Ghitza, 2011; Giraud & Poeppel, 2012).

Despite the apparent importance and versatility of the neural tracking of speech, it remains unclear which features of speech are essential to allow tracking to occur. Three aspects of the speech signal have been identified as candidate features: First, early investigations of the neural tracking of speech focused on modulations in the broadband temporal amplitude envelope of speech as the primary feature that enables speech tracking (Ahissar et al., 2001; Aiken & Picton, 2008; Luo & Poeppel, 2007). In this view, it is the amplitude of the acoustic signal itself that contains low-level cues that convey information about the dynamic contents of the speech. Second, researchers have suggested that the brain tracks higher-level spectrotemporal features of the speech acoustics such as modulations within discrete frequency bands (Ghitza, Giraud, &

Poeppel, 2013; Obleser et al., 2012). Third, it is also proposed that neural tracking of speech reflects entrainment to linguistic features of speech that are characterized by complex and variable conjunctions of acoustic features such as phonemes, syllables, words, or hierarchical prosodic or grammatical structures (Di Liberto, O'Sullivan, & Lalor, 2015; Ding, Melloni, et al., 2016; Mesgarani et al., 2014). The goal of the present study was to elucidate this question by temporally smoothing the low-level acoustic modulations of speech, leaving only higher-level spectrotemporal and linguistic features intact.

A prior investigation by Zoefel and VanRullen (2016) suggested at the importance of each of these features for speech tracking, however the stimulus conditions and behavioral task led to some difficulty with interpretation. Briefly, their experiment consisted of three distinct speech presentations: 1) speech presented by itself, 2) speech presented in a background of noise that obscured the broadband acoustic envelope but retained high-level spectrotemporal and linguistic features, and 3) a time-reversed speech-in-noise that retained high-level spectrotemporal modulations but eliminated envelope and linguistic features. Crucially, their behavioral task was not related to the presented speech; rather than listening to the speech, listeners were tasked with monitoring the acoustic environment for brief tone pips. It is well established that the processing of unattended speech for linguistic content is limited (Cherry, 1953; Dalton & Fraenkel, 2012; Lachter et al., 2004; Treisman, 1964) and the limited pre-attentive processing of speech is believed to be based on stimulus-memory-trace comparisons rather than the extensive processing of speech that is the object of attention

(Pulvermüller & Shtyrov, 2006). There is substantial evidence that attended speech is tracked by the brain to a significantly greater degree (Ding & Simon, 2012a; Hambrook & Tata, 2014; Kerlin et al., 2010; Mesgarani & Chang, 2012; Zion Golumbic, Ding, et al., 2013; Zion Golumbic et al., 2012) and that active listening to speech enhances brain activity in response to speech under poor listening conditions (Wild et al., 2012). Thus, Zoefel and VanRullen (2016) reported no significant differences between the neural tracking of speech-in-noise regardless of whether it was presented normally or time-reversed.

In the present study we investigated whether acoustic envelope modulations are necessary for the neural tracking of speech during active listening. We compared the neural response to speech presented alone with the response to speech presented against a background of amplitude-modulated noise that effectively smoothed the broadband envelope, while listeners were actively attending to the speech stimuli. Further, if envelope modulations are necessary for speech tracking, we tested the theory that tracking can be restored by spatially separating the speech from the masking noise. If the brain is able to track the low-frequency dynamics of speech, despite a smooth amplitude envelope, then other higher-level mechanisms must account for the speech tracking phenomenon.

3.2 Methods

3.2.1 Participants

48 undergraduates participated in the experiment. Two between-subjects conditions were tested: one in which target speech and masking noise were co-located at the same speaker, and one in which the target and masker were presented from different locations. Each group had 24 participants. The participants in the colocalized group had an age range of 18-29 years, with a mean age of 20.3 years; 2 were left-handed; and 13 were female. The participants in the spatially separated group had an age range of 18-24 years, with a mean age of 20.6 years; 2 were left-handed; and 15 were female. All participants were University of Lethbridge students and were recruited and participated for course credit. Participants provided informed written consent. Procedures were in accordance with the Declaration of Helsinki and were approved by the University of Lethbridge Human Subjects Review Committee. Participants were neurologically normal and reported normal hearing.

3.2.2 Stimuli

All stimuli were presented in free-field by an Apple iMac with a firewire audio interface (M-Audio Firewire 410). Participants sat 1.1 meters from a near-field studio monitor (Mackie HR624 MK-2) located on the front auditory midline. For the colocalization group, both speech and noise-maskers were presented from this midline speaker. For the spatial separation group, noise-maskers were presented from an identical studio monitor 30° to the right or left of the auditory midline with the location of the masker being pseudorandomly chosen and balanced between trials and

conditions. Stimulus presentation was controlled by a script custom coded using MATLAB (MATLAB version 7.10.0; The Mathworks Inc., 2010, Natick, MA, USA) and Psychophysics Toolbox Version 3 (Brainard, 1997).

Speech stimuli were constructed by concatenating sentences from the Pacific Northwest/Northern Cities (PN/NC) corpus (McCloy et al., 2013). The PN/NC corpus consists of recordings of male and female speakers reading 180 sentences from the IEEE "Harvard" set ("IEEE Recommended Practice for Speech Quality Measurements," 1969). Speech samples were created by concatenating three unrelated sentences read by a male voice to create a sample of speech roughly 6.5 s long.

Noise maskers were broadband noise with spectral composition matching the roughly 1/f, long-term average spectral composition of the speech stimuli. Spectrally matched noise was generated by randomly time-shifting each original speech segment and adding the resulting signals together 10 000 times and finally scaling the resulting signal to 2.5 times the original average RMS amplitude of the speech sample, resulting in a target-to-masker ratio of -8 dB. This procedure resulted in stationary noise that matched the average spectrum of the original speech samples. Two types of masker were tested: In the *Flat* Mask condition the noise masker consisted of a sample of noise with constant amplitude added to the acoustic scene. In the *Complementary* Mask condition, the amplitude of the noise masker was modulated by the inverse of the low-frequency amplitude envelope of the concurrently presented speech signal, effectively eliminating low-level amplitude modulations in the scene (Figure 3.1).

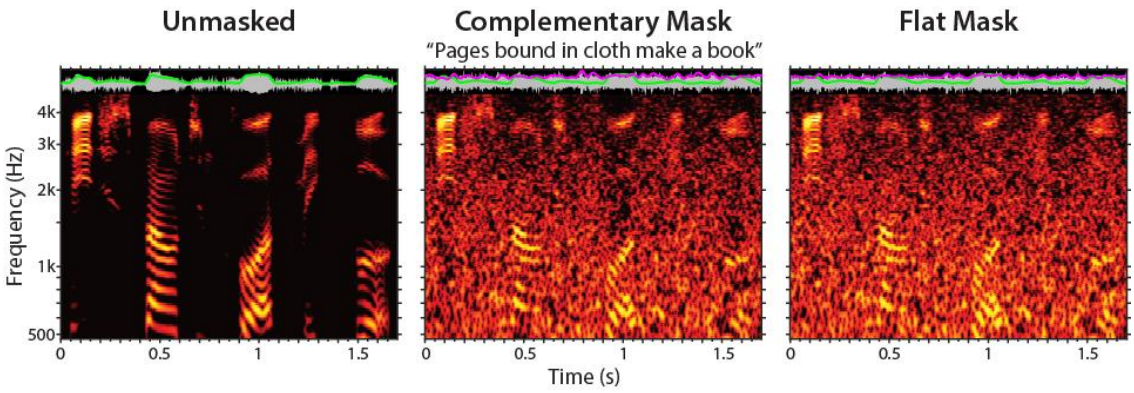

Figure 3.1: Waveform (grey) and spectrogram of an example sentence from the three experimental conditions. The speech envelope (green) and the masked stimulus envelope (magenta) are superimposed on the sound waveform.

3.2.3 Procedures

At the start of each session listeners were presented with a speech sample without a masker, followed by that same speech sample accompanied by an unmodulated masker, and finally the same sample presented with an envelope-modulated masker, to familiarize listeners to the stimuli and ensure they could hear the speech in the presence of masking noise. Listeners were instructed to listen carefully to the speech sample and maintain fixation at a fixation cross displayed on a monitor just below eyeline. In each trial listeners were presented with a speech sample either alone (in the *Unmasked* condition) or with a simultaneously presented noise masker (in the *Flat* and *Complementary* Mask conditions). Following each trial, listeners were given up to 25 seconds to recall as many words from the preceding speech sample as they could by

typing them into a keyboard; listeners could move on to the next trial before 25 seconds had elapsed if they could not recall anymore words. Each experiment consisted of three blocks, corresponding to the three experimental conditions, and each block contained 30 trials. The presentation order of the three conditions was counterbalanced using a Latin-square design.

Performance on individual trials was scored as a proportion of correctly transcribed words divided by the total number of words in the speech sample, excluding common articles "the", "a", and "an".

3.2.4 EEG analysis

EEG was recorded with 128 Ag/Ag-Cl electrodes in an elastic net (Electrical Geodesics Inc., Eugene, OR, USA). Scalp voltages were recorded at a 500 Hz sampling rate and impedances were maintained under 100 kΩ. Data were first analyzed using the BESA software package (Megis Software 5.3, Grafelfing, Germany). Data were visually inspected for bad channels and the signal from a small number of electrodes (10 or fewer) was replaced with an interpolated signal. Because of the length each trial, eye movement artifacts occurred in a majority of trials, therefore eye movement artifacts were corrected using an adaptive artifact correction algorithm (Ille et al., 2002). Data were interpolated to an 81-channel 10-10 montage and further analyzed in MATLAB (MATLAB version 7.10.0; The Mathworks Inc., 2010, Natick, MA, USA) using custom scripts and EEGLAB functions (Delorme & Makeig, 2004).

To isolate EEG activity phase-locked to the competing speech samples, the first derivative of the acoustic envelope for each sample was calculated and cross-correlated with the EEG. The acoustic envelope for each speech sample was calculated by taking the absolute value of the Hilbert transform of the sample and low-pass filtering the resulting waveform at 25 Hz. The acoustic envelope was then down-sampled to match the sample rate of the EEG data. The first-derivative of the resulting signal was calculated, half-wave rectified, and normalized such the sum of the signal across the whole epoch equaled 1 (Hambrook & Tata, 2014; Hertrich et al., 2012). Thus, we obtained a signal that captures transient energy increases, an aspect of acoustic stimuli to which the auditory system is known to be tuned (Fishbach et al., 2001; Howard & Poeppel, 2010). The first 500 ms of EEG and acoustic signal from each trial was discarded to minimize the effect of strong responses to the sudden onset of sound. The speech envelope dynamics signal was then cross-correlated with each channel of the time-aligned EEG data to arrive at a cross-correlation function, which reflects electrical activity phase-locked to the acoustic dynamics of that speech signal.

The frequency content of the phase-locked neural activity captured by the cross-correlation function was analyzed by a wavelet decomposition for a range of cross-correlation lags [-300 700] ms. The evoked power was calculated as the power in the trial-averaged cross-correlation function, normalized by the power in the [-300 -100] ms lag epoch in which the EEG signal is presumed to precede the speech signal.

3.3 Results

3.3.1 Behavioral data

Listener's ability to recall words from the speech stream was reduced by the addition of a noise masker to the acoustic scene (Figure 3.2). A 2x3 mixed ANOVA with mask location (colocalized, spatially separated) as a between subject factor and mask type (unmasked, complementary mask, flat mask) as a within-subject factor revealed a significant main effect of mask type ($F(1.64,75.63)=251.1$, $p<0.001$, $\eta^2=0.85$, Greenhouse-Geisser adjusted, $\varepsilon=0.82$), a significant effect of mask location ($F(1,46)=23.10$, $p<0.001$, $\eta^2=0.33$), as well as a significant interaction between mask type and location ($F(1.64,75.63)=17.77$, $p<0.001$, $\eta^2=0.28$, Greenhouse-Geisser adjusted, $\varepsilon=0.82$). Analysis of the simple main effects identified significant effects of mask type for both the colocalized speech-masker group ($F(2,45)=151.2$, $p<0.001$ Benjamini-Hochberg adjusted, $\eta^2=0.87$) and the spatially separated speech-masker group ($F(2,45)=56.86$, $p<0.001$ Benjamini-Hochberg adjusted, $\eta^2=0.72$). The analysis also identified a significant increase in word recall rate when the speech sample and masker were spatially separated in both the complementary mask ($F(1,46)=28.15$, $p<0.001$ Benjamini-Hochberg adjusted, $\eta^2=0.38$) and flat mask ($F(1,46)=58.78$, $p<0.001$ Benjamini-Hochberg adjusted, $\eta^2=0.56$) conditions. As there was no masker present in the *Unmasked* condition, the comparison between location groups in Figure 3.2 simply reflects the between groups comparison given identical unmasked stimuli ($F(1,46)=2.36$, $p=0.13$ Benjamini-Hochberg adjusted, $\eta^2=0.049$).

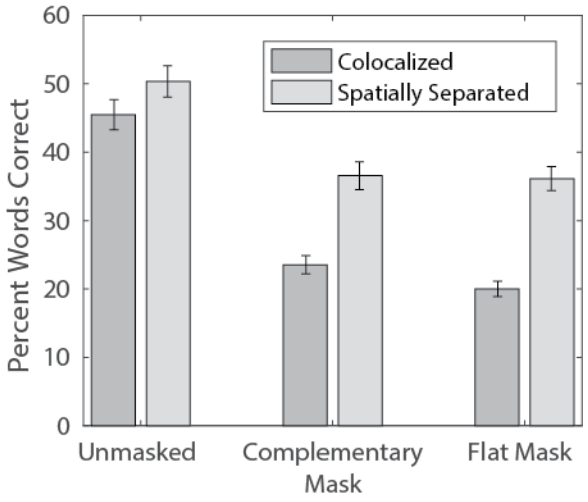

Figure 3.2: Behavioral results for the listening task for colocalized and spatially separated speech and noise maskers. Error bars indicate standard error.

3.3.2 EEG results

Cross-correlation of the first-derivative of the speech envelope with recorded EEG revealed robust neural tracking of speech, even in the presence of masking noise that obscured the low-level amplitude envelope (Figure 3.3). The correlation between speech dynamics and EEG activity remained consistent whether speech was presented alone as in the *Unmasked* condition, or whether the speech was presented with a noise masker; examination of the cross-correlation function revealed that the addition of noise to the scene produced robust speech-locked activity at later lags [200 350] ms which had a similar scalp-topographical distribution as the earlier activity seen across all conditions (Figure 3.3, bottom).

