ANALYSIS OF PATTERNS OF NEURAL ACTIVITY IN RESPONSE TO AUDITORY STIMULI

DILLON A. HAMBROOK

Bachelor of Science, University of Lethbridge, 2011

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

MASTER OF SCIENCE

Neuroscience
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

Abstract

It is usually easy to understand speech, but when several people are talking at once it becomes difficult. The brain must select one speech stream and ignore distracting streams. This thesis tested a theory about the neural and computational mechanisms of attentional selection. The theory is that oscillating signals in brain networks phase-lock with amplitude fluctuations in speech. By doing this, brain-wide networks acquire information from the selected speech, but ignore other speech signals on the basis of their non-preferred dynamics. Two predictions were supported: first, attentional selection boosted the power of neuroelectric signals that were phase-locked with attended speech, but not ignored speech. Second, this phase selectivity was associated with better perception for the attended speech. We also describe a novel analysis of neuroelectric responses in complex auditory scenes, and suggest a new model of auditory distraction that is consistent with some unexpected results.

Acknowledgements

Foremost, I would like to thank my supervisors Dr. Matthew Tata and Dr. Artur Luczak for their support, guidance, and freedom to follow my curiousity. I would also like to thank the other members of my graduate committee: Dr. David Euston and Dr. Kevin Grant; their feedback has been invaluable. I am immensely grateful to my labmates, past and present: Amanda McMullan, Karla Ponjavic-Conte, Sebastian Pavlovic, Scott Oberg, Erin Zelinski, Sheena MacInnis, Sam Dodic, and Ryutaro Uchiyama; it was through conversations with you guys that I honed my own understanding.

Finally, I am forever indebted to my friends and family for their love, patience, understanding, and most of all their ability to distract me from the mysteries of the brain.

Table of Contents

List of Figures

1 Introduction

1.1 Background and History

Making sense of a world filled with sound is a complex perceptual task that has been the subject of much investigation over the past 60 years. The recognition of speech in particular is an extremely difficult perceptual problem because speech is a spectrotemporally dynamic series of sounds which, simply based on its physical properties, is difficult to predict. The difficult problem of speech perception is further exasperated when the brain must deal with multiple speech streams.

This problem was astutely identified and conceptualized 60 years ago as the "cocktail party problem" (Cherry, 1953). The problem is simply: How do we identify what one person is saying when others are speaking at the same time? This question has formed the basis for extensive study. Research has been typically divided into three themes. Studies of selective listening focus on the properties and consequences of auditory selective attention (Broadbent, 1958; Treisman, 1964). Extensive research has been carried out to characterize how sounds are integrated or segregated into coherent streams relating to discrete sound sources (Bregman, 1990). Studies of the effects of competing sounds on speech intelligibility focus on how speech perception is impaired by other environmental sounds (Bronkhorst, 2000; Kidd, Mason, Richards, Gallun, & Durlach, 2007). Although these three themes are theoretically and empirically related, this thesis adopts the topic of selective listening in the context of speech perception as a primary focus.

Early studies of selective attention used a common paradigm in which listeners were presented with two different speech signals, usually into different ears. Their task

was to listen and encode information from these speech streams (Cherry, 1953). The key finding in Cherry's study was that listeners could follow the target speech with high accuracy, provided that the other speech signal was almost completely ignored. This finding directly led to the synthesis of Broadbent's theory of early attentional selection (Broadbent, 1958).

Broadbent's theory maintained that, because human information processing capacity is less than the bandwidth of sensory channels, a selective filter is necessary to prevent overwhelming the system. Thus, Broadbent proposed that information from attended sensory channels passes through the filter, while information from unattended sensory channels is blocked from accessing higher-order processing systems. While such a model of attention explains the effect of attention on a single target, it fails to explain how unattended signals with highly salient content – for example hearing one's name – is capable of capturing attention.

An alternate, but closely related theory is an attenuation model of attention, proposed by Anne Tresiman (1969). The model proposed that attentional selection operated by attenuating the signal from unattended sensory channels. This model allows for both attended and unattended stimuli to be processed; however, unattended stimuli must overcome a higher threshold to gain access to processing mechanisms. As we will discuss later, recent electrophysiological evidence may provide evidence of brain mechanisms that support such a model.

The study of sound segregation is intimately related to the study of the cocktail party effect. In order to successfully recognize what one person is saying we must necessarily group their speech sounds into a coherent stream of sounds, while excluding

unrelated sounds. A series of experiments by Albert Bregman examined how sounds are perceptually organized (Bregman, 1990). Bregman conceptualized the sound integration processes as occurring either sequentially or simultaneously. Sequential integration was used to describe the process by which sequential sounds were integrated to form a continuous percept of a single sound source. Bregman identified temporal regularity, spectral relationships between sounds, and attentional focus as primary factors influencing the sequential integration of sounds. In general, similar sounds are sequentially grouped together, given that the listener's task is to perceive such a stream. Sounds are grouped simultaneously based on factors including harmonicity, shared modulation in frequency and amplitude, shared spatial localization, and based on schemas developed before the sounds are heard. The principles of sound integration based on similarity are borne out by studies that have explored speech intelligibility by essentially breaking sound segregation.

There is a broad literature that has studied the degradation of speech intelligibility when target speech is presented against a competing "masker" stimuli (see Bronkhorst, 2000; Kidd et al., 2007 for review). The literature differentiates between two broad categories of masking: When target speech is masked by broadband noise the degradation of speech intelligibility is primarily due to interference at the basilar membrane; such interference in the sensory periphery is commonly referred to as energetic masking. When speech is masked by other speech or by a spectrotemporally similar dynamic signal, it is known as informational masking (Pollack, 1975). At the same average loudness energetic masking is far more effective at degrading the intelligibility of speech (Miller, 1947). Unsurprisingly, the masking effectiveness of competing speech and

simultaneous sound integration depend on similar factors such as frequency proximity. Masking sounds comprised of other speech streams greatly reduce intelligibility when the masker is localized near the target (Arbogast, Mason, & Kidd, 2002), when the masker shares a similar pitch (Brungart, 2001), and when the target and masker share a similar temporal profile (Bronkhorst, 1992). While factors influencing speech intelligibility have been widely studied, only recently have neuroimaging techniques allowed for explorations of the mechanisms supporting speech perception.

1.2 Selective Entrainment of Neural Oscillations as a Solution to the Cocktail Party Problem

Neuroimaging studies have shown that cortical responses to speech stimuli reflect spectral and temporal content not only of individual words, but also of entire speech segments (Ahissar et al., 2001; Giraud et al., 2000; Luo, Wang, Poeppel, & Simon, 2006; Suppes, Han, Epelboim, & Lu, 1999). A key study by Luo & Poeppel (2007) found evidence that low-frequency (4 - 8 Hz) auditory cortex oscillations share temporal dynamics with the speech stimuli that drive them. The phase of these low frequency oscillations, as measured by the magnetoencephalogram (MEG) and electroencephalogram (EEG) signal, reliably discriminated different speech samples. Furthermore, they found that degrading the intelligibility of the speech samples by creating a speech-noise chimera reduced the discriminability of the resulting brain signals – suggesting that this phase tracking of speech reflects the neural encoding of speech. Subsequent studies found that speech comprehension was not necessary for phase tracking (Howard & Poeppel, 2010); however, comprehension does enhance the degree

of tracking of speech (Peelle, Gross, & Davis, 2013). Phase-tracking of speech stimuli is also enhanced by the inclusion of matching, synchronized visual stimuli (Luo, Liu, & Poeppel, 2010; Zion Golumbic, Cogan, Schroeder, & Poeppel, 2013), suggesting that a multimodal network might be temporally entrained by speech dynamics. Many of these studies converge on the acoustic envelope of speech as the key factor driving the phase-entrainment phenomenon (Aiken & Picton, 2008; Hertrich, Dietrich, Trouvain, Moos, & Ackermann, 2012; Lalor & Foxe, 2010; but see Obleser, Herrmann, & Henry, 2012 for a dissenting opinion). While these results suggest a link between phase-tracking and speech perception they do little to explain the perceptual importance of the entrainment of neural oscillations to speech.

It has been proposed that entrainment of neural oscillations by discrete acoustic streams may reflect the selection of those streams, in a manner consistent with Broadbent's and Treisman's models of early attentional selection (Malsburg & Schneider, 1986; Schroeder & Lakatos, 2009; Zion Golumbic, Poeppel, & Schroeder, 2012). This is known as the *selective entrainment hypothesis*. This hypothesis is based on evidence that: sensitivity of spiking neural assemblies is modulated by the phase of low-frequency oscillations as measured in local-field potentials (LFP) and EEG (Lakatos et al., 2005); perceptual sensitivity is modulated by neural oscillations (Kayser, Petkov, & Logothetis, 2008; Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007); and that phase-selection of oscillations may provide a means of attentional selection (Fries, 2005; Lakatos, Karmos, Mehta, Ulbert, & Schroeder, 2008). The theory is that by entraining neural oscillations to the acoustic envelope of a target speech stream, neural assemblies are made most sensitive to important acoustic events in that attended speech stream.