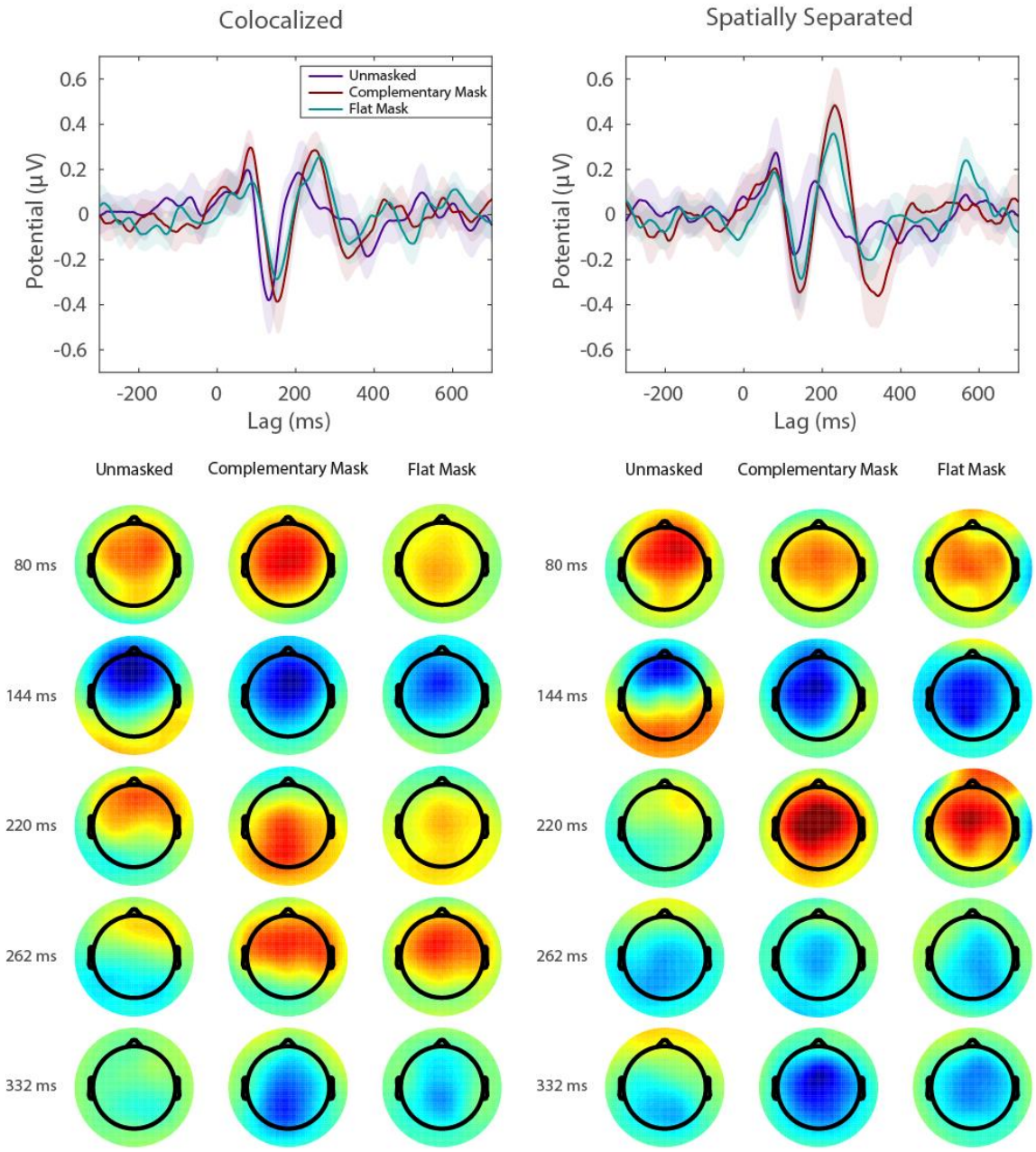

Figure 3.3: Cross-correlation of the EEG and the first-derivative of the speech envelope at a representative electrode, FCz (top), and scalp topographies (bottom) corresponding to local peaks in the cross-correlation function when masking noise was colocalized (left) and spatially separated (right) relative to target speech. Shaded area indicates 95% confidence interval.

Time-frequency decomposition of the cross-correlation function revealed that speech-locked theta-band EEG activity increased relative to baseline. For the colocalized speech-masker group there was a significant increase from baseline in speech-locked theta power in all conditions for a range of latencies (See Figure 3.4; one-tailed *t*-test, Benjamini-Hochberg FDR corrected, $p<0.05$). A 2x3 mixed ANOVA with mask location (colocalized, spatially separated) as a between-subject factor and mask type (unmasked, complementary mask, flat mask) as a within-subject factor reveals that mask type has a significant effect ($F(2,92)=9.66$, $p<0.001$, $\eta^2=0.17$) on the latency of the maximum speech-locked theta-band power; there was no effect of mask location ($F(1,46)=0.21$, $p=0.81$, $\eta^2=0.004$). Post hoc pairwise comparisons confirm that peak theta power occurs later for the masked conditions (242 ± 19 ms *Complementary* Mask; 256 ± 18 ms *Flat* Mask) than the unmasked (161 ± 19 ms) condition ($p<0.004$, Bonferroni corrected) while there was no difference between masking conditions ($p=1$, Bonferroni corrected). This shift in the latency of maximum speech-locked theta power was due to the component in the cross-correlation function that was present at later lags. This component occurred only for speech presented simultaneously with noise, and likely reflects a stream segregation or attention processes.

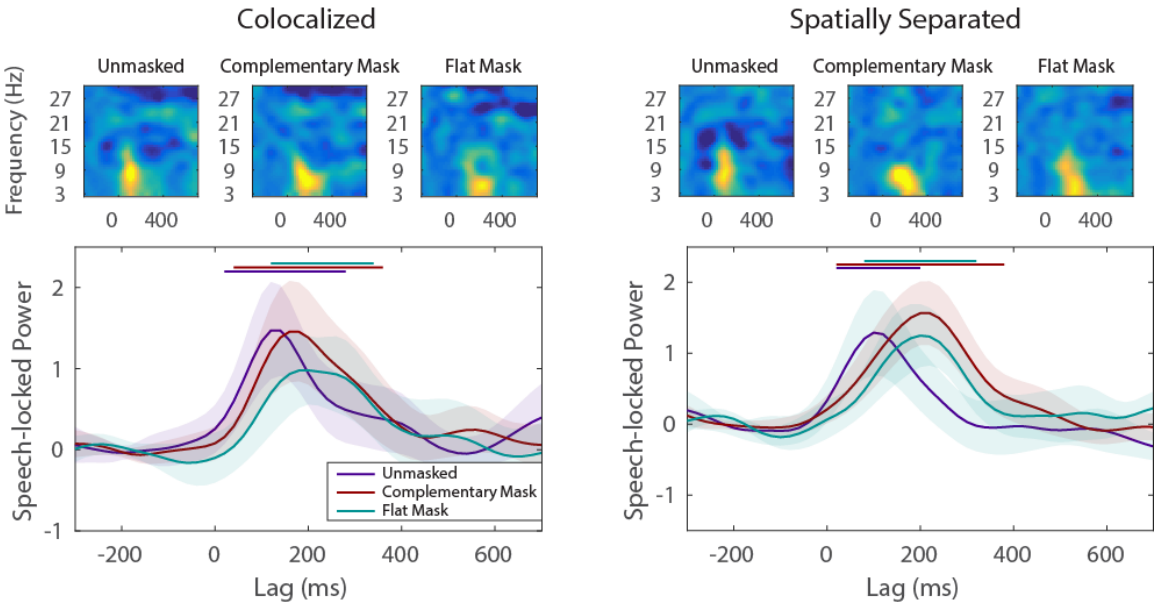

Figure 3.4: (Top) Speech-locked evoked power in the speech-EEG cross-correlation function for masked and unmasked speech. (Bottom) Speech-locked theta (4-8 Hz) cross-correlation power. Horizontal bars indicate latencies at which speech-locked theta power was significantly greater than during the [-300 -100] ms lag baseline (one-tailed *t*-test, P<0.05, Benjamini-Hochberg adjusted). Shaded area indicates 95% confidence interval.

3.4 Discussion

Our behavioral results show a robust effect of noise-maskers on the word recall rate of speech. Listeners could recall fewer words when speech was presented simultaneously with envelope modulated and unmodulated noise. Spatially separating the noise masker from the speech somewhat reduced the effect of the masker, consistent with the effect of spatial release from masking described in speech intelligibility experiments (Arbogast et al., 2002).

The cross-correlation of the EEG with the speech envelope dynamics showed that the brain tracks periodic features of speech despite the presence of noise that obscures

the broadband speech envelope. This result strongly suggests that intact broadband speech envelope dynamics is not a necessary feature that enables the neural tracking of speech. Instead, higher-level features such as modulations within discrete frequency bands and/or linguistic features with fuzzily defined acoustic attributes must provide a sufficient basis for the brain's ability to track the speech signal. The fact that speech tracking is unaffected by the spatial location of the noise masker suggests that the observed neural tracking of speech reflects an object-based tracking of the speech signal after it has been separated from other competing sounds in the scene, rather than simple monitoring and integration of the acoustic energy in the scene.

Robustness of cortical tracking of speech-in-noise could be explained in part by a contrast gain control mechanism (Ding & Simon, 2013). The addition of noise to the acoustic scene effectively compresses the dynamic range of speech in both broadband (i.e. the acoustic envelope) and narrow-band frequency ranges (See Figure 4.1). Such a passive mechanism, in which the sensitivity of the auditory periphery is adapted based on the statistics of the acoustic scene, does not necessarily entail cortical involvement and can be observed in anaesthetized animals (Dean, Harper, & McAlpine, 2005; Rabinowitz, Willmore, Schnupp, & King, 2011; Wen, Wang, Dean, & Delgutte, 2009). However, passive contrast gain control fails to explain the difference between the tracking of natural and vocoded speech-in-noise observed by Ding et al. (2014). They found that, while background noise reduced the intensity contrast for both speech types (which would be compensated for by a contrast gain control mechanism), only the tracking of vocoded speech-in-noise is reduced, which suggests that the observed

cortical tracking mechanism is sensitive to spectrotemporal fine-structure cues that are not present in vocoded speech. Thus, contrast gain control is not a sufficient mechanistic explanation for the observed neural tracking of speech-in-noise.

Our results contrast in interesting ways with those of a related study by Zoefel and VanRullen (2016). They employed a cross-correlation procedure using acoustically similar speech samples as background sound in a tone-detection task. They reported differences between the brain's response to natural speech and speech-in-noise: the cross-correlation function to speech-in-noise lacked a peak between 100-150 ms lag and was attenuated relative to the cross-correlation function to natural speech. We found no such differences in our results; the cross-correlation function to speech was similar across masking conditions for lags ≤200 ms. We attribute the difference in results to the established effect of attention on the neural tracking of speech. Previous studies have found that attending to a speech stream in the presence of distractors enhances the neural synchronization to that speech stream even in the presence of competing sounds (Ding & Simon, 2012a; Hambrook & Tata, 2014; Kerlin et al., 2010; Mesgarani & Chang, 2012; Zion Golumbic, Ding, et al., 2013). Furthermore, Baltzell et al. (2017) reported an effect of task demands on speech tracking even when speech was presented in isolation. We therefore propose that active listening to speech among competing acoustic streams engages higher-level linguistic and attentional mechanisms that make the tracking of periodic features more robust.

The speech-EEG cross-correlation function for the masked conditions showed a speech-phase-locked response at lags >200 ms which we propose reflects a stream segregation and attentional selection process in which the speech stream is separated from the background noise. This segregation mechanism seems not to be dependent on spatial separation, as the co-localized and spatially separated groups did not substantially differ in our study, though we note a study focused on distinguishing the two with more experimental power may be able to identify a difference. Other experiments using a two-talker paradigm have reported robust activity at lags >200 ms phase-locked to attended speech (O'Sullivan, Power, et al., 2015; Power et al., 2012). O'Sullivan et al. (2015b) reported a similar effect in the neural response to stochastic figure-ground stimuli. Stochastic figure-ground stimuli contain a "figure" consisting of a series of tones that emerge from a background of unrelated tones based on consistent temporal coherence of frequency components over time. They found that actively listening to the acoustic stimuli evoked significantly greater phase-locked activity peaking at around 210 ms, consistent with the cross-correlation component peaking around from 220-260 ms lag in our results. Importantly, the topography of this late-latency activity appears to be consistent across studies and is itself consistent with neural sources in temporal cortex, suggesting that it may be a signature of a general sound segregation and attention mechanism. O'Sullivan et al. (2015b) suggest activity at these latencies reflects a general mechanism for segregating acoustic streams on the basis of their temporal coherence.

Pre-attentive stream segregation appears to be reflected by early latency cortical activity phase-locked to temporally coherent acoustic streams that occurs even during passive listening (O'Sullivan, Shamma, et al., 2015; Zoefel & VanRullen, 2016). The effects of selective attention and stream segregation emerge at latencies around 200 ms, though it remains unclear what this activity represents. It may be the case that this later-latency activity represents a continuation of stimulus-driven activity. Alternatively, we propose that this activity reflects a response to top-down feedback activity, potentially originating in frontal and motor areas that have also been shown to track speech (Ding, Melloni, et al., 2016; Park et al., 2015; Zion Golumbic, Ding, et al., 2013). Such feedback signals may encode predictions about upcoming stimulus features in the quasi-periodic speech stream or represent the response of error encoding predictive neural units (Feldman & Friston, 2010). Further research examining the causal relationship between speech-tracking activity in auditory areas and higher-order brain areas is called for to identify the mechanistic underpinnings of attentional effects on the speech-tracking response.

Our results show that the neural tracking of speech does not rely solely on tracking the broadband acoustic envelope. We found that the acoustic dynamics of speech are tracked even when loud background noise eliminates broadband amplitude modulations in the acoustic scene. Two higher-level features in the speech stream may provide the periodic cues that enable tracking: it is possible that the brain tracks fast modulations within discrete frequency bands (Ghitza et al., 2013). Likewise, higher-level linguistic or grammatical features may allow the brain to temporally align oscillatory

activity with linguistic features even in the absence of periodic acoustic cues. Indeed, the observed entrainment between EEG oscillations and speech features across many studies might not be driven entirely by bottom-up acoustic features. Instead, entrainment between brain oscillatory activity and quasi-periodic speech features might reflect temporal coherence between top-down signaling among speech-related cortical regions, or coherence between top-down and bottom-up afferent signals. For example, predictive coding is proposed as a computational mechanism for auditory perception (Bastos et al., 2012; Bendixen, 2014; Gagnepain, Henson, & Davis, 2012; Winkler, Denham, Mill, Bohm, & Bendixen, 2012). In this theory, predictions about afferent features are projected down to lower levels. For dynamic stimuli such as speech, there must be coherence between predicted features and bottom-up evidence. The fact that familiarity with the grammatical or lexical structure of the language (Ding, Melloni, et al., 2016), and the present result that robust speech tracking can occur in the absence of a discrete broadband envelope suggests that this phenomenon reflects the temporal coherence called for by a top-down dynamic prediction mechanism.

4 The Effects of Periodic Interruptions on Cortical Entrainment to Speech

4.1 Introduction

Speech sounds like a smooth stream of words separated by gaps but, acoustically speech consists of periodic bursts of acoustic energy interleaved with silences that do not necessarily correspond to word boundaries. This is evident when hearing a foreign language: we easily recognize that the sound we hear is speech, but instead of segmented words we hear only staccato bursts of sound without clear word boundaries. Unfamiliar speech sounds like this in part because many languages share a quasi-periodic 5-hz amplitude envelope corresponding to the syllable rate (Chandrasekaran et al., 2009; Poeppel, 2003). In fact, speech is a temporally and spectrally complex acoustic signal modulated at several time scales which also include high frequency modulations (30-50 Hz) corresponding to phonemic features and a lower-frequency intonation contour (1-2 Hz) (Chait, Greenberg, Arai, Simon, & Poeppel, 2015; Ghitza & Greenberg, 2009; Giraud & Poeppel, 2012).

Although the brain is remarkably good at stitching these dynamic acoustic events together into a coherent stream of words, speech perception mechanisms can be disrupted. One such disruption occurs when brief segments of speech are replaced with silent gaps –also known as the "picket fence" effect. In this case, speech processing mechanisms fail to recover the content of the interrupted speech. It is possible to restore perception somewhat by filling these gaps with broadband noise (Miller & Licklider, 1950; Warren, 1970). This is called *phonemic restoration* and it depends on several factors: the intensity of the noise bursts, with louder noise being more effective; spectral overlap

between the noise and speech (Bashford & Warren, 1987); the proportion of speech occluded by interruptions (Bashford, Riener, & Warren, 1992; Cooke, 2003; X. Wang & Humes, 2010); linguistic context (X. Wang & Humes, 2010); and agreement with visual cues (Shahin, Bishop, & Miller, 2009). Importantly, speech perception is more resilient to interruptions when the envelope dynamics are preserved (Bashford, Warren, & Brown, 1996; Başkent, Eiler, & Edwards, 2009; Fogerty, 2013; Fogerty & Humes, 2012; Fogerty, Kewley-Port, & Humes, 2012; Gilbert, Bergeras, Voillery, & Lorenzi, 2007; Shinn-Cunningham & Wang, 2008), which suggests that low-frequency envelope modulations are not merely epiphenomena, but rather encode information that can be used by the brain to facilitate speech perception.

The phase of low-frequency (4-8 Hz) neuroelectric oscillations tracks modulations in the acoustic envelope. This phase-tracking of speech has been associated with improved intelligibility of degraded speech and improved effectiveness of selective attention in studies employing electroencephalography (EEG) (Ding & Simon, 2009; Hambrook & Tata, 2014; Kerlin et al., 2010; Peelle & Davis, 2012), electrocorticography (ECoG) (Mesgarani & Chang, 2012; Zion Golumbic, Ding, et al., 2013), and magnetoencephalography (MEG) (Cogan & Poeppel, 2011; Ding & Simon, 2012a; Doelling et al., 2014). Studies using advanced statistical techniques have found that the brain also tracks speech features beyond the broadband envelope (Di Liberto et al., 2015; Mesgarani et al., 2014). Converging lines of evidence from studies using perceptual entrainment paradigms (Hickok, Farahbod, & Saberi, 2015; Zoefel & VanRullen, 2015a) and periodically modulated electrical stimulation (Riecke, Formisano, Herrmann, &

Sack, 2015) further suggest that neural entrainment provides perceptual benefits. These results have led to the theory that low-frequency neural oscillations play a computational role in optimally parsing speech (Ghitza, 2011; Giraud & Poeppel, 2012). The goal of the present study was to investigate the importance of the speech phase-tracking phenomenon in the neural mechanisms that restore removed linguistic information in the picket-fence effect.

Entraining oscillatory activity to temporal modulations in speech may connect low-level acoustic representations to brain-wide speech processing networks (Fries, 2005; Schroeder & Lakatos, 2009). Thus, degradation of speech perception in the picket-fence effect may result not only from removal of phonemic information, but also because interruptions introduce sharp acoustic transients that do not align with real syllable boundaries. These transients probably disrupt neural speech tracking. In this theory, phonemic restoration occurs because the continuity of the speech envelope is restored. Two predictions follow: first, that inserting gaps into continuous speech degrades EEG speech tracking. Second, that filling those gaps with carefully modulated noise that restores the acoustic envelope (Bashford et al., 1996; Fogerty & Humes, 2012; Shinn-Cunningham & Wang, 2008) will restore speech-related brain responses along with speech perception. In the present study we show that phonemic restoration is facilitated by restoring speech envelope dynamics, which in turn restores the dynamics of cortical speech-processing networks.

4.2 Methods

4.2.1 Subjects

Twenty undergraduates (12 female; 2 left-handed; mean age 19.5 years) from the University of Lethbridge were recruited and participated for course credit. Participants were neurologically normal and reported normal hearing. All participants provided informed written consent and procedures were in accordance with the Declaration of Helsinki and were approved by the University of Lethbridge Human Subjects Review Committee.

4.2.2 Presentation

All stimuli were presented in free-field by an Apple iMac with a firewire audio interface (M-Audio Firewire 410). Participants sat 1 meter from a near-field studio monitor (Mackie HR624 MK-2) located on the front auditory midline. Stimulus presentation was controlled by a script custom coded using MATLAB and Psychophysics Toolbox Version 3 (Brainard, 1997; Pelli, 1997).

4.2.3 Stimuli

The stimuli consisted of 60 speech samples of continuous speech generated from the Pacific Northwest/Northern Cities (PN/NC) corpus (McCloy et al., 2013). The PN/NC corpus consists of recordings of male and female speakers reading 180 sentences from the IEEE "Harvard" set and their time-aligned phonetic transcripts (“IEEE Recommended Practice for Speech Quality Measurements,” 1969). Each speech sample contained three unrelated sentences, read by one of two male voices, concatenated

together to create a sample of speech ~6.5s long. Each speech sample contained 20-27 total words (mean: 22.93 ± 1.8). Each individual sentence was presented twice for each participant: once read by each speaker used with non-identical partner sentences to create a unique sentence triplet. Four versions of each speech segment, corresponding to the four experimental conditions, were created: *Original* speech segments consisted of uninterrupted continuous speech, *Gap* speech segments were generated based on original speech segments that had been interrupted by 166 ms silences inserted every 333 ms with 40 ms of jitter, *Burst* speech segments were created from *Gap* speech segments in which the silent periods were filled with loud (+4 dB relative to average speech level) bursts of spectrally matched, uniform intensity noise. Previous studies suggest that phonemic restorations are more likely to occur if the masking noise shares spectral similarities with the interrupted speech, therefore we used noise samples that matched the spectral properties of the original speech (Bashford & Warren, 1987). Spectrally matched noise was generated by randomly time-shifting each original speech segment and adding the resulting signals together; this process was repeated 10 000 times and the resulting signals were combined and scaled to match the original average RMS amplitude. This procedure resulted in stationary noise which matched the average periodogram of the original speech samples. Finally, *Smooth* speech segments were created from *Gap* speech segments in which the silent periods were filled with spectrally matched noise that had been scaled by the low-pass (<25 Hz) filtered acoustic envelope of the original speech (Figure 4.1). This procedure resulted in speech samples that had

been interrupted yet retained the same low-frequency amplitude dynamics as the original speech.

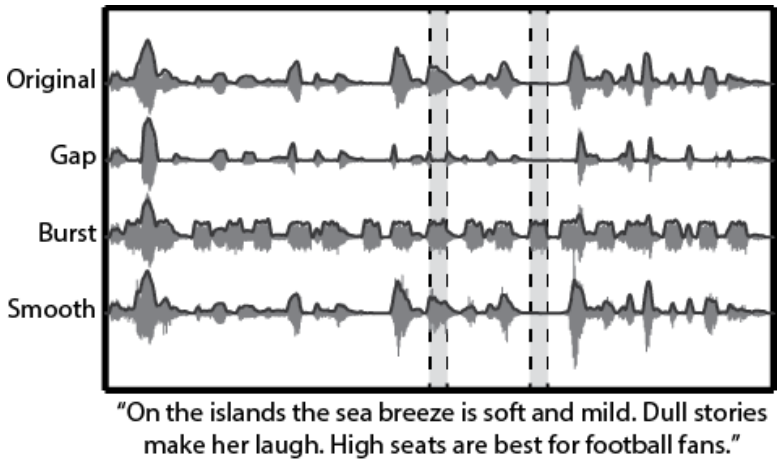

Figure 4.1: Example speech waveform and envelope for *Original*, *Gap*, *Burst*, and *Smooth* speech conditions. Both the *Gap* and *Burst* conditions substantially alter the envelope of the original speech. The envelope is restored in the *Smooth* condition. Highlighted regions show the effect of adding gaps or bursts to epochs within syllables (left vertical column) and between syllables (right vertical column).

4.2.4 Experimental paradigm

Each participant completed 80 trials, 20 in each condition, pseudorandomly ordered, with one break after 40 trials. Each trial consisted of the presentation of a single speech sample after which participants were given 30 seconds to recall and type as many of the words from the speech sample as possible. Performance on each trial was scored as a proportion of correctly recalled words divided by the total of number of words in the speech sample excluding common articles "the", "a", and "an'".

4.2.5 EEG recording and analysis

EEG was recorded with 128 Ag/Ag-Cl electrodes in an elastic net (Electrical Geodesics Inc., Eugene, OR, USA). Scalp voltages were recorded at a 500 Hz sampling rate and impedances were maintained under 100 kilo-ohms. Data were first analyzed using the BESA software package (Megis Software 5.3, Grafelfing, Germany). Data were visually inspected for bad channels and the signal from a small number of electrodes (10 or less) was replaced with an interpolated signal. Because of the length of the trials, eye movement artifacts occurred in a majority of trials, therefore eye movement artifacts were corrected using the adaptive artifact correction algorithm (Ille et al., 2002). The EEG data from three subjects was not included in the final analysis: one subject had >10 identified bad channels, and two subjects had large non-eye movement artifacts that could not be corrected. Data were re-referenced to an average reference, interpolated to an 81-channel 10-10 montage, digitally filtered between 1-15 Hz, and exported to MATLAB (MATLAB version 8.3.0.532; The Mathworks Inc., 2014, Natick, Massachusetts, USA) where further analysis was performed using custom scripts, EEGLAB functions, and the mTRF toolbox (Crosse, Di Liberto, Bednar, & Lalor, 2016; Delorme & Makeig, 2004).

Brain activity related to speech processing was isolated using linear regression to determine multivariate temporal response functions (mTRFs) which describe a mapping between the EEG and three different representations of the original, uninterrupted speech signal. 1) The envelope dynamics representation was calculated by taking the absolute value of the Hilbert transform of the signal to extract the acoustic envelope,

low-pass filtering the envelope at 25 Hz, down-sampling the envelope to match the sample rate of the EEG data, then taking the first-derivative of the signal and half-wave rectifying it to create a signal that captured the low-frequency dynamics of the speech signal. This first-derivative envelope dynamics representation is preferred to the envelope itself because the auditory system is tuned to transient changes in sound captured by the first-derivative of the envelope (Doelling et al., 2014; Fishbach et al., 2001; Hambrook & Tata, 2014; Hertrich et al., 2012; Howard & Poeppel, 2010). 2) The spectrogram representation was computed at 19 bark-scale frequencies using the VOICEBOX toolbox (Brookes, 2003). 3) The phonetic features representation was obtained by mapping the phonetic transcript of speech samples in the PN/NC corpus onto a space of 19 articulatory-acoustic features (Di Liberto et al., 2015; Mesgarani et al., 2014). For the three interruption conditions the same original speech representations were used for generating TRFs and reconstructing EEG data. This choice relates to the fundamental hypothesis of our study: that restored perception of interrupted speech is supported by neural mechanisms active during those interrupted segments. We therefore sought to measure the extent to which the normal neural mechanisms of speech perception engaged during uninterrupted speech were abolished or restored in the three interruption conditions.

The relationship between a speech representation and it's encoding in the EEG can be quantified using a model-based approach. Essentially, an mTRF and the speech representation for a given trial are combined to predict the EEG signal on that given trial. The accuracy of that prediction, measured and reported here as a Pearson correlation

coefficient between the predicted and measured EEG signals, reflects the degree to which the EEG encodes the information contained in the speech representation.

Generic mTRFs for each experimental condition and speech representation were generated based on procedures described by Di Liberto & Lalor (2017). EEG data was down-sampled to 100 Hz and converted to z-scores prior to regression to improve computational efficiency (Crosse et al., 2016). First, mTRFs were generated for each subject using a leave-one-out cross-validation approach in which an mTRF was trained on 19 trials and used to predict EEG signal for the remaining trial. This was repeated 20 times, until all trials had been predicted. This cross-validation procedure was repeated 25 times to empirically tune the ridge regression regularization parameter, λ, across a range of logarithmically spaced values from $[10^{-1}, 10^{6}]$. The regularization parameter is used to optimize the prediction accuracy for each condition-representation combination and is described in detail by Crosse et al. (2016). Second, a subset of 12 symmetrical fronto-central electrodes (FC5, F3, FC3, F1, FC1, Fz, FCz, F2, FC2, F4, FC4, FC6) with the highest correlation coefficients across all conditions were identified and selected for further analysis. Exemplar mTRFs and reported reconstruction accuracies are the average results over these 12 electrodes. Finally, the EEG signal for each subject was predicted by averaging the optimized subject-specific mTRFs from all other subjects to create a "generic" mTRF which is more effective than subject-specific models. All mTRFs were computed using a peri-stimulus time-window of lags ranging from -100 to 400 ms. Thus, the prediction accuracies reported reflect the prediction of single-subject EEG data based on mTRF models trained using the data from all other subjects in the experiment.