Furthermore, because competing speech will have a different temporal profile, it will be functionally suppressed because key events in those streams arrive during periods of non-maximal neural sensitivity.

The selective entrainment hypothesis makes two very clear predictions: First, there should be greater low-frequency phase-tracking of attended speech, compared to unattended speech. Second, greater phase-tracking of the attended speech is associated with some perceptual benefit, such as better encoding in memory or more sensitive discrimination of incoming words. A number of recent studies support the first prediction that phase-tracking of attended speech is enhanced (Ding & Simon, 2012; Kerlin, Shahin, & Miller, 2010; Power, Foxe, Forde, Reilly, & Lalor, 2012; Zion Golumbic, Ding, et al., 2013). However, these studies did not report evidence in support of the second prediction.

In this thesis we will present two experiments that test the predictions of the selective entrainment hypothesis in two-talker and multi-talker situations. Particular emphasis is placed on examining the possible connection between electrophysiological data and behavioural data to evaluate if there is a perceptual advantage afforded by phase-entrainment to a target speech stream.

1.3 Methodology

The envelope of human speech is temporally dynamic, with periodicities ranging from milliseconds to seconds. Of particular interest is an amplitude modulation at approximately 5 Hz that is thought to be related to the rate of syllable boundaries (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; MacNeilage,

1998). This low-frequency modulation has been under particular scrutiny with respect to entrainment between speech dynamics and neuroelectric dynamics.

In the basic case of a single speech stream, the auditory scene contains only one acoustic envelope with only one amplitude modulating signal. What makes the cocktail party problem particularly challenging is that, in scenes with multiple speakers, the physical vibrations that make up sounds mix and interfere as they propagate. Thus, the ear (or recording sensor) necessarily receives the superposition of a number of individual signals. The problem of unmixing this superposition of signals has been a defining constraint over the past several decades of auditory cognitive neuroscience. This constraint has caused the substantial majority of auditory EEG and MEG studies to use simple discrete tones presented in isolation against a nearly silent noise background. Even studies that use a speech background and a task that requires listeners to attend to a competing speech stream used discrete tones co-localized to a target speech stream and subsequently generated event-related potentials (ERP) to the tones (Lambrecht, Spring, & Münte, 2011; Münte, Spring, Szycik, & Noesselt, 2010). Although experimentally tractable, such an auditory scene is profoundly unnatural and the ecological validity of such studies is questionable. One key goal of the present thesis was to develop a novel approach to computationally "unmix" the superposition of neuroelectric responses to realistically complex auditory scenes - scenes comprised of multiple speech streams with independent dynamics.

Despite the superposition problem described above, EEG and MEG remain the preferred methodologies for studying attentional dynamics in responses to acoustic stimuli. This preference is rooted almost entirely in the high temporal precision that can

be obtained (in principal) with these techniques. Metabolic techniques such as fMRI and PET necessarily blur events in time, whereas EEG and MEG allow temporally accurate measurements of intracranial current flows, at the cost of spatial resolution. For this reason, the studies described here made extensive use of the 128-channel dense-array EEG system at the University of Lethbridge to investigate EEG dynamics, but put little emphasis on characterizing the underlying functional anatomy of speech perception.

The EEG acquisition system is integrated into a multi-speaker virtual audio space such that EEG and audio signals can be co-registered in time with millisecond precision. This time-alignment allowed us to develop an analysis approach in which the positive half of the first-derivative of the acoustic envelopes of individual speech segments were cross-correlated with EEG signals to extract a record of phase-locked EEG power during selective listening. This signal captures acoustic onset transients, to which the auditory system is particularly sensitive, and which may provide important cues for the parsing of speech (Fishbach, Nelken, & Yeshurun, 2001; Howard & Poeppel, 2010; Seither-Preisler, Krumbholz, Patterson, Seither, & Lutkenhoner, 2004). It is this cross-correlation signal, along with simple psychophysical measures of perception that provides the basis by which we may test predictions about selective listening to speech in complex scenes.

2 Theta-band phase tracking in the two-talker problem

2.1 Introduction

The human auditory system has a striking ability to selectively perceive a single sound source out of a complex mixture. This general phenomenon and the associated computational challenges have been termed the "cocktail party problem" (Cherry, 1953). This problem emerges in any acoustic scene with more than one sound source., The perceptual consequence of failing to maintain selection in a complex scene has been called auditory information masking (Kidd et al., 2007), or more generally, *distraction*.

The neural mechanisms by which we deal with complex scenes have been under intense investigation in recent years. A promising recent theory, called *selective entrainment* (Schroeder & Lakatos, 2009; Zion Golumbic et al., 2012), proposes that this problem is solved in part by phase matching between neuroelectric oscillations of the brain and low-frequency dynamics of acoustic signals. It is known now that neuroelectric oscillatory activity can "track" spectrotemporal modulations in speech (Abrams, Nicol, Zecker, & Kraus, 2008; Ahissar et al., 2001; Hertrich et al., 2012; Luo & Poeppel, 2007). Furthermore, selective attention modulates the selectivity or strength of this tracking process (Ding & Simon, 2012; Kerlin et al., 2010; Mesgarani & Chang, 2012). By selectively tracking the phase of a single audio source, oscillating ensembles might preferentially represent the tracked signal and reject signals that are not phase locked.

Evidence for such a theory has begun to emerge: theta-band phase tracking of speech is more pronounced when the speech signal is well comprehended relative to when it is degraded and difficult to understand (Peelle et al., 2013). Thus phase-tracking is a correlate of successful perception. Furthermore, using intracranial electrocorticography (ECoG), (Zion Golumbic, Ding, et al., 2013) showed that oscillatory signals in auditory cortex track the acoustic envelope of speech in a non-selective manner - both attended and unattended speech signals were similarly tracked. By contrast, Medial Frontal Gyrus (MFG) exhibited selective tracking such that the attended speech was preferentially tracked. Since this region of cortex is also known to engage in auditory working memory tasks (Arnott, Grady, Hevenor, Graham, & Alain, 2005; Crottaz-Herbette, Anagnoson, & Menon, 2004), these data suggest a role for phase tracking in linking sensory and memory regions. Finally, theta-band phase tracking of speech was more pronounced when the speech signal was accompanied by video of the talker's lip movements (Zion Golumbic, Cogan, et al., 2013) - suggesting that phase-tracking is associated with communication between ensembles of neurons that are anatomically distinct but functionally linked.

Selective attention in a complex scene is well-known to enhance perception and memory encoding (Broadbent, 1952; Treisman, 1964). If phase tracking of speech dynamics is a mechanism for implementing selective attention, then variation in perceptual performance should mirror variation in the strength of speech-locked EEG signals. In the present study we report that selective listening in a free-field "two-talker" situation strengthens a theta-band signal that tracks the acoustic envelope of selected speech, relative to ignored speech. Furthermore, by reassigning trials on the basis of

correct or erroneous recall of a probe word, we found evidence that selective phase tracking of an attended stream supports perception.

Briefly, participants listened to two different, simultaneously presented, 15-second audio book clips read by different speakers, presented 60° to either side of the acoustic midline while EEG was recorded. Before each block of 15-second trials participants were cued to attend to one of the two speakers. Following each trial participants were presented a probe word from the target clip, the distractor clip, or a clip that was not presented on that trial (catch probe). The participants' task was a two-alternative forced choice task to indicate if the probe word was present or absent in either of the previously played clips. EEG data from each trial were cross-correlated with the first derivatives of the speech envelopes of the target and distractor speech clips played on that trial. The first 1000 ms of the EEG data for each trial was excluded as it contained transient responses due to the sudden onset of sound. This cross-correlation function selectively separated brain activity that was phase-locked to energy transients in either speech stream. We tested the prediction that EEG signals independently phase locked to target and distractor streams would be differentiated when the target was successfully encoded, but not when encoding of the target was compromised by the distracting stream.

2.2 Methods

2.2.1 Participants

19 undergraduates from the University of Lethbridge 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. 2 participants were excluded for failing to respond on a significant number of trials (3 standard deviations outside the mean across all trials). Only EEG data from participants who correctly responded at a rate higher than chance (>50% correct) to the target stream were analyzed, thus 16 participants contributed to the data analysis (12 female; 2 left-handed; average age: 22.2 years).

2.2.2 Stimuli & Task

All stimuli were presented in free field by an Apple Mac Pro with a firewire audio interface (M-Audio Firewire 410). Participants sat between two near-field studio monitors (Mackie HR624 MK-2) arranged 1 metre away and 60° from the front auditory midline. Stimulus presentation was controlled by a program custom coded using Apple Computer's Core Audio framework (Mac OS 10.6).

The stimuli consist of 20 segments from the book *World War Z* by Max Brooks, narrated by 20 different readers (1 female). Each segment was 15 seconds long and normalized to the same average RMS sound amplitude. Three unique probe words were selected from each of the 20 speech segments and audio clips of the selected words were obtained from an online dictionary.

Each participant completed 20 blocks of 5 trials each. Blocks were of 98 seconds duration. Each speech segment was the target on five trials. Within each block the presentations of speech segments were randomized and an individual speech segment did not occur twice within a single block. Prior to each block participants were instructed to attend to either the left or right speaker. The target and distractor streams were presented simultaneously from separate speakers for 15 seconds, followed by a 1 second silence,

followed by a probe word presented from both speakers. Participants were given 3.5 ± 0.25 seconds following the probe word to respond before the start of the next trial. Probe words were drawn from the target stream, distractor stream, or a stream that was not presented on that trial (probe absent or "catch" trials). Participants performed a two-alternative forced choice task to indicate if the probe word was present or absent in either of the speech clips.

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 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, Berg, & Scherg, 2002). Data were interpolated to an 81-channel 10-10 montage and exported from BESA and further analyzed in MATLAB (MATLAB version 7.10.0; The Mathworks Inc., 2010, Natick, Massachusetts, 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 was calculated. The acoustic envelope of each sample was calculated by taking the absolute value of the Hilbert transform of the sample and

low-pass filtering 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 that the sum of the signal across the whole epoch equaled 1 (Hertrich et al., 2012). Thus a signal which captures transient energy increases, an aspect of acoustic stimuli to which the auditory system is known to be tuned, was obtained (Fishbach et al., 2001; Howard & Poeppel, 2010). This signal was then cross-correlated with each channel of the time-aligned EEG data to arrive at a cross-correlation function which reflects activity that is phase-locked to acoustic transients in either stream.

To determine the frequency content of the observed phase-locked activity wavelet decomposition was performed on the cross-correlation function. Evoked power was calculated as the power in the trial-averaged cross-correlation function, normalized by the mean evoked power across the whole [-200, 800] ms epoch.

2.3 Results

Repeated measures t-tests were conducted to compare differences in response rates (Figure 2.1) when the probe was drawn from the target stream, the distractor stream, or a stream that was not heard on that particular trial (i.e. a "catch" trial which ensures that participants monitor the stimuli and do not respond affirmatively for all trials). Participants successfully detected the presence of the probe when it was in the target stream ("responded present" vs. "responded absent", $t=11.16$, $p<0.0001$), but not when it was in the distractor stream ($t=-0.72$, $p=0.4846$). Participants also successfully noted the absence of the probe on "catch" trials ("responded present" vs. "responded absent", t=-

6.4, $p<.0001$). The proportion of correct detections ("responded present") was greater when the probe was present in the target stream relative to the distractor stream ($t=4.89$, $p=.0003$).

Figure 2.2A shows the grand averaged cross-correlation of 14-seconds of speech dynamics and recorded EEG, with the first second following stimulus onset removed to exclude transient activity due to the initial speech onset, at electrode Fz. The most robust difference occurred at ~150ms lag. The cross-correlation functions for target and distractor speech had similar scalp topographies at this lag (Figure 2.2B), suggesting that the difference was due to an increase in power that was phase-locked to the first-derivative of the acoustic envelope, rather than a spatial reconfiguration of cortical generators.

Previous studies suggested that EEG signals were maximally phase-locked to speech in the theta band (4 – 8 Hz). We used a wavelet time-frequency decomposition to explore the frequency content of the cross-correlation function for target and distractor speech streams (Figure 2.3A). Phase-locked power was maximal in the theta and alpha frequency bands (4-14 Hz), for both target and distractor speech. Because previous studies have identified differences primarily in the theta band we sought to assess differences between the theta-band response to the target and distractor speech, a Monte Carlo permutation test was performed on the time-frequency data, averaged across 4-8 Hz, with a correction to preserve false-discovery rate (FDR) (Benjamini & Hochberg, 1995). There was significantly more theta-band power phase-locked to the target speech than the distractor speech at lags 140-240 ms ($p<0.0011$, FDR corrected $\alpha=0.05$) (Figure 2.3B). To assess the behavioral importance of this difference in phase-locked theta

power, we sorted probe-present trials according to performance: Correct responses were those trials on which participants correctly detected a probe in the target stream and errors were those trials on which the participant missed a probe in the target stream. Phase-locked theta power was averaged across latencies from 140 – 240 ms (Figure 2.3C). Differences were assessed using Wilcoxon signed rank tests. Target-locked theta power was significantly greater than distractor-locked theta power on correct trials ($Z=-2.10$, $p=0.0353$), but not on error trials ($Z=-1.48$, $p=0.153$). Importantly, there was significantly more power phase-locked to the target stream on correct trials than on error trials ($Z=-1.98$, $p=0.0494$), but no difference between distractor-locked theta power on correct and error trials ($Z=-1.16$, $p=0.2676$).

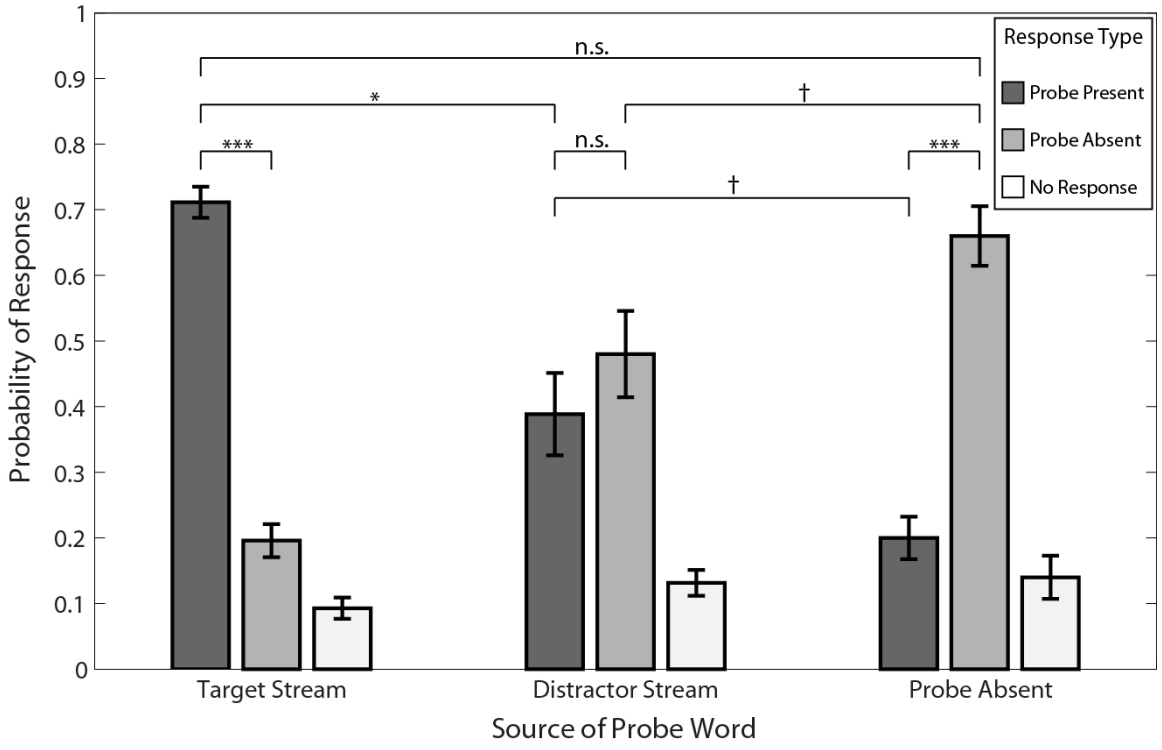

Figure 2.1: Probability of response types when the probe word is from the target stream, distractor stream, or an unheard stream. Probability of a correct response is lower when the probe is drawn from the distractor stream, compared to when the probe is from the target stream or from an unheard stream († $p < 0.05$, uncorrected; * $p < 0.01$, Bonferroni corrected; *** $p < 0.0001$, Bonferroni corrected). Error bars indicate standard error of the mean.