For plotting TRF models for different conditions, which may have different amplitudes due to different λ-values used in their generation, TRF weights were normalized by subtracting the mean pre-stimulus baseline and dividing by the standard deviation across the whole time-window.

To further understand the differences in speech-locked brain activity between the four experimental conditions we analyzed the topographic distribution of the temporal response functions derived from the envelope dynamics representation of speech based on the methodology described by Murray et al. (2008). Because different λ parameters were used to arrive at the TRF for each condition, and because the λ-value acts as a smoothing factor that modulates the variance of the TRF across time, comparing global field power across conditions yields results that cannot be readily interpreted, so we instead focus on the global topographic dissimilarity (DISS) between conditions. To calculate the topographic dissimilarity the TRF weights at each electrode are normalized by subtracting the mean TRF weight across all electrodes and dividing by the instantaneous global field power (the standard deviation of TRF weights across all electrodes). From these normalized TRF weights the topographic dissimilarity is computed for a given time lag by computing the square root of the mean of the squared differences between the TRF weights at each electrode.

4.2.6 Statistical analysis

The significance of differences in behavioral performance and TRF-based reconstruction of EEG signals using repeated measures ANOVAs performed in SPSS

(IBM SPSS Statistics version 20.0.0; IBM Corp., 2011, Armonk, New York, USA). Designs for each individual ANOVA can be found in the Results describing Figure 4.2 (for behavioral results) and Figure 4.4 (for EEG reconstruction results). Assumptions of sphericity were assessed using Mauchly's test of sphericity. For factors that violated the assumption of sphericity original and adjusted degrees of freedom are reported along with the *p*-value based on the adjusted degrees of freedom.

The significance of DISS values was assessed using a non-parametric permutation test in which, at the within-subject level, TRF weights were randomly assigned to an experimental condition and a new DISS value was computed based on the reassigned TRF weights; this process was repeated 100 000 times to generate an empirical distribution of DISS values and p-values were assigned based on where the actual DISS ranked within this distribution. The false discovery rate of this analysis was controlled by Bonferroni correcting for the number of pairwise between-condition comparisons (six) and using the Benjamini-Hochberg procedure to account for the comparisons at each computed time lag (51) (Benjamini & Hochberg, 1995).

4.3 Results

Listeners were able to recall fewer words from interrupted speech compared to uninterrupted speech, however restoring the acoustic envelope in the *Smooth* interruption condition remediated the detrimental effect of interruption (Figure 4.2). Note that speech stimuli consisted of three unconnected simple sentences and contained an average of 23 words. At the end of each trial, participants performed a free-recall of

these words. Thus, performance was limited by the working memory capacity of listeners (Cowan, 2001), and we would not expect near-ceiling performance even in the *Uninterrupted* condition. A repeated measures ANOVA on the percentage of words recalled per trial showed a significant main effect of interruption condition (F(3,57; adj:1.9,36)=258.68, p<0.001, η^2=0.93, Greenhouse-Geisser adjusted, ε=0.64). Post hoc pairwise comparisons revealed significant differences in task performance between conditions (p<0.001) with the exception of the *Gap* and *Burst* conditions, though the difference between those conditions trended towards significance (p=0.098).

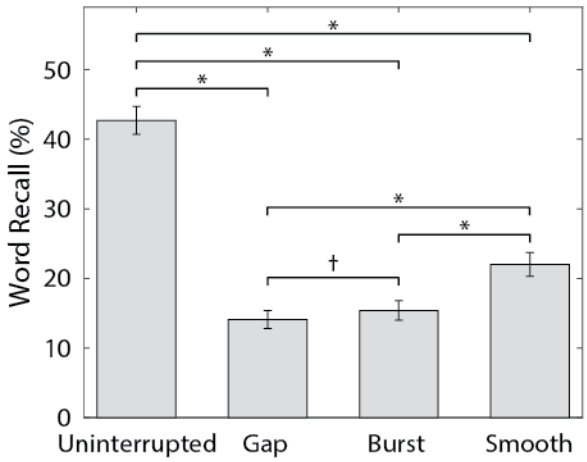

Figure 4.2: Percentage of correctly recalled words from sentences that were uninterrupted, interrupted by silent gaps (*Gap*), interrupted by bursts of noise (*Burst*), or interrupted by noise that followed the original speech envelope (*Smooth*). Interruptions significantly decreased word recall. Word recall was significantly improved by the restoration of the original acoustic envelope in the *Smooth* condition (error bars indicate standard error; † p<0.1; * p<0.0001).

The encoding of speech features by the brain was measured using a forward TRF modelling approach in which EEG data was reconstructed based on optimized TRFs and

representations of the envelope dynamics of speech, the spectrogram of speech, and the phonetic features of speech respectively. Qualitatively, the TRFs for the envelope dynamics and spectrogram representations of speech contained a common peak-trough-peak neural activation pattern occurring at post-stimulus lags from 0 to 300 ms, across all interruption conditions (Figure 4.3). The TRFs for the phonetic feature representation of speech are less straightforward to interpret, possibly due to the relatively limited recording time in the current study compared to previous studies (130 seconds of data per condition in the current study compared to a minimum of 600 seconds of data used by Di Liberto et al. (2017)). Increased recording time can improve the discriminability of phonetic features based on their TRFs (Di Liberto & Lalor, 2017), it is possible that our current data cannot support the generation of distinct TRFs using a phonetic feature representation.

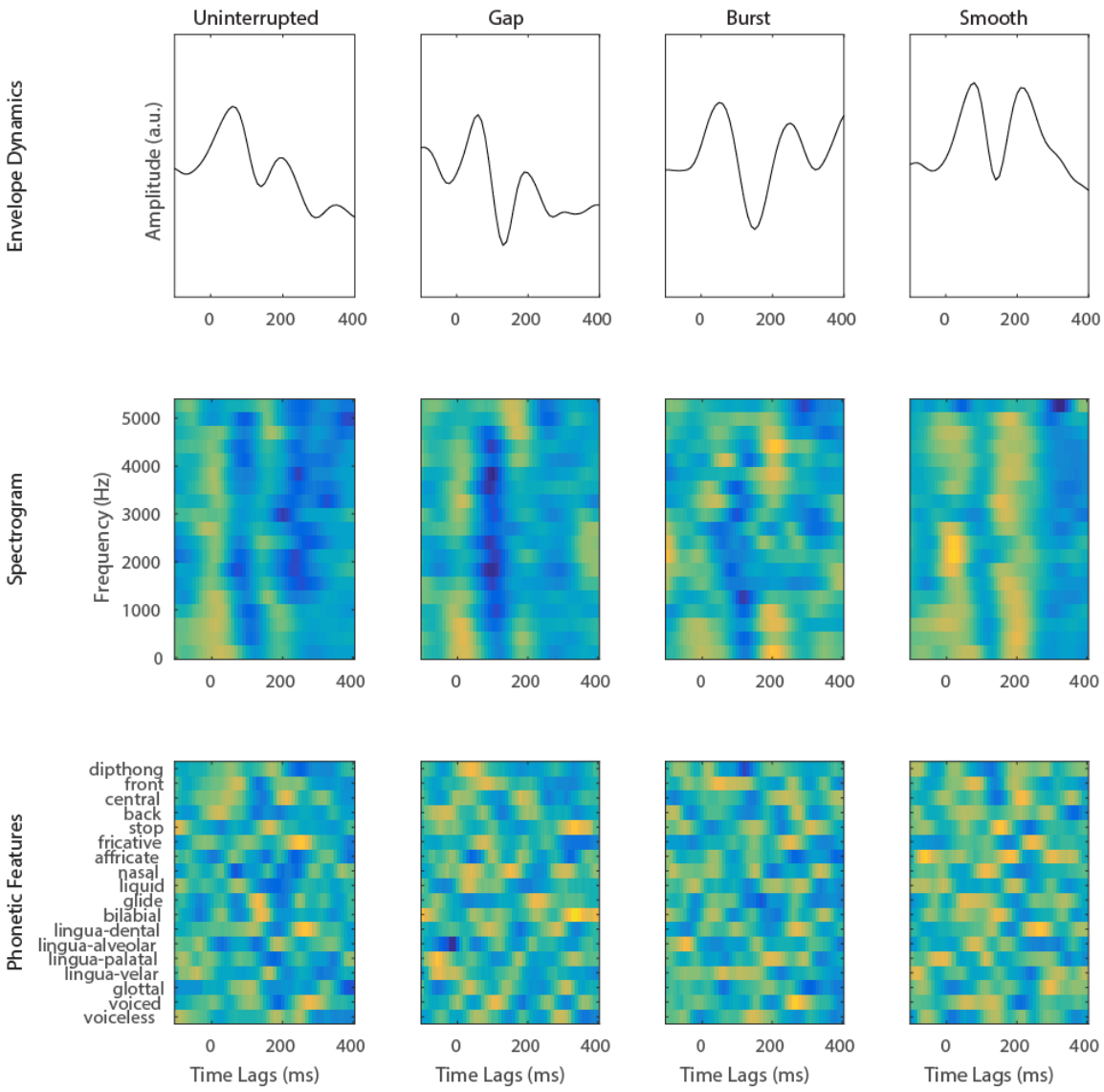

Figure 4.3: Grand-average multivariate temporal response functions (mTRFs) generated for envelope dynamic, spectrogram, and phonetic feature models at peri-stimulus time-lags from -100 to 400 ms for uninterrupted and interrupted speech, averaged over 12 fronto-central electrodes.

How well the brain encodes speech features under each interruption condition is quantifiable by measuring the correlation (Pearson's *r*) between the measured EEG and the EEG signal reconstructed based on the speech representation and the TRF (Figure 4.4A). For all combinations of conditions and speech representations the reconstruction

accuracy was significantly greater than zero (one-sample t-test; t>2.81, p<0.007 Benjamin-Hochberg FDR corrected). A 3x4 repeated measure ANOVA with speech representation (envelope dynamics, spectrogram, phonetic features) and interruption (uninterrupted, gap, burst, smooth) as factors reveals a significant main effect of interruption (F(3,48)=11.12, p<0.001, η^2=0.41), while there was not a significant effect of speech representation (F(2,32; adj:1.5,24)=2.51, p=0.11, η^2=0.13, Greenhouse-Geisser adjusted, ε=0.75) nor a significant interaction between factors (F(6,96)=1.06, p=0.39, η^2=0.062). Post hoc pairwise comparisons of interruption conditions reveals significant differences between the *Uninterrupted* condition and both the *Gap* (p=0.045) and *Burst* (p<0.001) conditions while there was not a significant difference between the *Uninterrupted* and *Smooth* conditions (p=0.61); there were significant differences between the *Smooth* condition and both the *Gap* (p=0.014) and *Burst* (p<0.001) conditions; there was a trend towards significance for the difference between *Gap* and *Burst* conditions (p=0.093). It is worthwhile to note that because we have used the original, uninterrupted speech representations for all conditions these electrophysiological results reflect speech-related responses in the brain, rather than responses to the interruptions. Because the interruptions were not time-locked to any feature of the speech itself the averaged EEG response captured by the mTRF analysis for each condition reflects responses to the speech signal and not the interruptions themselves.

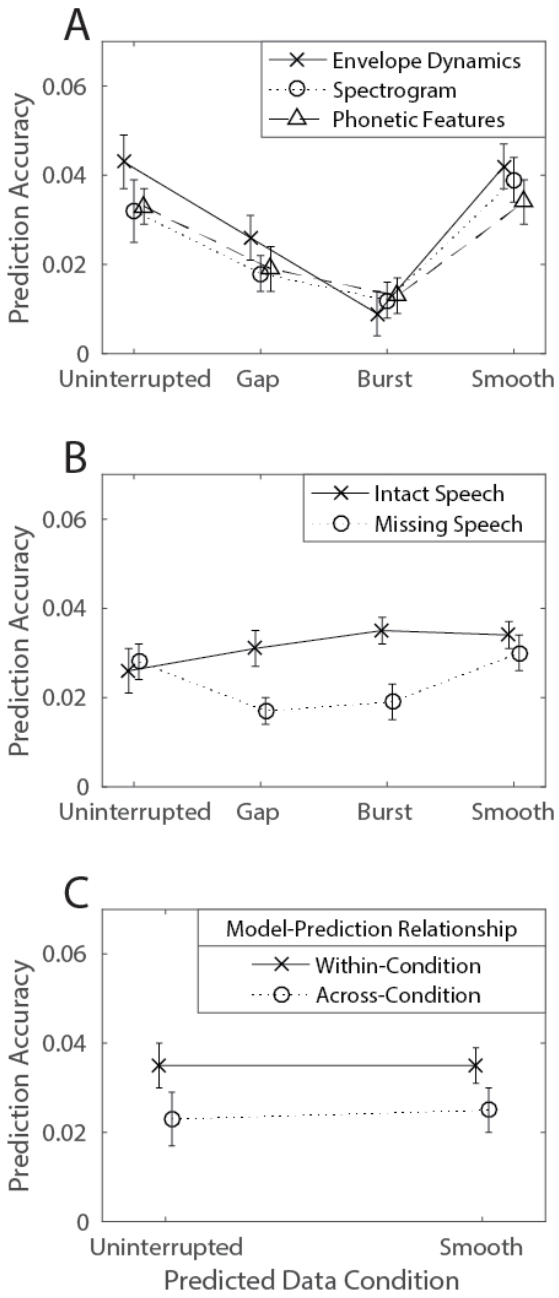

Figure 4.4: A) Grand-average EEG prediction accuracy (Pearson's *r*) for each speech model for uninterrupted and interrupted speech over 12 frontocentral electrodes. Reconstruction of EEG activity is poor for *Gap* and *Burst* interrupted speech, but reconstruction is improved in the *Smooth* condition in which the acoustic envelope is restored. B) Grand-average EEG prediction accuracy for intact and missing segments of speech, collapsed across speech representation. C) Grand-average EEG prediction accuracy for *Uninterrupted* and *Smooth* speech conditions over 12 frontocentral electrodes, collapsed across speech representation, using TRFs trained on data from either *Uninterrupted* or *Smooth* conditions. Error bars indicate standard error.