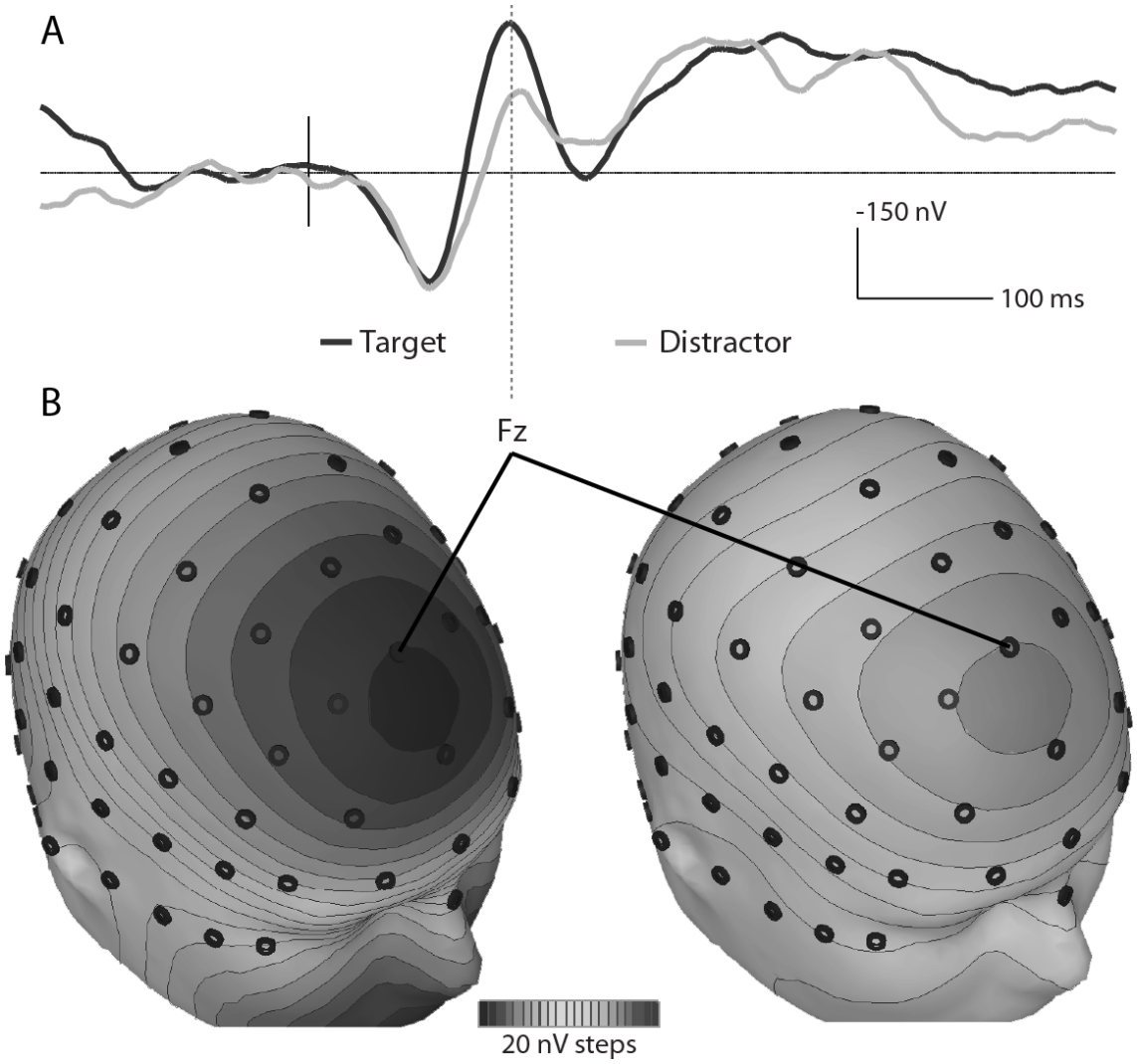

Figure 2.2: (A) Cross-correlation of the EEG signal and the first derivative of the speech envelope of target and distractor speech streams at electrode Fz. (B) Isopotential maps of the cross-correlation function for target (left) and distractor (right) speech streams.

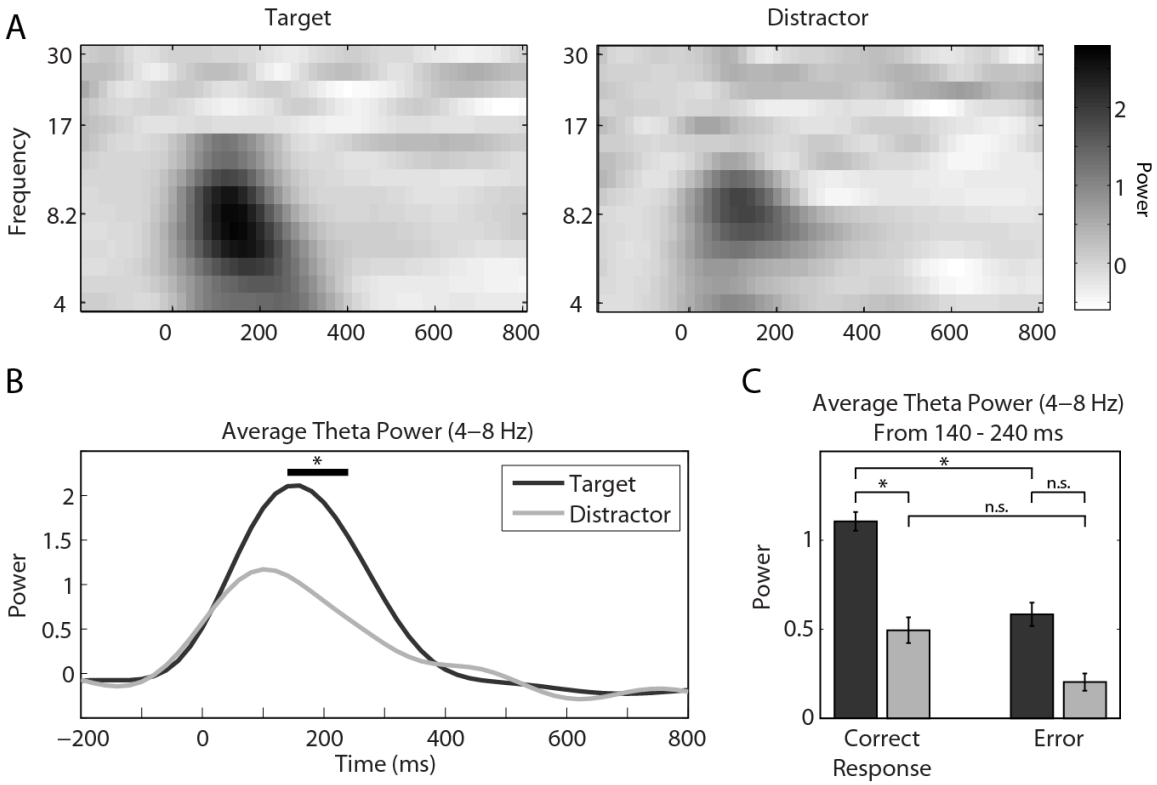

Figure 2.3: (A) Time frequency plots of grand averaged evoked power in the cross-correlation between the EEG signal and the first derivative of the speech envelope of the target (left) and distractor (right) streams. (B) Grand averaged evoked power for the theta band (4-8 Hz). Black bar indicates time bins in which there was a significant difference between the response to the target and distractor streams ($p < 0.0011$, FDR corrected $\alpha = 0.05$). (C) Grand averaged evoked power for the theta band (4-8 Hz) averaged across latencies from 140 – 240 ms. Trials were split on the basis of the participant's successful identification of the subsequent probe word. (* $p < 0.05$).

2.4 Discussion

2.4.1 Effects of Attention

Our behavioral results show a clear effect of attention on participants' ability to correctly recall words from recently heard speech streams. Participants were significantly more likely to correctly recall words than to miss words when they were presented in the target speech stream. By contrast, participants were not more likely to correctly recall words from the distractor stream (in fact they were slightly more likely to miss words in the distractor stream). However, it is important to note that they were more likely to recall the probe word when it was present in either stream relative to when it was absent. This suggests that some aspects of the distractor speech were at least occasionally encoded.

Our electrophysiological results showed an enhancement of the theta band EEG response that was phase locked to the acoustic dynamics of target compared to distractor speech. This was most evident at latencies between 140 – 240ms. This result converges with other studies that used different analysis techniques to identify the effect of attention on low frequency, speech-locked activity (Ding & Simon, 2012; Kerlin et al., 2010; Power et al., 2012; Zion Golumbic, Ding, et al., 2013). This result provides confirmation that correlating the acoustic dynamics of competing speech streams with the electrophysiological response to the sound mixture is an effective way to separate the neural response to the individual streams. This is a way to "unmix" the concurrent dynamics of competing speech representations.

The cross-correlation method employed in the present study acts as a low-pass filter with a kernel unique to each speech envelope. It therefore retains only the low-frequency activity that is phase-locked to the acoustic dynamics of individual speech segments. It is therefore not surprising that we observed no effects at higher frequencies. Our results suggest that low-frequency amplitude modulations in the acoustic envelope of speech are cues that allow for the entrainment of neuroelectric activity to speech dynamics. However we do not rule out a role for higher-frequency signals in speech selection and scene analysis.

Phase tracking of speech dynamics is a relatively unexplored phenomenon and, while studies have been conducted to understand the mechanism (Ding & Simon, 2012; Zion Golumbic, Ding, et al., 2013) there has been less effort applied to understanding its behavioral correlates. The combination of our behavioral and electrophysiological data illustrates the perceptual importance of phase-tracking. We showed that enhanced phase tracking of the dynamics of speech preceded correct recall of probe words in the selected speech stream. There was not a significant enhancement of phase-tracking the target stream preceding failures of recall. This result aligns with the results of Mesgarani and Chang (2012). That study reconstructed the spectrogram of speech based on local field potentials. They found that on successful trials the target speech stream was preferentially represented whereas on error trials competing streams were equally represented. These results suggest that an increase in phase tracking of target speech is associated with solving the cocktail party problem.

2.4.2 Selective Entrainment

The *selective entrainment* hypothesis, proposed by Schroeder and colleagues (Schroeder & Lakatos, 2009; Zion Golumbic et al., 2012) suggests that entrainment of neural oscillations to the temporal dynamics of a single behaviorally relevant stream is a mechanism for attentional selection. What perceptual benefit might phase tracking provide to the listener? The theory is well aligned with the notion that communication between two or more neural ensembles is facilitated when their oscillatory behavior is synchronized - a model known as *communication through coherence* (Fries, 2005). In this view, phase coherence enables a transmitting neuron (or group of neurons) to optimally drive a receiving neuron by aligning their pre-synaptic activity with temporal windows of maximal sensitivity to post-synaptic depolarization. By modulating phase, such a mechanism might provide a means for some neural assemblies to "ignore" inputs from non-selected cells (Engel, Fries, & Singer, 2001).

Taken together, *selective entrainment* and *communication by coherence* provide a possible framework in which to explore perception and distraction in complex scenes. Selective attention has been conceptualized for over a century as the preferential representation of a single source of information for enhanced perception, at the exclusion of other sources (James, 1890). This implies that attended acoustic signals selectively contribute information to, for example, working-memory, reward-processing, and response-planning mechanisms. At the neuronal level, this implies that networks of cells representing the features of attended sensory input should not only be bound together within sensory cortex (Malsburg & Schneider, 1986; Singer & Gray, 1995), but should also have preferential access to brain-wide non-sensory areas. To be selectively attended,

an auditory signal should have exclusive access to brain-wide networks, while representations of distracting signals should be unable to communicate outside of sensory cortex. Selective entrainment might create a computational bias in the cortex by enabling a phase-predictive process: by entraining brain-wide neural oscillations to the modulation frequency of a single speech stream, neural ensembles across multiple cortical systems can become biased to events in that stream.

This view of the selective entrainment predicts that the EEG signal should exhibit little or no phase tracking of unattended speech. In this way our data were not entirely consistent with the selective entrainment hypothesis. We did find phase-locked activity associated with both streams, with phase-locked theta power being enhanced for the target relative to distractor stream. An important consideration here is that the phase relationship between target and distractor envelopes might be critically important. Purely by chance, the competing speech streams should exhibit periods of transient coherence. Such coherence between target and distractor might make it particularly difficult to maintain selection and would appear as transient phase-locking between the EEG and the distractor stream. A kind of active distraction might result, in which events in the distractor stream could intrude into representations of the target stream. Our study was not designed to differentiate errors of intrusion from simple failures of recall.

The theta-band signal in the EEG that is phase-locked with the target envelope might exhibit a change in power for various reasons: This increase may be due to increased gain in a fixed-latency evoked response triggered by increments in sound energy within the acoustic signal. Alternatively, ongoing theta-band oscillations might be better phase entrained to the attended envelope, without any modulation in the amplitude

of those oscillations. Either of these situations, or a combination of the two, would appear as a modulation of phase-locked power. Distinguishing between phase entrainment and additive phase-locked activity as mechanisms in the generation of phase-locked signals in scalp-recorded EEG is highly problematic (Telenczuk, Nikulin, & Curio, 2010). Therefore we can conclude only that the neuroelectric dynamics of the brain exhibit components that are phase-aligned with the envelopes of speech in the auditory scene, that these signals are modulated by attentional selection, and that this modulation is reflective of variation in perceptual performance.

3 Theta-band phase tracking in a multi-talker environment

3.1 Introduction

A typical acoustic environment consists of a complex mixture of sounds emitted by a number of discrete sources. The human auditory system routinely selects and perceives a single source from the mixture. This phenomenon and the associated computational challenges are known colloquially as the "cocktail party problem" (Cherry, 1953). Competing streams can occasionally disrupt the perception of the selected stream, even when the streams are separated at the sensory periphery. This disruption has previously been conceptualized as a failure of the selectivity of attention, informational masking (Kidd et al., 2007), or more generally distraction (Ponjavic-Conte, Hambrook, Pavlovic, & Tata, 2013).

There has been an increased effort in understanding the neural mechanisms that allow the parsing of complex auditory scenes. The selective entrainment hypothesis (Schroeder & Lakatos, 2009; Zion Golumbic et al., 2012), proposes that phase-matching of neuroelectric oscillations to low-frequency dynamics of acoustic signals selectively increases cortical sensitivity to the target acoustic stream. Numerous studies have shown that the phase of neuroelectric oscillatory activity can track spectrotemporal modulations in speech (Abrams et al., 2008; Ahissar et al., 2001; Hertrich et al., 2012; Luo & Poeppel, 2007). Further, the focus of attention modulates the selectivity or strength of this tracking process (Ding & Simon, 2012; Kerlin et al., 2010; Mesgarani & Chang, 2012). It has been suggested that matching the phase of oscillatory activity to the dynamics of a target

acoustic stream ensures the brain is in a maximally excitable state when acoustic events in the target stream occur.

There is mounting evidence supporting the selective entrainment hypothesis. Cortical theta-band phase tracking of speech is more pronounced when the speech signal is comprehensible, when speech that is made difficult to understand through acoustic degradation is minimally tracked – phase-tracking of speech is a correlate of successful speech perception. Using intracranial electrocorticography (ECoG), (Zion Golumbic, Ding, et al., 2013) showed that oscillatory activity in auditory cortex track the acoustic envelopes of both target and non-target; contrastingly, activity in medial frontal gyrus exhibited selective tracking of the target stream. This region is known to engage in auditory working memory tasks (Arnott et al., 2005; Crottaz-Herbette et al., 2004) which suggests a role for phase-tracking in linking sensory and memory areas. Phase-tracking was also enhanced when speech was presented simultaneously with video of the speaker suggesting an association between phase-tracking and multimodal sensory integration.

Selective attention in a crowded acoustic scene is known to enhance perception and memory of the attended stream (Broadbent, 1952; Treisman, 1964). Factors that impair perception of a target stream have been widely studied. Relative differences in loudness, spectral, and spatial separation all influence the discriminability of target speech in environments with two competing speakers (Arbogast et al., 2002; Brungart, 2001). These factors, in addition to the overall number of distractors in the scene, also influence perception in environments with more competing speakers (Brungart, Simpson, Ericson, & Scott, 2001; Ericson, Brungart, & Brian, 2004). Interestingly, there is also

evidence that similarity between the temporal envelopes of target and distractor also impairs perception of the target stream (Bronkhorst, 2000).

If phase tracking of speech dynamics is a mechanism for implementing selective attention, then variation in perceptual performance should mirror variation in the strength of speech-locked EEG signals. Furthermore, we should expect a reduction in the strength of tracking as more distractors are added to the scene. In the present study we report that selective listening in a free-field situation, with multiple distractors strengthens a theta-band signal that tracks the acoustic envelope of selected speech, relative to ignored speech. Phase-tracking of the attended signal was reduced as distractor number increased for correct trials but not for errors.

3.2 Methods

3.2.1 Participants

19 undergraduates from the University of Lethbridge 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. 2 participants were excluded from data analysis for failing to respond on a large number of trials (miss rate > 3 standard deviation over the mean across distraction conditions). Thus, 17 participants contributed to the data analysis (12 female; 3 left-handed; average age: 21.0 years).