Disrupting the low-frequency envelope in the *Gap* and *Burst* conditions appears to disrupt both the perception (Figure 4.2) and the cortical representation (Figure 4.4A) of the interrupted speech. Restoring the envelope in the *Smooth* condition appears to restore both perception and the brain's response to the interrupted speech. However, it is possible that interruptions disrupt brain responses globally, across the entire speech sample, or only locally within the missing segments themselves. To assess the extent of such local disruptions we considered the differences in response to intact and missing segments of speech. The mTRF analysis was repeated using speech representations that included data only from the intact or removed portions of the signal from each trial. For the *Uninterrupted* condition the "missing" segments were in fact intact speech that would have been removed in an equivalent interrupted condition (Figure 4.4B). A 2x4 repeated measures ANOVA with speech intactness (intact, missing) and interruption condition (uninterrupted, gap, burst, smooth) collapsed across speech feature models revealed a significant main effect of intactness ($F(1,16)=18.49$, $p<0.001$, $\eta^2=0.54$), no effect of interruption ($F(3,48)=1.02$, $p=0.39$, $\eta^2=0.06$), as well as a significant intact*interruption interaction ($F(3,48)=3.00$, $p<0.001$, $\eta^2=0.27$). Consideration of the simple main effects showed a significant effect of intactness for the *Gap* ($F(1,16)=16.28$, $p=0.002$ Benjamini-Hochberg adjusted, $n^2=0.50$) and *Burst* ($F(1,16)=16.18$, $p=0.002$ Benjamini-Hochberg adjusted, $\eta^2=0.50$) conditions and found no effect of intactness for the *Uninterrupted* or *Smooth* conditions ($F(1,16)<1.10$, $p>0.62$ Benjamini-Hochberg adjusted, $\eta^2<0.064$). Thus, the response to intact speech segments remains similar regardless of interruption type

while the speech information encoded by the EEG is substantially reduced during the missing segments of speech for *Gap* and *Burst*, but not *Smooth* interrupted speech.

The lack of a difference in reconstruction accuracy between the *Uninterrupted* and *Smooth* conditions suggests that the brain responds similarly to uninterrupted speech and interrupted speech with an intact envelope. Restoring the envelope in the *Smooth* condition could restore the "canonical" response to speech; alternatively, reconstruction accuracy might be improved in the *Smooth* condition due to the encoding of some different, yet still speech-locked, set of features. If the speech features encoded by the EEG in both conditions are the same, then we could predict that the TRFs are interchangeable between conditions – that is to say that the EEG from *Uninterrupted* trials can be predicted just as well using TRF models trained on data from *Smooth* trials as TRFs trained based on data from *Uninterrupted* trials and vice versa. To test this hypothesis, we used *Uninterrupted* TRFs to reconstruct the EEG from *Smooth* trials and *Smooth* TRFs to reconstruct *Uninterrupted* data and compared the reconstruction accuracy to the reconstruction accuracy of a purely within-condition reconstruction. A 2x2 repeated measure ANOVA was performed with interruption (uninterrupted, smooth) and TRF model-prediction relationship (within-condition, across-condition) as within-subject factors, collapsed across speech representation, failed to find a significant main effect of interruption ($F(1,16)=0.03$, $p=0.86$, $\eta^2=0.002$), however there was a significant effect of TRF model-prediction relationship ($F(1,16)=13.55$, $p=0.002$, $\eta^2=0.45$) as the across-condition reconstruction accuracy was poorer for both conditions and for all speech representations (Figure 4.4B). There was no significant interaction effect ($F(1,16)=0.035$,

p=0.86, η^2=0.002). This result suggests that, while the degree of speech feature encoding in the EEG signal may be similar across *Uninterrupted* and *Smooth* trials, the specific features encoded in the EEG signal are systematically different.

Topographic analysis of the TRFs based on the envelope dynamics of the speech revealed significant topographic differences between the topography of TRF weights for the *Uninterrupted* and both the *Burst* and *Smooth* interruption conditions which suggests different neural generators are activated in response to speech interrupted by noise (Figure 4.5). The earliest topographic differences occurred for a range of lags from 60-130 ms characterized by a bilateral pattern of relatively stronger TRF weights at frontal electrodes for the interrupted conditions and weaker TRF weights at occipital electrodes. A second difference occurred for a later range of lags from 230-330 ms characterized by a pattern of strongly right-lateralized activity in response to the noise interrupted speech. There were no significant differences found for other pairwise comparisons between conditions.

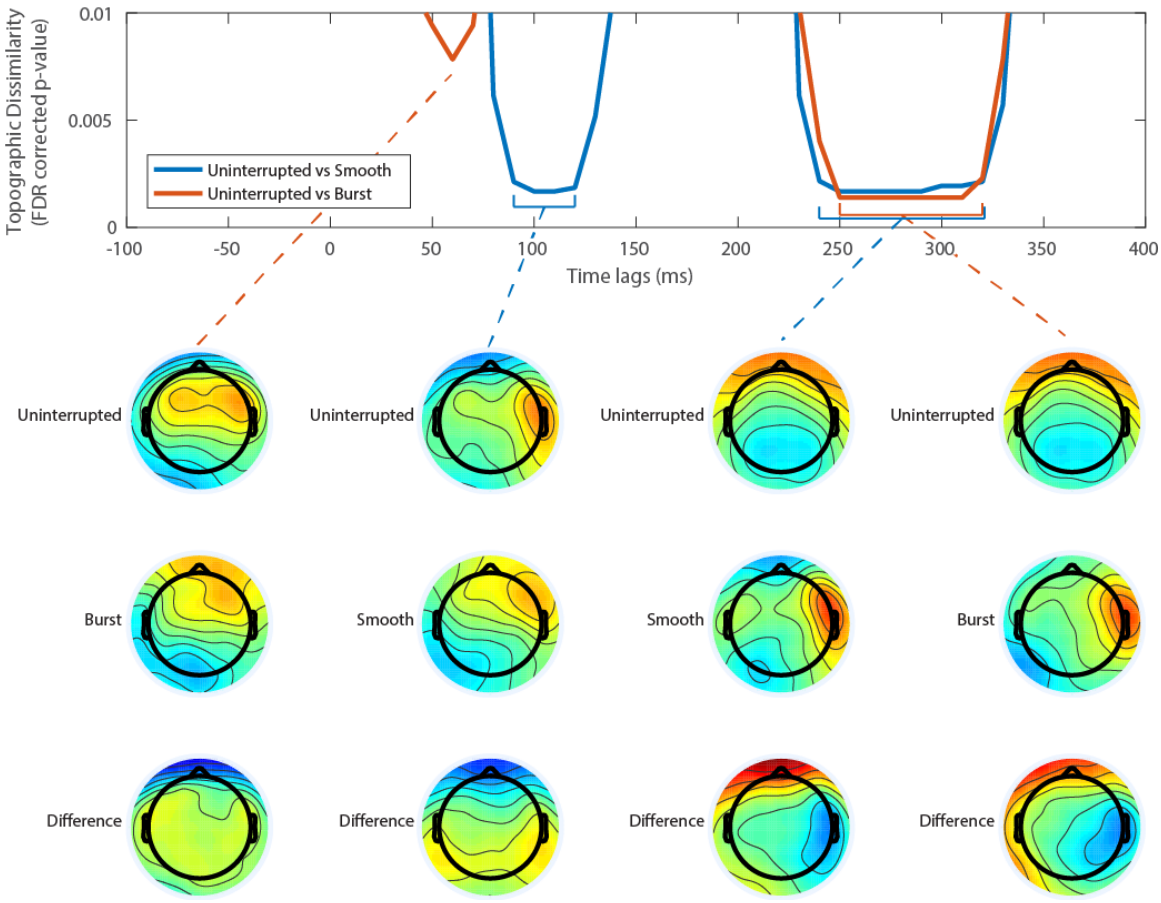

Figure 4.5: Results of topographic dissimilarity analysis for pairwise comparisons of envelope dynamics TRFs. Topographies of TRFs from noise-interrupted conditions significantly differed from the TRF from uninterrupted speech at early lags (60-130 ms) and later lags (230-330 ms).

4.4 Discussion

We found that interrupting speech with brief, repeated gaps of silence impaired perception and subsequent recall. Filling those gaps with noise restored perception to some degree - this is the classical phonemic restoration or "picket fence" effect. Filling those gaps such that the low-frequency acoustic envelope was restored resulted in significantly better recovery of perception. Our EEG results showed some evidence of

correspondence to our behavioral results: The *Gap* and *Burst* conditions showed the worst perceptual performance (Figure 4.2) and had the worst EEG reconstruction accuracy (Figure 4.4A). In the *Smooth* condition, perception was moderately restored; interestingly, EEG reconstruction accuracy was as good as it was for the *Uninterrupted* condition, however further analysis suggested that this does not mean that the original neural response to speech was restored by smoothing the acoustic envelope.

Recent work has demonstrated the importance of the relationship between the acoustic dynamics of speech and the electrical dynamics of the brain as revealed in EEG/MEG studies. We hypothesized that interruptions in speech might impair perception not only by removing phonemic information, but also by destroying the original target envelope of the utterance. Two predictions followed: First, that interruptions that disrupt the speech envelope should disrupt the neural tracking of speech as measured by the accuracy of reconstruction of the EEG signal constructed from the original uninterrupted speech representations. Second, that restoring the appropriate envelope, without restoring the phonemes themselves, should restore both the neural response to speech and speech perception.

Our electrophysiological results do not fully support our first prediction and call for a more nuanced account of speech tracking in the picket-fence effect. If acoustic transients unrelated to syllable boundaries disrupted speech responses across the entire utterance, then we should expect reduced speech tracking during both missing and intact speech segments. Figure 4.4B shows that for speech interrupted by gaps or bursts,

there was a reduction in the amount of speech information encoded in the EEG signal only during those interruptions when the speech signal had been completely removed. Speech information during intact segments was encoded similarly regardless of whether other segments of the utterance had been interrupted. These results suggest that the speech tracking response is not simply a bottom-up, stimulus-driven response to acoustic transients. Intracranial studies of single phonemic restorations have found that the bilateral auditory cortex response to restored phonemes was predicted by pre-interruption activity in left frontal cortex which suggests a top-down mechanism that biases auditory activity to support the perception of continuous speech (Leonard et al., 2016). We speculate that the canonical EEG-speech tracking response to uninterrupted speech represents, to some degree, agreement between bottom-up activity driven by acoustic transients and dynamic top-down activity reflecting an attempt to predict the content of the incoming speech signal. This notion is consistent with models of predictive coding based on intracranial recordings (Leonard, Bouchard, Tang, & Chang, 2015). For the intact speech, in all conditions, the speech tracking response is consistent because the speech signal, and thus the bottom-up signal, is intact and matches the top-down prediction. The initial acoustic transient at the interruption offset in the *Gap* and *Burst* do not affect the response to subsequent intact speech because, even though this transient is a highly salient sound feature in the bottom-up signal sweep, it is not accounted for in the top-down prediction and so the response to that transient is filtered out. During the *Gap* and *Burst* interruptions the acoustic signal and the brain's predictions are mismatched resulting in a reduced speech tracking response.

Our second prediction, that restoring the smooth envelope should restore normal speech-related processes, was partly supported by our electrophysiological results. Restoring the acoustic envelope of speech in the *Smooth* condition appeared to restore the neural encoding of the speech signal even during segments of missing speech. However, TRFs were not interchangeable between uninterrupted and smooth conditions, suggesting that speech features encoded in these two conditions are different when the low-frequency acoustic envelope is restored without the spectral fine-structure of the original phonemes. Furthermore, as shown in Figure 4.5, restoring the low-frequency envelope in the smooth condition lead to a dissimilar topographic distribution of TRF weights when compared to the uninterrupted condition. This suggests that, while speech perception may be restored to some degree, the neural mechanisms mediating that perception are systematically different that those engaged during normal speech perception.

The roles of auditory and speech-specific brain regions in parsing continuous dynamic stimuli have been under investigation and provide some framework for interpreting the present results. For example, neurophysiological studies of the continuity illusion using simple tone stimuli in primates and humans suggest that continuity is reflected physiologically at the level of primary auditory cortex (Petkov, O'Connor, & Sutter, 2007; Riecke, van Opstal, Goebel, & Formisano, 2007). Other studies using speech stimuli interrupted by bursts of noise identified complementary networks that act to repair the interrupted stimulus and maintain the percept of continuity involving, respectively, the left inferior frontal gyrus, left pre-supplementary motor area,

and bilateral insula and; left posterior angular gyrus and superior temporal sulcus, right superior temporal sulcus, bilateral superior frontal sulcus, and precuneus (Shahin et al., 2009). The core area of right Heschl's gyrus (HG) has also been implicated in maintaining perceptual continuity of interrupted stimuli (Riecke et al., 2007; Shahin et al., 2009). Our topographic analysis found a strongly right-lateralized TRF component at lags from 230-330 ms occurring for conditions in which interruptions were filled with noise which could plausibly arise from the activation of right HG. This putative activation of right HG is time-locked to the interrupted speech signal and is the second significant deviation (with the first occurring at lags from 60-130 ms) from the brain activity associated with uninterrupted speech processing, which suggests that it represents a secondary stage of processing mediated by feedback based on the surrounding intact speech signal. Other studies of illusory continuity have found that low-frequency activity related to the interruption onset is suppressed when the illusion is perceived (Kaiser, Senkowski, Roa Romero, Riecke, & Keil, 2018; Riecke, Esposito, Bonte, & Formisano, 2009). While our data do not speak to the brain's response to the interruption, we found evidence for an enhanced or additional response to the interrupted speech itself only when the interruptions are filled with noise that could potentially mask a continuous speech signal.

Our data provide limited insight into the broader question of what relationship exists between speech-tracking neural responses and speech intelligibility. The relationship between speech intelligibility and cortical entrainment to speech remains controversial (Haegens & Zion Golumbic, 2018; Zoefel & VanRullen, 2015b). Recent

experiments have used priming paradigms in which degraded speech is rendered intelligible following the presentation of un-degraded speech and suggest that the linguistic cues present in intelligible speech support spectrotemporal tuning to speech features within auditory cortex (Holdgraf et al., 2016) and increase top-down signaling to auditory areas from association areas (Di Liberto, Lalor, et al., 2018). Our results indirectly support the idea that speech-tracking responses are related to intelligibility based on the correspondence between perceptual performance and our EEG reconstruction results; however, it was not the goal of our study to parameterize intelligibility. The main goal of this study was to characterize the cortical response to interrupted speech dynamics and explore a potential mechanistic explanation for a well-known perceptual illusion. While the correspondence between intelligibility and the EEG speech-tracking response in our results is certainly suggestive of a relationship between cortical entrainment to speech and speech intelligibility, further experiments that exercise tight control over acoustic differences while parameterizing intelligibility in other ways will provide more definitive answers.

Entrainment of brain electrical dynamics to the low-frequency acoustic dynamics in speech has been proposed as a mechanism that allows the brain to rhythmically improve perceptual sensitivity. One theory is that this allows the brain to effectively parse speech into perceptual units. It might also play a role in maintaining selective attention to a single stream (Giraud & Poeppel, 2012; Hambrook & Tata, 2014; Hickok et al., 2015; Riecke et al., 2015; Schroeder & Lakatos, 2009). Those theories suggest a mechanism that aligns temporal windows of enhanced neural sensitivity with important

spectrotemporal events in the speech stream. A related theory, called Communication-through-Coherence, posits that the brain electrical dynamics of disparate regions should be entrained when those regions need to effectively exchange information (Engel et al., 2001; Fries, 2005). Taken together, these ideas suggest that, to optimally process speech, language-processing networks should entrain to the frequency of important spectrotemporal events in speech.

Smoothing the low-frequency amplitude envelope substantially restored perception beyond simply filling gaps with noise. This is remarkable considering that, because only low-frequency modulations were restored, only minimal phonemic-level information was restored to the signal. The *Smooth* condition, by design, retained the low-frequency amplitude cues of normal speech. These might allow the brain to optimally process speech-related spectrotemporal events. This improved processing of sound may aid top-down mechanisms that the brain employs to repair the percept of degraded speech. It might also allow for the more efficient use of contextual information about the sound surrounding the interruptions. Such restorative neural processes are likely to be dynamic, since the speech signal itself is dynamic, and an intact acoustic envelope may provide an important temporal cue that coordinates the dynamical activity between the distributed brain areas responsible for the successful perception of interrupted speech.

5 Conclusions

Oscillatory electrical activity in the brain entrains to rhythmic fluctuations of sensory inputs. The computational purpose of coupling internal, neural activity to external stimulus dynamics is not yet understood. The entrainment of cortical activity to acoustic and abstract features of speech has been proposed as a physiological mechanism involved in processing continuous speech. This thesis has described three studies that consider top-down, cognitive effects on the entrainment of auditory cortical activity to speech. We found that entrainment to speech cannot be explained as a simple bottom-up stimulus-response; it also reflects top-down influences of attentional and linguistic processing mechanisms. These studies contribute valuable insights into how the brain processes speech - especially in real-world, noisy environments.

In Chapter 2 we considered the cognitive process of selective attention in relation to the EEG speech-tracking response. We recreated an ecologically valid "cocktail party" environment in which listeners had to attend and respond to a single target stream against a background of up to six other competing speakers. Behaviorally, listeners were less likely to report words from a target speech stream as the number of distractor voices in the scene was increased. They were also more likely to report words occurring in the distractor streams suggesting that the informational content of distractors actively interfered with perception of the target speech. Our electrophysiological results showed enhanced tracking of attended versus ignored speech but only during epochs surrounding correct responses to target words in the attended speech, in agreement with existing literature (Hambrook & Tata, 2014; Mesgarani & Chang, 2012). We tested three

hypotheses: (a) Speech-tracking activity related to the target speech stream is reduced as the attentional challenge is increased by increasing the distractor set-size. As predicted, the attentional enhancement effect was reduced as more distractors were added to the acoustic scene. (b) Spectral overlap in the auditory periphery results in increased energetic masking, thus speech-tracking is reduced due to a degraded low-level stimulus representation. We invalidate this hypothesis based on the observation that tracking of the target speech stream was also reduced when Interference in the auditory periphery was ruled out by spectrally separating the target from distractor streams in the vocoded speech group. (c) We explicitly tested a theory of "active distraction" which suggests that distractor streams may intrude on perception by transiently capturing attention, however we found no evidence to suggest that such a mechanism occurs in our experimental paradigm.

The main result from Chapter 2 is that the attentional enhancement of entrainment to target speech is reduced as the distractor set-size in the acoustic scene is increased. This reduction in attentional effect cannot be explained solely based on interference in the auditory periphery: the tracking of target speech was also reduced by adding distractors when target and distractor speech was spectrally separated through a vocoding process. Thus, we can conclude that interference between competing speech streams occurs in cortex. We further speculate, based on reports of reduced interference within the spike-train representations of spectrally separated sounds in auditory cortex (Larson et al., 2012; Narayan et al., 2007), that increased interference within higher-order, potentially speech-specific, cortical areas drives the reduction in attentional

enhancement of speech-tracking responses in auditory areas with increasing distractor set-size. We believe that the results presented in Chapter 2 strongly suggest that neural networks responsible for the more general cognitive tasks of attentional selection and language processing exert some top-down influence on auditory cortical activity to enhance the tracking of a target speech signal.

In Chapter 3 we examined the role of the broadband acoustic envelope in the cortical entrainment to speech. One possible explanation for the effect observed in Chapter 2 is that broadband acoustic envelope cues related to the target speech are obscured by the addition of more speech signals to the scene. If broadband envelope fluctuations are important for tracking speech we should expect that systematically removing broadband envelope cues from the scene will likewise abolish cortical entrainment to speech. We used temporally modulated, spectrally matched noise to obscure the broadband speech envelope and found that the broadband envelope modulations were not necessary for speech tracking to occur. Thus, higher-level acoustic modulations (such as energy fluctuations within discrete frequency bands) and linguistic features are taken to be sufficient for entraining cortical activity to speech. This finding reinforces our interpretation of the results presented in Chapter 2, that competing speech interferes with the representation of higher-level features of attended speech in cortex. Furthermore, we identify a component in the EEG speech-tracking response that appears to be related to the selection of a speech stream in the presence of background noise. This component appears to be analogous to a component observed in experiments where listeners had to selectively listen to one speech stream in a two-talker dichotic listening

paradigm (Power et al., 2012), and a component related to separating a salient tone-sequence from a background of other tones (O'Sullivan, Shamma, et al., 2015). We propose that this component may reflect activity related to the cognitive processes of selective attention and stream segregation. We speculated that this component may reflect a response to feedback from higher-order neural ensembles generating top-down predictions about the nature of the target acoustic stream.

In Chapter 4 we investigated the brain's response to acoustic and linguistic features of interrupted speech, seeking a mechanistic explanation for the phonemic restoration effect. The study described in Chapter 3 suggests a speech-tracking mechanism that operates by entraining activity to the high-level spectrotemporal transients that occur quasi-periodically in normal speech. We had predicted that interrupting speech by replacing segments of speech with silent gaps or bursts of noise would disrupt speech perception by disrupting such a neural mechanism – the transients that mark the on-off cycle of the interruptions introduce spectrotemporal boundaries unrelated to the temporal organization of the target speech. Thus, we predicted that the tracking of interrupted speech would be reduced, relative to intact speech. We further predicted that restoring the acoustic envelope of interrupted speech by inserting envelope-modulated noise into gaps would restore both perception of the interrupted speech and neural entrainment to speech. Our first prediction was not supported by our data: while cortical tracking of speech was reduced for speech interrupted by silence or noise bursts, this reduction is more parsimoniously attributed to the absence of the speech information itself rather than the disruption of a continuous processing

mechanism. We found that the tracking of segments of speech that had not been not been removed remained intact regardless of interruption; thus, the spurious spectrotemporal transients at the beginning and ending of intact speech segments did not appear to influence entrainment to the acoustic or linguistic features of speech. This called for a more nuanced account of speech-tracking: entrainment to speech does not simply reflect a bottom-up, stimulus-driven response to spectrotemporal modulations; rather, it reflects the alignment between the bottom-up sensory signal and top-down predictions about the sensory signal, with those predictions putatively coming from association areas (Davis & Johnsrude, 2003, 2007, Leonard et al., 2016, 2015; Peelle & Davis, 2012). Our second hypothesis was only partly supported by our results: the magnitude of speech-tracking responses was restored by restoring the acoustic envelope of interrupted speech, however the speech-tracking response itself was topographically different from the response to uninterrupted speech which suggests that a different set of neural ensembles are responsible for encoding interrupted speech. Interestingly, this difference occurred at around the same latency as the selection-related component observed in Chapter 3. This response may reflect a similar feedback-driven response to signals from higher-order areas responsible for generating predictions about an incoming behaviorally relevant stimulus. It should be noted that, while we do observe components in the EEG speech-tracking response at latencies around 200–300 ms in both experiments, they exhibit markedly different topographies which may indicate that the components reflect different underlying processes (i.e. there is not a shared feedback mechanism), or it may be due to the differential activation of brain areas by a similar

feedback mechanism – the illusory continuity in the phonemic restoration is known to activate distinct brain areas that are not activated by intact speech (Shahin et al., 2009). Since we cannot distinguish between these alternatives based on our current data further experiments are called for.

These experiments describe top-down influences in the brain's response to speech and suggest that cortical entrainment to speech reflects the interplay of high-level attentional and linguistic processes. The cognitive goal of attention to speech – to select one stream of speech for enhanced processing while excluding other, un-related stimuli – can be achieved by two complementary, oscillatory neural mechanisms: (a) sensory selection of the rhythmic speech signal (Schroeder & Lakatos, 2009; Zion Golumbic et al., 2012), and (b) temporally organizing neural activity within anatomically diverse neural ensembles to define functional networks that can optimally communicate and encode high-level or abstract stimulus features (Fries, 2005, 2015; Helfrich & Knight, 2016; Voloh & Womelsdorf, 2016; Voytek et al., 2015). Under current theories of speech perception these mechanisms work together. Oscillation-based sensory selection effectively segments and parses the continuous speech signal (Ghitza, 2011) while also providing a temporal frame-of-reference to organize activity throughout a language processing network that spans areas in frontal, prefrontal, and temporal cortices (Giraud & Poeppel, 2012). However, these theories of speech perception are limited because they describe speech perception in terms of a bottom-up stimulus decoding process, related to the speech-tracking electrophysiological response, and do not account for the role of top-down feedback processes. These theories cannot explain, for example, why interference

in the linguistic representation of speech results in an attenuated speech-tracking response as we observed in Chapter 2. Therefore, we propose an expansion to these theories of speech perception, based on a predictive coding perspective, which maintains that the cortical speech-tracking response reflects, to some degree, the agreement between bottom-up sensory responses and top-down predictions. Under this expanded theory, adding competing speakers to a scene reduces tracking of an attended signal by degrading the neural representation of high-level linguistic features, which in turn results in less robust predictive feedback to the sensory areas producing the EEG speech-tracking response. Some limited support for this theory, from studies using causality analysis, already exists: Park et al. (2015) found that low-frequency auditory cortical activity tracked speech better as a function of increased top-down signaling from frontal, motor, and posterior temporal areas; Di Liberto et al. (2018) found increased cortical entrainment to speech rendered intelligible through prior knowledge that was mediated by increased top-down signaling from left inferior frontal gyrus. These studies represent useful models for testing the predictions of a predictive coding account of speech-tracking. Future studies, preferably leveraging the superior spatial localization capabilities of intracranial recording or MEG, should be conducted to determine the causal relationship underlying activity associated with the selective attention, stream segregation, and perceptual restoration effects described in this thesis. The results of such studies would elucidate the neurobiological mechanisms that connect sound perception, attentional selection, and language processing.

References

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *28*(15), 3958–65. http://doi.org/10.1523/JNEUROSCI.0187-08.2008

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13367–72. http://doi.org/10.1073/pnas.201400998

Aiken, S. J., & Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear and Hearing*, *29*(2), 139–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18595182

Arbogast, T. L., Mason, C. R., & Kidd, G. (2002). The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America*, *112*(5 Pt 1), 2086. http://doi.org/10.1121/1.1510141

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

Aydelott, J., Jamaluddin, Z., & Nixon Pearce, S. (2015). Semantic processing of unattended speech in dichotic listening. *The Journal of the Acoustical Society of America*, *138*(2), 964–975. http://doi.org/10.1121/1.4927410

Baltzell, L. S., Srinivasan, R., & Richards, V. M. (2017). The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. *Journal of Neurophysiology*, (2010), jn.00023.2017. http://doi.org/10.1152/jn.00023.2017

Bashford, J. A., Riener, K. R., & Warren, R. M. (1992). Increasing the intelligibility of speech through multiple phonemic restorations. *Perception & Psychophysics*, *51*(3), 211–217. http://doi.org/10.3758/BF03212247

Bashford, J. A., & Warren, R. M. (1987). Multiple phonemic restorations follow the rules for auditory induction. *Perception & Psychophysics*, *42*(2), 114–121. http://doi.org/10.3758/BF03210499

Bashford, J. A., Warren, R. M., & Brown, C. A. (1996). Use of speech-modulated noise adds strong "bottom-up" cues for phonemic restoration. *Perception & Psychophysics*, *58*(5), 342–350. http://doi.org/10.3758/BF03206810

Başkent, D., Eiler, C., & Edwards, B. (2009). Effects of envelope discontinuities on perceptual restoration of amplitude-compressed speech. *The Journal of the Acoustical Society of America*, *125*(6), 3995–4005. http://doi.org/10.1121/1.3125329

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J.

(2012). Canonical Microcircuits for Predictive Coding. *Neuron*, *76*(4), 695–711. http://doi.org/10.1016/j.neuron.2012.10.038

Bee, M. A., & Klump, G. M. (2005). Auditory Stream Segregation in the Songbird Forebrain: Effects of Time Intervals on Responses to Interleaved Tone Sequences. *Brain, Behavior and Evolution*, *66*(3), 197–214. http://doi.org/10.1159/000087854

Bendixen, A. (2014). Predictability effects in auditory scene analysis: a review. *Frontiers in Neuroscience*, *8*, 60. http://doi.org/10.3389/fnins.2014.00060

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B - Methodological*, *57*(1), 289–300. http://doi.org/10.2307/2346101

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, *107*(2), 1065. http://doi.org/10.1121/1.428288

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Bregman, A. S. (1990). *Auditory scene analysis : the perceptual organization of sound*. Cambridge, MA: MIT Press.