3.2.2 Stimuli & Task

All stimuli were presented in free field by an Apple Mac Pro with a firewire audio interface (M-Audio Firewire 410). Participants sat of a circular array of near-field studio monitors (Mackie HR624 MK-2) arranged 1 metre away. The target stimuli were presented from a speaker directly in front of the participant. Distractor stimuli were presented from two, four, or six speakers in symmetric locations around the circular array with a minimum separation of 30° of arc between target and maskers (Figure 3.1A). Stimulus presentation was controlled by a program custom coded using Apple Computer's Core Audio framework (Mac OS 10.6).

Each stimulus consisted of the concatenation of 8 sentences, spoken by the same speaker, from the Coordinate Response Measure (CRM) Corpus (Bolia, Nelson, Ericson, & Simpson, 2000). The CRM corpus consists of stereotyped sentences of the format: "Ready <call sign> go to <colour> <number> now," spoken by 4 male and 4 female speakers; "white", "red", "blue", and "green" were the four possible colour targets. On each trial participants were simultaneously played one target stimulus and up to six distractor stimuli, each spoken by a unique speaker. Each block contained 12, 15.5 second stimuli which were divided into pseudo-randomly ordered sub-blocks of 4 stimuli at each level of distraction.

Participants were instructed to listen carefully to the target stream and report, via button press, colour words from the target stream only. Colour words occurred in all streams in close temporal proximity; the standard deviation from the mean latency of colour word onset on a given trial was 60 ms. The CRM corpus contains samples with four different colour words; on any given trial one colour word occurred in the target

stream, two colour words occurred in the distractor streams, and one colour word was left out. In the four and six distractor conditions distractor colour words occurred in more than one stream. Trials for which participants reported the colour from the target stream were considered hits; responses in which participants reported the colour from a distractor stream were labeled intrusion errors; responses in which participants reported a colour that was not present in any stream were labeled insertion errors.

3.2.3 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 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). Data were interpolated to an 81-channel 10-10 montage and exported from BESA and further analyzed in MATLAB (MATLAB version 7.10.0; The Mathworks Inc., 2010, Natick, Massachusetts, 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 was calculated. The acoustic envelope of each sample was calculated by taking the absolute value of the Hilbert transform of the sample and

low-pass filtering 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 that the sum of the signal across the whole epoch equaled 1 (Hertrich et al., 2012). Thus a signal which captures transient energy increases, an aspect of acoustic stimuli to which the auditory system is known to be tuned, was obtained (Fishbach et al., 2001; Howard & Poeppel, 2010). This signal was then cross-correlated with each channel of the time-aligned EEG data to arrive at a cross-correlation function which reflects activity that is phase-locked to acoustic transients in either 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 recorded data to obviate normalization of the cross-correlation function. Trials were labeled based on task performance relative to each target.

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 levels of distraction the power from all 2, 4, or 6 distractor streams was averaged before comparison with power phase-locked to the target stream.

3.3 Results

Participant's ability to identify colour words from the target stream was impaired as the number of distractors in the auditory scene increased (Figure 3.1B). A 3x3 within-subject ANOVA reveals significant main effects of distractor number ($F(2,64)=6.03$,

p=0.006); and response type ($F(2,64)=469.98$, $p<0.001$). There was also a significant interaction between distractor number and response type ($F(4,64)=331.24$, $p<0.001$). Participants were less likely to correctly identify the colour in the target stream, much more likely to make an intrusion error, and slightly more likely to make an insertion error as more distractors were added to the auditory scene.

Previous studies suggested that EEG signals were maximally phase-locked to speech in the theta band (4 – 8 Hz). We used a wavelet time-frequency decomposition to explore the frequency content of the cross-correlation function for target and distractor speech streams. There is a peak in phase-locked power in the theta band around 120 ms lag for all levels of distraction, and for both the target and distractor streams (Figure 3.2). To assess the behavioural importance of this peak in theta power, trials were separated into hits and intrusion errors based on the participant's response for that trial (Figure 3.3). A 3x2x2 within-subject ANOVA performed on theta power at 120 ms lag reveals significant main effects of attention ($F(1,16)=9.70$, $p=0.007$), and performance ($F(1,16)=4.91$, $p=0.042$). There was not a significant main effect of distractor number ($F(2,32)=0.146$, $p=0.865$); however, there was a significant performance*distractor number interaction ($F(2,32)=5.64$, $p=0.008$). The interaction was driven by the decrease in theta power with increasing distraction for hits, and the increase in theta power with increasing distraction for intrusion errors.

The selective entrainment hypothesis states that entrainment to an attended speech stream aligns windows of maximal neuronal excitability to important events in that speech stream. It follows that acoustic events in competing streams will be most distracting when they arrive within these temporal-windows opened by the attended

stream. Thus, we could predict that when the acoustic dynamic signals are more correlated participants would be more likely to make intrusion errors. For each trial we calculated the correlation coefficient between the first-derivative of the acoustic envelope of the attended stream and each of the distractor streams (Figure 3.4). A 3x2 within-subjects ANOVA shows a significant main effect of distractor number on the correlation between the target stream and the most correlated distractor stream ($F(1,16)=1976.68$, $p<0.001$). There was not a significant effect of performance on target-distractor correlation ($F(1,16)=1.124$, $p=0.305$); however, there was a significant interaction between participant performance and distractor number ($F(2,32)=11.276$, $p<0.001$). At low levels of distraction target-distractor correlation is higher on hits, but at higher levels of distraction intrusion errors are associated with higher target-distractor correlation.

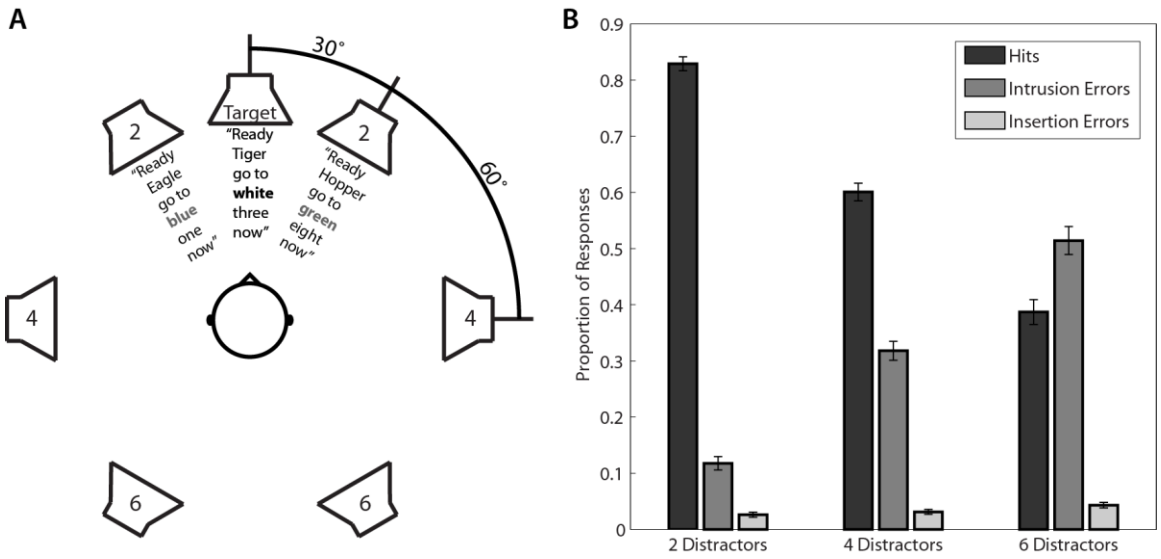

Figure 3.1: (A) Diagram of stimulus presentation array. The frontal midline speaker was the target on all trials. Speakers are labeled with the lowest number of distractors at which they become active. An example trial with 2 distractors is illustrated: a response of "white" would be a hit, a response of "blue" or "green" would be an intrusion error, and a response of "red" would be an insertion error. (B) Mean proportion of responses across subjects at different levels of distraction. Identification of colours in the target stream is impaired as more distractors are added to the auditory scene. Chance probability is 0.5 for intrusion errors and 0.25 for hits and insertion errors.

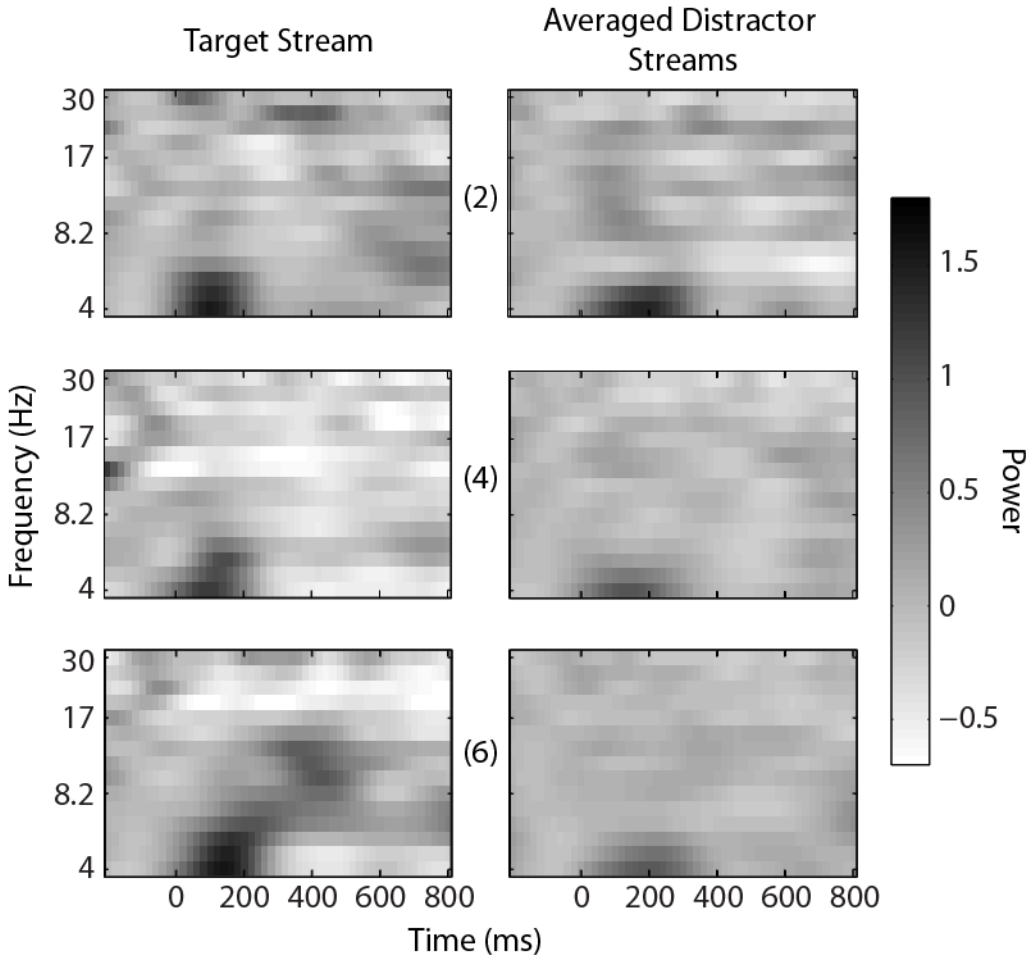

Figure 3.2: Grand average power phase-locked to the acoustic dynamics of target and distractor streams when there are 2, 4, or 6 distractor streams in the acoustic stream. At all levels of distraction a peak in power occurs in the theta band (4-8.2 Hz) at roughly 120 ms.

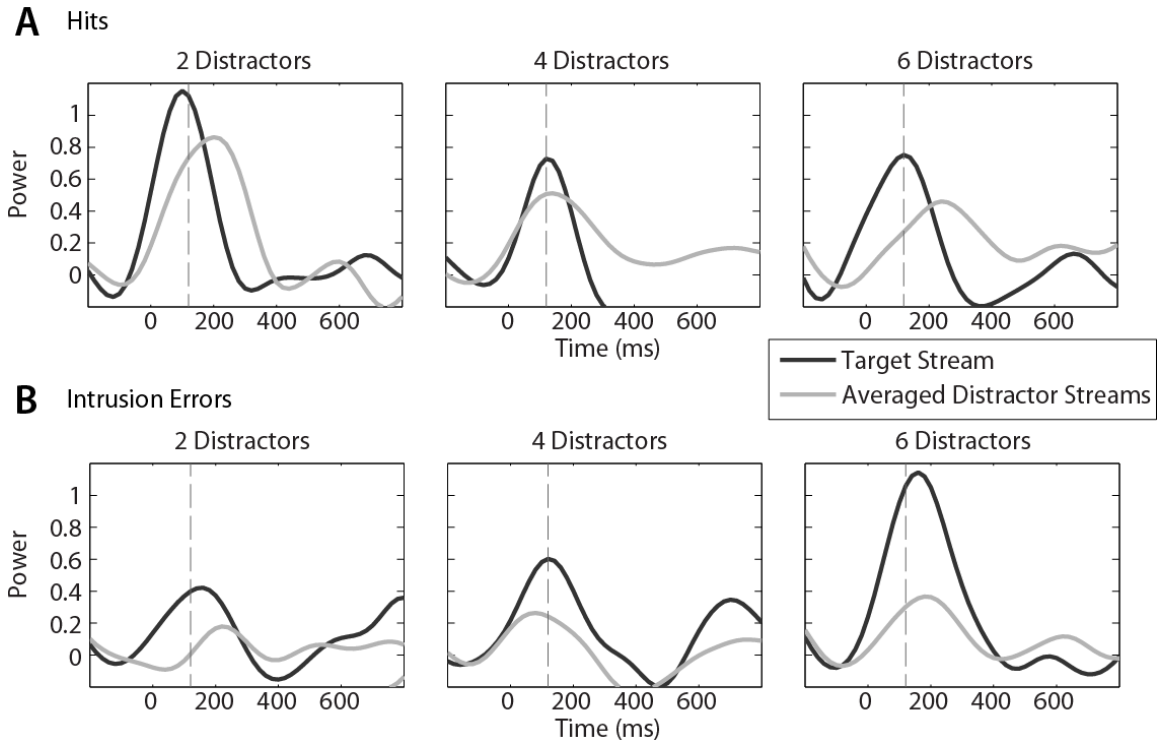

Figure 3.3: Grand average theta (4-8.2 Hz) power phase-locked to target and distractor streams at electrode FCz for (A) hits and (B) intrusion errors. Average peak power latency at 120 ms is indicated with a dashed line. More power is consistently phase-locked to the dynamics of the target stream at all levels of distraction and regardless of performance. Theta power phase-locked to the attended stream decreases as the number of distractors increase for hits, but increases with number of distractors for intrusion errors.

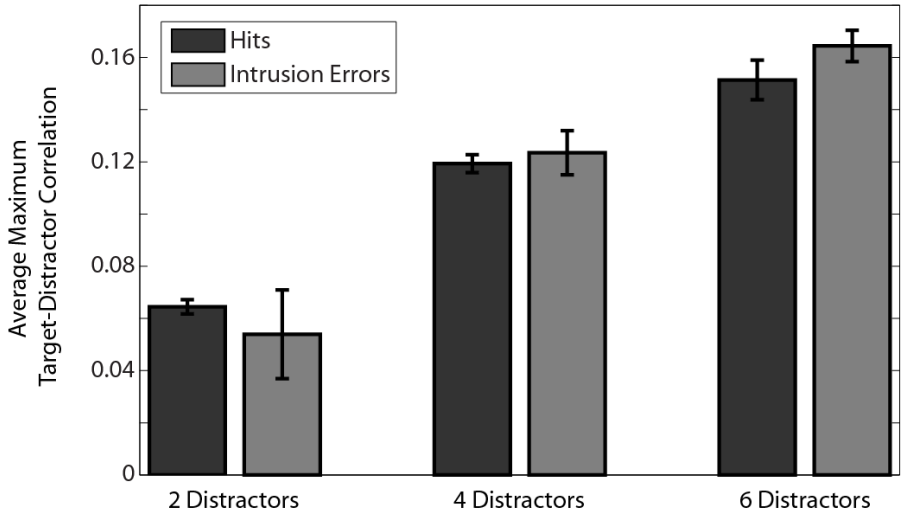

Figure 3.4: Grand average correlation between the acoustic dynamics of the target stream and the dynamics of the most correlated distractor stream for hits and intrusion errors. Target-distractor correlation increases in more crowded acoustic scenes. At higher levels of distraction, higher target-distractor correlation is associated with intrusion errors. Error bars indicate standard error of the mean.

3.4 Discussion

Our behavioral results show a clear effect of distractor number on participants' ability to identify words in a target speech stream. Participants were significantly more likely to identify words from the target stream when there were fewer active distractors, and more likely to make intrusion errors when there were more distractors. Crucially, although insertion errors also increased with the number of distractors in the scene, the relative increase was much less than the increase in intrusion error rate. Because insertion errors involve the apparent percept of an absent stimulus, the insertion error rate indicates the amount of perceptual interference at the sensory periphery. This suggests that while intrusion rate may be influenced by interference at the periphery, the marked increase in intrusion error rate must be due to increased interference at the cortical level.

Our electrophysiogical results show an enhancement of theta band EEG power phase-locked to the acoustic dynamics of target, compared to distractor speech. This enhancement peaked at a lag of 120 ms. This result agrees with previous studies that have found that attention enhances low-frequency activity evoked by continuous speech (Ding & Simon, 2012; Kerlin et al., 2010; Power et al., 2012; Zion Golumbic, Ding, et al., 2013). This enhancement was maintained, even in a very crowded acoustic scene with 6 distractors suggesting that phase-tracking of the acoustic dynamics of a target stream is a generalized mechanism for maintaining the neural representation of that stream.