Broadbent, D. E. (1952). Listening to one of two synchronous messages. *Journal of Experimental Psychology*, *44*(1), 51–55. http://doi.org/10.1037/h0056491

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*(5), 1465–1487. http://doi.org/10.3758/s13414-015-0882-9

Brookes, M. (2003). VOICEBOX: A speech processing toolbox for MATLAB. 2006.

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America*, *131*(2), 1449–1464. http://doi.org/10.1121/1.3675943

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109. http://doi.org/10.1121/1.1345696

Brungart, D. S., Simpson, B. D., Darwin, C. J., Arbogast, T. L., & Kidd, G. (2005). Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task. *The Journal of the Acoustical Society of America*, *117*(1), 292–304. http://doi.org/10.1121/1.1835509

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–2538.

http://doi.org/10.1121/1.1408946

Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S., & Bradlow, A. R. (2013). Masking Release Due to Linguistic and Phonetic Dissimilarity Between the Target and Masker Speech. *American Journal of Audiology*, *22*(1), 157. http://doi.org/10.1044/1059-0889(2013/12-0072)

Calandruccio, L., Dhar, S., & Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America*, *128*(2), 860–869. http://doi.org/10.1121/1.3458857

Chait, M., Greenberg, S., Arai, T., Simon, J. Z., & Poeppel, D. (2015). Multi-time resolution analysis of speech: Evidence from psychophysics. *Frontiers in Neuroscience*, *9*(MAY), 1–10. http://doi.org/10.3389/fnins.2015.00214

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436. http://doi.org/10.1371/journal.pcbi.1000436

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. http://doi.org/10.1121/1.1907229

Cogan, G. B., & Poeppel, D. (2011). A mutual information analysis of neural coding of speech by low-frequency MEG phase information. *Journal of Neurophysiology*, *106*(2), 554–63. http://doi.org/10.1152/jn.00075.2011

Cooke, M. (2003). Glimpsing speech. *Journal of Phonetics*, *31*(3–4), 579–584. http://doi.org/10.1016/S0095-4470(03)00013-5

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), S0140525X01003922. http://doi.org/10.1017/S0140525X01003922

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*(November), 1–14. http://doi.org/10.3389/fnhum.2016.00604

Culling, J. F., & Summerfield, Q. (1995). Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *The Journal of the Acoustical Society of America*, *98*(2), 785–797. http://doi.org/10.1121/1.413571

Dalton, P., & Fraenkel, N. (2012). Gorillas we have missed: Sustained inattentional deafness for dynamic events. *Cognition*, *124*(3), 367–372. http://doi.org/10.1016/j.cognition.2012.05.012

Darwin, C. J. (1997). Auditory grouping. *Trends in Cognitive Sciences*, *1*(9), 327–333.

http://doi.org/10.1016/S1364-6613(97)01097-8

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *114*(5), 2913. http://doi.org/10.1121/1.1616924

Darwin, C. J., & Hukin, R. W. (2000). Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *The Journal of the Acoustical Society of America*, *107*(2), 970–977. http://doi.org/10.1121/1.428278

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *23*(8), 3423–3431. http://doi.org/23/8/3423 [pii]

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147. http://doi.org/10.1016/j.heares.2007.01.014

Dean, I., Harper, N. S., & McAlpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. *Nature Neuroscience*, *8*(12), 1684–1689. http://doi.org/10.1038/nn1541

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. http://doi.org/10.1016/j.jneumeth.2003.10.009

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(1), 193–222. http://doi.org/10.1146/annurev.ne.18.030195.001205

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology*, *25*(19), 2457–65. http://doi.org/10.1016/j.cub.2015.08.030

Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Cortical Measures of Phoneme-Level Speech Encoding Correlate with the Perceived Clarity of Natural Speech. *Eneuro*, ENEURO.0084-18.2018. http://doi.org/10.1523/ENEURO.0084-18.2018

Di Liberto, G. M., & Lalor, E. C. (2016). Isolating Neural Indices of Continuous Speech Processing at the Phonetic Level. In P. van Dijk, D. Başkent, E. Gaudrain, E. de Kleine, A. Wagner, & C. Lanting (Eds.), *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing* (pp. 337–345). Cham: Springer International Publishing.

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

Di Liberto, G. M., Lalor, E. C., & Millman, R. E. (2018). Causal cortical dynamics of a

predictive enhancement of speech intelligibility. *NeuroImage*, *166*, 247–258. http://doi.org/10.1016/j.neuroimage.2017.10.066

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

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

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

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2016). Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews*. http://doi.org/10.1016/j.neubiorev.2017.02.011

Ding, N., & Simon, J. Z. (2009). Neural representations of complex temporal modulations in the human auditory cortex. *Journal of Neurophysiology*, *102*(5), 2731–2743. http://doi.org/10.1152/jn.00523.2009

Ding, N., & Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. http://doi.org/10.1073/pnas.1205381109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1205381109

Ding, N., & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78–89. http://doi.org/10.1152/jn.00297.2011

Ding, N., & Simon, J. Z. (2013). Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech. *Journal of Neuroscience*, *33*(13), 5728–5735. http://doi.org/10.1523/JNEUROSCI.5297-12.2013

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Speech Comprehension By Facilitating Perceptual Parsing. *Neuroimage*, *85*(15). http://doi.org/http://dx.doi.org/10.1016/j.neuroimage.2013.06.035

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *The Journal of the Acoustical Society of America*, *102*(4), 2403–2411. http://doi.org/10.1121/1.419603

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. *Neuron*, *61*(2), 317–329. http://doi.org/10.1016/j.neuron.2008.12.005

Elhilali, M., Xiang, J., Shamma, S. A., & Simon, J. Z. (2009). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biology*, *7*(6). http://doi.org/10.1371/journal.pbio.1000129

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nat Rev Neurosci*, *2*(10), 704–716. http://doi.org/10.1038/35094565 35094565 [pii]

Ericson, M. A., Brungart, D. S., & Brian, D. (2004). Factors That Influence Intelligibility in Multitalker Speech Displays. *The International Journal of Aviation Psychology*, *14*(3), 313–334. http://doi.org/10.1207/s15327108ijap1403

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*(December), 215. http://doi.org/10.3389/fnhum.2010.00215

Fishbach, A., Nelken, I., & Yeshurun, Y. (2001). Auditory Edge Detection: A Neural Model for Physiological and Psychoacoustical Responses to Amplitude Transients. *Journal of Neurophysiology*, *85*(6), 2303–2323. http://doi.org/10.1152/jn.2001.85.6.2303

Fishman, Y. I., Arezzo, J. C., & Steinschneider, M. (2004). Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. *The Journal of the Acoustical Society of America*, *116*(3), 1656–1670. http://doi.org/10.1121/1.1778903

Fishman, Y. I., Reser, D. H., Arezzo, J. C., & Steinschneider, M. (2001). Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hearing Research*, *151*(1–2), 167–187. http://doi.org/10.1016/S0378-5955(00)00224-0

Fogerty, D. (2013). Acoustic predictors of intelligibility for segmentally interrupted speech: temporal envelope, voicing, and duration. *Journal of Speech, Language, and Hearing Research : JSLHR*, *56*(5), 1402–8. http://doi.org/10.1044/1092-4388(2013/12-0203)

Fogerty, D., & Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, *131*(2), 1490–1501. http://doi.org/10.1121/1.3676696

Fogerty, D., Kewley-Port, D., & Humes, L. E. (2012). The relative importance of consonant and vowel segments to the recognition of words and sentences: effects of age and hearing loss. *The Journal of the Acoustical Society of America*, *132*(3), 1667–78. http://doi.org/10.1121/1.4739463

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America*, *106*(6), 3578–3588. http://doi.org/10.1121/1.428211

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through

neuronal coherence. *Trends Cogn Sci*, *9*(10), 474–480. http://doi.org/S1364-6613(05)00242-1 [pii] 10.1016/j.tics.2005.08.011

Fries, P. (2015). Rhythms for Cognition: Communication through Coherence. *Neuron*, *88*(1), 220–235. http://doi.org/10.1016/j.neuron.2015.09.034

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention - focusing the searchlight on sound. *Current Opinion in Neurobiology*, *17*(4), 437–455. http://doi.org/10.1016/j.conb.2007.07.011

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal Predictive Codes for Spoken Words in Auditory Cortex. *Current Biology*, *22*(7), 615–621. http://doi.org/10.1016/J.CUB.2012.02.015

Ghazanfar, A. a, Morrill, R. J., & Kayser, C. (2013). Monkeys are perceptually tuned to facial expressions that exhibit a theta-like speech rhythm. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(5), 1959–63. http://doi.org/10.1073/pnas.1214956110

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

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

Ghitza, O., Giraud, A.-L., & Poeppel, D. (2013). Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. *Frontiers in Human Neuroscience*, *6*(January), 340. http://doi.org/10.3389/fnhum.2012.00340

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1–2), 113–26. http://doi.org/10.1159/000208934

Gilbert, G., Bergeras, I., Voillery, D., & Lorenzi, C. (2007). Effects of periodic interruptions on the intelligibility of speech based on temporal fine-structure or envelope cues. *The Journal of the Acoustical Society of America*, *122*(3), 1336–1339. http://doi.org/10.1121/1.2756161

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–7. http://doi.org/10.1038/nn.3063

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

Greenberg, S. (1996). Understanding speech understanding: Towards a unified theory of speech perception. In *ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception* (pp. 1–8). Retrieved from https://www.isca-

speech.org/archive_open/absp_96/papers/asp6_001.pdf

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLoS Biology*, *11*(12). http://doi.org/10.1371/journal.pbio.1001752

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

Hambrook, D. A., & Tata, M. S. (2014). Theta-band phase tracking in the two-talker problem. *Brain and Language*, *135*, 52–56. http://doi.org/10.1016/j.bandl.2014.05.003

Helfrich, R. F., & Knight, R. T. (2016). Oscillatory Dynamics of Prefrontal Cognitive Control. *Trends in Cognitive Sciences*, *20*(12), 916–930. http://doi.org/10.1016/j.tics.2016.09.007

Hertrich, I., Dietrich, S., Trouvain, J., Moos, A., & Ackermann, H. (2012). 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. http://doi.org/10.1111/j.1469-8986.2011.01314.x

Hickok, G., Farahbod, H., & Saberi, K. (2015). The Rhythm of Perception: Entrainment to Acoustic Rhythms Induces Subsequent Perceptual Oscillation. *Psychological Science*, *26*(7), 1006–13. http://doi.org/10.1177/0956797615576533

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. http://doi.org/10.1038/nrn2113

Holdgraf, C. R., De Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J. J., … Theunissen, F. E. (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature Communications*, *7*(May). http://doi.org/10.1038/ncomms13654

Holender, D. (1986). *Semantic activation wihtout conscious identification in dichotic listening, parafoveal vision, and visual masking: a survey and appraisal*. *Behavioral & brain sciences* (Vol. 9).

Horton, C., D'Zmura, M., & Srinivasan, R. (2013). Suppression of competing speech through entrainment of cortical oscillations. *Journal of Neurophysiology*, *109*(12), 3082–3093. http://doi.org/10.1152/jn.01026.2012

Howard, M. F., & Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *Journal of Neurophysiology*, *104*(5), 2500–11. http://doi.org/10.1152/jn.00251.2010

Humes, L. E., Kidd, G. R., & Fogerty, D. (2017). Exploring Use of the Coordinate Response Measure in a Multitalker Babble Paradigm. *Journal of Speech Language and Hearing Research*, *60*(3), 741–754. http://doi.org/10.1044/2016_JSLHR-H-16-0042

IEEE Recommended Practice for Speech Quality Measurements. (1969). *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225–246. http://doi.org/10.1109/TAU.1969.1162058

Ihlefeld, A., & Shinn-Cunningham, B. (2008). Spatial release from energetic and informational masking in a selective speech identification task. *The Journal of the Acoustical Society of America*, *123*(6), 4369–79. http://doi.org/10.1121/1.2904826

Ille, N., Berg, P., & Scherg, M. (2002). Artifact Correction of the Ongoing EEG Using Spatial Filters Based on Artifact and Brain Signal Topographies. *Journal of Clinical Neurophysiology*, *19*(2), 113–124. http://doi.org/10.1097/00004691-200203000-00002

Kaiser, M., Senkowski, D., Roa Romero, Y., Riecke, L., & Keil, J. (2018). Reduced low-frequency power and phase locking reflect restoration in the auditory continuity illusion. *European Journal of Neuroscience*, (February), 1–8. http://doi.org/10.1111/ejn.13861

Kaya, E. M., & Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160101. http://doi.org/10.1098/rstb.2016.0101

Kayser, S. J., Ince, R. A. A., Gross, J., & Kayser, C. (2015). Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. *Journal of Neuroscience*, *35*(44), 14691–14701. http://doi.org/10.1523/JNEUROSCI.2243-15.2015

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(2), 620–8. http://doi.org/10.1523/JNEUROSCI.3631-09.2010

Knudsen, E. I. (2007). Fundamental Components of Attention. *Annual Review of Neuroscience*, *30*(1), 57–78. http://doi.org/10.1146/annurev.neuro.30.051606.094256

Kong, Y. Y., Mullangi, A., & Ding, N. (2014). Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hearing Research*, *316*, 73–81. http://doi.org/10.1016/j.heares.2014.07.009

Krishnan, L., Elhilali, M., & Shamma, S. A. (2014). Segregating complex sound sources through temporal coherence. *PLoS Computational Biology*, *10*(12), 1–10. http://doi.org/10.1371/journal.pcbi.1003985

Lachter, J., Forster, K. I., & Ruthruff, E. (2004). Forty-five years after broadbent (1958): Still no identification without attention. *Psychological Review*, *111*(4), 880–913. http://doi.org/10.1037/0033-295X.111.4.880

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science (New York,*

N.Y.), *320*(5872), 110–3. http://doi.org/10.1126/science.1154735

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron*, *77*(4), 750–61. http://doi.org/10.1016/j.neuron.2012.11.034

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. *Journal of Neurophysiology*, *102*(1), 349–359. http://doi.org/10.1152/jn.90896.2008

Larson, E., Maddox, R. K., Perrone, B. P., Sen, K., & Billimoria, C. P. (2012). Neuron-specific stimulus masking reveals interference in spike timing at the cortical level. *JARO - Journal of the Association for Research in Otolaryngology*, *13*(1), 81–89. http://doi.org/10.1007/s10162-011-0292-1

Leonard, M. K., Baud, M. O., Sjerps, M. J., & Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature Communications*, *7*, 1–9. http://doi.org/10.1038/ncomms13619

Leonard, M. K., Bouchard, K. E., Tang, C., & Chang, E. F. (2015). Dynamic Encoding of Speech Sequence Probability in Human Temporal Cortex. *Journal of Neuroscience*, *35*(18), 7203–7214. http://doi.org/10.1523/JNEUROSCI.4100-14.2015

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

Makov, S., Sharon, O., Ding, N., Ben-Shachar, M., Nir, Y., & Zion Golumbic, E. (2017). Sleep Disrupts High-Level Speech Parsing Despite Significant Basic Auditory Processing. *The Journal of Neuroscience*, *37*(32), 7772–7781. http://doi.org/10.1523/JNEUROSCI.0168-17.2017

McCloy, D. R., Souza, P. E., Wright, R. A., Haywood, J., Gehani, N., & Rudolph, S. (2013). The PN/NC corpus. Version 1.0. Retrieved from http://depts.washington.edu/phonlab/resources/pnnc/

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–6. http://doi.org/10.1038/nature11020

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, *343*(6174), 1006–1010. http://doi.org/10.1126/science.1245994

Micheyl, C., Tian, B., Carlyon, R. P., & Rauschecker, J. P. (2005). Perceptual Organization of Tone Sequences in the Auditory Cortex of Awake Macaques. *Neuron*, *48*(1), 139–148. http://doi.org/10.1016/j.neuron.2005.08.039

Middlebrooks, J. C., Dykes, R. W., & Merzenich, M. M. (1980). Binaural response-specific

bands in primary auditory cortex (AI) of the cat: Topographical organization orthogonal to isofrequency contours. *Brain Research*, *181*(1), 31–48. http://doi.org/10.1016/0006-8993(80)91257-3

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin*, *44*(2), 105–129. http://doi.org/10.1037/h0055960

Miller, G. A., & Licklider, J. C. R. (1950). The Intelligibility of Interrupted Speech. *The Journal of the Acoustical Society of America*, *22*(2), 167. http://doi.org/10.1121/1.1906584

Millman, R. E., Johnson, S. R., & Prendergast, G. (2015). 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. http://doi.org/10.1162/jocn_a_00719

Millman, R. E., Prendergast, G., Hymers, M., & Green, G. G. R. (2013). Representations of the temporal envelope of sounds in human auditory cortex: Can the results from invasive intracortical "depth" electrode recordings be replicated using non-invasive MEG "virtual electrodes"? *NeuroImage*, *64*, 185–196. http://doi.org/10.1016/j.neuroimage.2012.09.017

Morillon, B., Liégeois-Chauvel, C., Arnal, L. H., Bénar, C.-G., & Giraud, A.-L. (2012). Asymmetric function of theta and gamma activity in syllable processing: an intra-cortical study. *Frontiers in Psychology*, *3*, 248. http://doi.org/10.3389/fpsyg.2012.00248

Mrsic-Flogel, T. D., King, A. J., & Schnupp, J. W. H. (2005). Encoding of Virtual Acoustic Space Stimuli by Neurons in Ferret Primary Auditory Cortex. *Journal of Neurophysiology*, *93*(6), 3489–3503. http://doi.org/10.1152/jn.00748.2004

Murray, M. M., Brunet, D., & Michel, C. M. (2008). Topographic ERP analyses: A step-by-step tutorial review. *Brain Topography*, *20*(4), 249–264. http://doi.org/10.1007/s10548-008-0054-5

Narayan, R., Best, V., Ozmeral, E., McClaine, E., Dent, M., Shinn-Cunningham, B., & Sen, K. (2007). Cortical interference effects in the cocktail party problem. *Nature Neuroscience*, *10*(12), 1601–7. http://doi.org/10.1038/nn2009

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., … Brugge, J. F. (2009). Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex. *Journal of Neuroscience*, *29*(49), 15564–15574. http://doi.org/10.1523/JNEUROSCI.3065-09.2009

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., … Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. http://doi.org/10.1093/cercor/bht355

O'Sullivan, J. A., Shamma, S. A., & Lalor, E. C. (2015). Evidence for Neural Computations of Temporal Coherence in an Auditory Scene and Their Enhancement during

Active Listening. *Journal of Neuroscience*, *35*(18), 7256–7263. http://doi.org/10.1523/JNEUROSCI.4973-14.2015

Obleser, J., Herrmann, B., & Henry, M. J. (2012). Neural Oscillations in Speech: Don't be Enslaved by the Envelope. *Frontiers in Human Neuroscience*, *6*(August), 250. http://doi.org/10.3389/fnhum.2012.00250

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2015). Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners. *Current Biology*, *25*(12), 1649–1653. http://doi.org/10.1016/j.cub.2015.04.049

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., … Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, *10*(1). http://doi.org/10.1371/journal.pbio.1001251

Peelle, J. E., & Davis, M. H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology*, *3*, 320. http://doi.org/10.3389/fpsyg.2012.00320

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. *Cerebral Cortex*, *23*(6), 1378–1387. http://doi.org/10.1093/cercor/bhs118

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-Language Perspective on Speech Information Rate. *Language*, *87*(3), 539–558. http://doi.org/10.1353/lan.2011.0057

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*. http://doi.org/10.1163/156856897X00366

Peña, M., & Melloni, L. (2012). Brain Oscillations during Spoken Sentence Processing. *Journal of Cognitive Neuroscience*, *24*(5), 1149–1164. http://doi.org/10.1162/jocn_a_00144

Pérez, A., Carreiras, M., Gillon Dowens, M., & Duñabeitia, J. A. (2015). Differential oscillatory encoding of foreign speech. *Brain and Language*, *147*, 51–57. http://doi.org/10.1016/j.bandl.2015.05.008

Petkov, C. I., O'Connor, K. N., & Sutter, M. L. (2007). Encoding of Illusory Continuity in Primary Auditory Cortex. *Neuron*, *54*(1), 153–165. http://doi.org/10.1016/j.neuron.2007.02.031

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time.' *Speech Communication*, *41*(1), 245–255. http://doi.org/10.1016/S0167-6393(02)00107-3

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *The European Journal of Neuroscience*, *35*(9), 1497–503. http://doi.org/10.1111/j.1460-

9568.2012.08060.x

Pressnitzer, D., Sayles, M., Micheyl, C., & Winter, I. M. (2008). Perceptual Organization of Sound Begins in the Auditory Periphery. *Current Biology*, *18*(15), 1124–1128. http://doi.org/10.1016/j.cub.2008.06.053

Pulvermüller, F., & Shtyrov, Y. (2006). Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes. *Progress in Neurobiology*, *79*(1), 49–71. http://doi.org/10.1016/j.pneurobio.2006.04.004

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., & King, A. J. (2011). Contrast Gain Control in Auditory Cortex. *Neuron*, *70*(6), 1178–1191. http://doi.org/10.1016/J.NEURON.2011.04.030

Riecke, L., Esposito, F., Bonte, M., & Formisano, E. (2009). Hearing Illusory Sounds in Noise: The Timing of Sensory-Perceptual Transformations in Auditory Cortex. *Neuron*, *64*(4), 550–561. http://doi.org/10.1016/j.neuron.2009.10.016

Riecke, L., Formisano, E., Herrmann, C. S., & Sack, A. T. (2015). 4-Hz Transcranial Alternating Current Stimulation Phase Modulates Hearing. *Brain Stimulation*, *8*(4), 777–83. http://doi.org/10.1016/j.brs.2015.04.004

Riecke, L., van Opstal, A. J., Goebel, R., & Formisano, E. (2007). Hearing Illusory Sounds in Noise: Sensory-Perceptual Transformations in Primary Auditory Cortex. *Journal of Neuroscience*, *27*(46), 12684–12689. http://doi.org/10.1523/JNEUROSCI.2713-07.2007

Rimmele, J. M., Zion Golumbic, E., Schröger, E., & Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex*, *68*(5234), 144–154. http://doi.org/10.1016/j.cortex.2014.12.014

Rivenez, M., Guillaume, A., Bourgeon, L., & Darwin, C. J. (2008). Effect of voice characteristics on the attended and unattended processing of two concurrent messages. *European Journal of Cognitive Psychology*, *20*(6), 967–993. http://doi.org/10.1080/09541440701686201

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, *32*(1), 9–18. http://doi.org/10.1016/j.tins.2008.09.012

Shahin, A. J., Bishop, C. W., & Miller, L. M. (2009). Neural mechanisms for illusory filling-in of degraded speech. *NeuroImage*, *44*(3), 1133–1143. http://doi.org/10.1016/j.neuroimage.2008.09.045

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123. http://doi.org/10.1016/j.tins.2010.11.002

Shannon, R. V, Zeng, F.-G. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech

recognition with primarily temporal cues. *Science (New York, N.Y.)*, *270*(5234), 303–4. http://doi.org/10.1126/science.270.5234.303

Sheldon, S., Pichora-Fuller, M. K., & Schneider, B. A. (2008). Effect of age, presentation method, and learning on identification of noise-vocoded words. *The Journal of the Acoustical Society of America*, *123*(1), 476–488. http://doi.org/10.1121/1.2805676

Shinn-Cunningham, B. G., Ihlefeld, A., Satyavarta, & Larson, E. (2005). Bottom-up and top-down influences on spatial unmasking. *Acta Acustica United with Acustica*, *91*(6), 967–979. http://doi.org/10.1121/1.2996336

Shinn-Cunningham, B. G., & Wang, D. (2008). Influences of auditory object formation on phonemic restoration. *The Journal of the Acoustical Society of America*, *123*(1), 295. http://doi.org/10.1121/1.2804701

Simpson, S. A., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *The Journal of the Acoustical Society of America*, *118*(5), 2775–2778. http://doi.org/10.1121/1.2062650

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

Steinmetzger, K., & Rosen, S. (2017). Effects of acoustic periodicity and intelligibility on the neural oscillations in response to speech. *Neuropsychologia*, *95*(December 2016), 173–181. http://doi.org/10.1016/j.neuropsychologia.2016.12.003

Steinschneider, M., Nourski, K. V., & Fishman, Y. I. (2013). Representation of speech in human auditory cortex: Is it special? *Hearing Research*, *305*, 57–73. http://doi.org/10.1016/j.heares.2013.05.013

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, *111*(4), 1872–91. http://doi.org/10.1121/1.1458026

Treisman, A. M. (1964). The Effect of Irrelevant Material on the Efficiency of Selective Listening. *The American Journal of Psychology*, *77*(4), 533. http://doi.org/10.2307/1420765

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America*, *121*(1), 519–526. http://doi.org/10.1121/1.2400666

Vander Ghinst, M., Bourguignon, M., Op de Beeck, M., Wens, V., Marty, B., Hassid, S., … De Tiege, X. (2016). Left Superior Temporal Gyrus Is Coupled to Attended Speech in a Cocktail-Party Auditory Scene. *Journal of Neuroscience*, *36*(5), 1596–1606. http://doi.org/10.1523/JNEUROSCI.1730-15.2016

Volgushev, M., Chistiakova, M., & Singer, W. (1998). Modification of discharge patterns

of neocortical neurons by induced oscillations of the membrane potential. *Neuroscience*, *83*(1), 15–25. http://doi.org/10.1016/S0306-4522(97)00380-1

Voloh, B., & Womelsdorf, T. (2016). A Role of Phase-Resetting in Coordinating Large Scale Neural Networks During Attention and Goal-Directed Behavior. *Frontiers in Systems Neuroscience*, *10*(March), 1–19. http://doi.org/10.3389/fnsys.2016.00018

Voytek, B., Kayser, A. S., Badre, D., Fegen, D., Chang, E. F., Crone, N. E., … D'Esposito, M. (2015). Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. *Nature Neuroscience*, *18*(July), 1–10. http://doi.org/10.1038/nn.4071

Wang, X., & Humes, L. E. (2010). Factors influencing recognition of interrupted speech. *The Journal of the Acoustical Society of America*, *128*(4), 2100–11. http://doi.org/10.1121/1.3483733

Wang, Y., Ding, N., Ahmar, N., Xiang, J., Poeppel, D., & Simon, J. Z. (2012). Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: MEG evidence. *Journal of Neurophysiology*, *107*(8), 2033–2041. http://doi.org/10.1152/jn.00310.2011

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science (New York, N.Y.)*, *167*(917), 392–393. http://doi.org/10.1126/science.167.3917.392

Wen, B., Wang, G. I., Dean, I., & Delgutte, B. (2009). Dynamic range adaptation to sound level statistics in the auditory nerve. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(44), 13797–808. http://doi.org/10.1523/JNEUROSCI.5610-08.2009

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful Listening: The Processing of Degraded Speech Depends Critically on Attention. *Journal of Neuroscience*, *32*(40), 14010–14021. http://doi.org/10.1523/JNEUROSCI.1528-12.2012

Winkler, I., Denham, S., Mill, R., Bohm, T. M., & Bendixen, A. (2012). Multistability in auditory stream segregation: a predictive coding view. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1591), 1001–12. http://doi.org/10.1098/rstb.2011.0359

Womelsdorf, T., & Everling, S. (2015). Long-Range Attention Networks: Circuit Motifs Underlying Endogenously Controlled Stimulus Selection. *Trends in Neurosciences*, *38*(11), 682–700. http://doi.org/10.1016/j.tins.2015.08.009

Zion Golumbic, E. M., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(4), 1417–26. http://doi.org/10.1523/JNEUROSCI.3675-12.2013

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A. A., McKhann, G.

M. M., … Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron*, *77*(5), 980–991. http://doi.org/10.1016/j.neuron.2012.12.037

Zion Golumbic, E. M., Poeppel, D., & Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. *Brain and Language*, *122*(3), 151–61. http://doi.org/10.1016/j.bandl.2011.12.010

Zoefel, B., & VanRullen, R. (2015a). Selective perceptual phase entrainment to speech rhythm in the absence of spectral energy fluctuations. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*(5), 1954–64. http://doi.org/10.1523/JNEUROSCI.3484-14.2015

Zoefel, B., & VanRullen, R. (2015b). The Role of High-Level Processes for Oscillatory Phase Entrainment to Speech Sound. *Frontiers in Human Neuroscience*, *9*(December), 1–12. http://doi.org/10.3389/fnhum.2015.00651

Zoefel, B., & VanRullen, R. (2016). EEG oscillations entrain their phase to high-level features of speech sound. *NeuroImage*, *124*, 16–23. http://doi.org/10.1016/j.neuroimage.2015.08.054