The selective entrainment hypothesis proposes that phase entrainment of neural oscillations to the temporal dynamics of a behaviourally relevant stream is a mechanism for attentional selection (Schroeder & Lakatos, 2009; Zion Golumbic et al., 2012). The hypothesis relies on the premise that neural sensitivity is modulated by the phase of

ongoing low-frequency neural oscillations. When taken together with the theory of communication by coherence – that is the theory that communication between neuronal assemblies is optimally efficient when graded potentials in pre- and post- synaptic cells are phase-aligned (Fries, 2005). This phase alignment ensures that synaptic transmission occurs within a temporal window of maximal sensitivity to post-synaptic depolarization. By modulating phase, a selective entrainment mechanism may enable a sort of filter – allowing some neural assemblies to ignore inputs from non-selected cells while enhancing sensitivity to selected cells.

Selective attention has been conceptualized as the preferential representation of a single information source for enhanced perception, at the cost impaired perception of other sources. The combination of selective entrainment and communication by coherence may provide a mechanistic explanation for auditory selective attention. Furthermore, it provides a framework by which attended acoustic signals may selectively contribute information to cognitive processes including working memory, reward-processing, and response-planning mechanisms. At the level of networks of cells, assemblies representing features of attended sensory input should be bound together within sensory cortex and receive preferential access to brain-wide associative areas (Malsburg & Schneider, 1986; Singer & Gray, 1995). Selective entrainment – when there is a single source being tracked – not only biases neural networks to respond to the selected stream, but also selectively blocks competing signal from gaining access to associative areas. Selective entrainment may create such a bias by entraining neural oscillations across multiple cortical areas; priming multiple systems to respond to the selected stream.

This view of selective entrainment predicts that perception of the target stream may be disrupted by phase-tracking of a distractor stream or, alternatively, by transient coherence between the acoustic dynamics of the target and the distractor streams. If such disruption is a result of tracking the wrong speech stream, we should expect to see an increase in theta-band power phase-locked to the distractor stream intruding on perception. We found no evidence of such an increase; however, the design of the experiment, in which distractor colour words may appear in more than one distractor stream, limits our ability to identify which stream is intruding on perception. We cannot rule out the possibility that intrusion errors are correlated with an increase in phase-tracking of a distractor stream. It is also possible that tracking of the target stream is maintained despite momentary disruptions in perception and that the intrusion of the distractor stream is due to transient coherence between the target and distractor streams. In such a situation events from the distractor stream may arrive coincident with target events within the same temporal window of maximal neuronal excitability, maintained by phase-tracking of the attended stream. We found limited evidence supporting this explanation: there is, on average, a significantly greater correlation between the acoustic dynamics of the target and most correlated distractor stream on intrusion errors when compared to hits. While this result is suggestive, subsequent experiments will be necessary to confirm that target-distractor coherence produces an active disruption of the perception of the attended stream.

4 Discussion

The goal of this thesis was to test two predictions of the selective entrainment hypothesis: first, that attended speech is preferentially tracked by the phase of theta-band oscillations; and second, that better phase-tracking of a speech stream should be associated with improved behavioural performance.

In an acoustic environment with two competing talkers we found that theta power phase-locked to a target speech stream was enhanced, relative to a non-target distractor speech stream. Furthermore, significant enhancement of phase-tracking of attended speech was limited to trials on which listeners successfully encoded the target speech. These results support two predictions of the selective entrainment hypothesis suggested by Schroeder and Lakatos (2009): that phase-tracking of speech should be stronger for attended speech, and that enhanced phase-tracking of target speech should be associated with improved recall of the target speech.

In a multi-talker environment we found that phase-tracking of target speech was enhanced for all levels of distraction, suggesting that phase-tracking of a speech stream is a generalized mechanism for maintaining the neural representation of that speech. When we considered the data for trials on which subjects made a hit or an intrusion error separately, a puzzling effect emerged. For hits, the tracking of the target stream is attenuated as more distractors are added to the auditory scene, suggesting that maintaining the representation of a target stream was less effective in a crowded environment. However, for intrusion errors, power phase-locked to the target stream appeared to increase as distractors were added. This result is counterintuitive, given the results of our two-talker experiment and other studies, which have suggested that the

strength of phase-tracking is correlated with perception (Luo & Poeppel, 2007; Mesgarani & Chang, 2012; Peelle et al., 2013). Such a result may be explained by a simple but often overlooked characteristic of complex acoustic scenes: as the number of temporally dynamic sound sources in an environment increase, there is a greater chance that one or more of the non-target streams will be – at least transiently – correlated with the target stream. Why might such a correlation give rise to distraction?

The selective entrainment hypothesis proposes that through phase-entrainment of network-wide neural oscillations to a single speech stream, neural assemblies are made maximally sensitive to, important events in that stream. Such entrainment is selective because events in competing streams will arrive at periods of non-optimal neural excitability because they are out of phase with the network. This mechanism is vulnerable to a type of active distraction by permitting a competing stream to access neural assemblies that are configured to respond to the target stream. This occurs if that competing stream shares similar temporal dynamics with the target stream. We refer to this theory as *distraction through coherence*. We suggest that the increase in the strength of phase-tracking of the target stream observed for intrusion errors is caused by transient correlation between the target and distractor streams. In this case the measured response may be amplified as it represents the superposition of acoustic energy from multiple streams. If active intrusions of distractor into perception are caused by transient correlations between the acoustic dynamics of the distractor stream and the tracked target stream, then we should expect to see a greater peak target-distractor correlation on intrusion error trials. This is indeed what we observed.

While these results are suggestive of a mechanism that allows for an active distraction process, more careful study is necessary. In the multi-talker paradigm, we were unable to identify precisely which distractor stream intruded on perception. This step is necessary to conclusively test the prediction that distraction is caused by target-distractor coherence. A future study in which target-distractor coherence is systematically manipulated will allow for the precise identification of the intruding distractor stream is necessary.

We also note that while we have used the terms phase entrained and phase-locked more or less interchangeably, there is a possibility that the observed affects are due to evoked phase-locked activity, rather than pure entrainment of ongoing oscillations. Put another way, the signals observed here could be conceptualized as an ongoing oscillation or as a train of ERP components in the more classical sense. Distinguishing between phase-entrainment and phase-locked evoked activity using EEG is problematic (Telenczuk et al., 2010); however, other studies have found evidence suggesting that the phase-tracking phenomenon is indeed due to phase-entrainment (Giraud & Poeppel, 2012; Obleser et al., 2012).

5 Conclusion

This thesis accomplished several goals: first it confirmed two predictions of the selective entrainment hypothesis: that neuroelectric oscillations in the theta-band of the EEG are more phase-locked to attended relative to unattended speech; and that perceptual performance was modulated by the degree of phase-tracking. In the course of this work, we also developed a novel approach to "unmix" the superposition of brain responses in complex auditory scenes by cross-correlating individual acoustic envelopes with the "mixed" EEG signal. Finally, the results of the experiment described in Chapter Three provide a first tentative suggestion of a novel theory of distraction: *distraction through coherence*. In this theory, "distraction" is not the same as "not attending". Instead it is a more active phenomenon that arises when the dynamics of target and distractor speech envelopes cannot be successfully resolved. This theory, aligned with the theories of selective entrainment (Schroeder & Lakatos, 2009) and communication through coherence (Fries, 2005) provides an exciting starting point for future investigations of auditory attention and distraction.

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. doi: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. doi: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. *J Acoust Soc Am*, *112*(5 Pt 1), 2086–2098. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12430820

Arnott, S. R., Grady, C. L., Hevenor, S. J., Graham, S., & Alain, C. (2005). The functional organization of auditory working memory as revealed by fMRI. *Journal of cognitive neuroscience*, *17*(5), 819–31. doi:10.1162/0898929053747612

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.

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. doi:10.1121/1.428288

Bregman, A. S. (1990). *Auditory scene analysis : the perceptual organization of sound* (p. xiii, 773 p.). Cambridge, MA: MIT Press.

Broadbent, D. E. (1952). Listening to one of two synchronous messages. *Journal of Experimental Psychology*, *44*, 51–55.

Broadbent, D. E. (1958). *Perception and Comunication*. Oxford: Pergamon Press.

Bronkhorst, A. W. (1992). Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *The Journal of the Acoustical Society of America*, *92*(6), 3132. doi:10.1121/1.404209

Bronkhorst, A. W. (2000). The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions. *Acta Acustica united with Acustica*, *86*(1), 117–

128. Retrieved from http://www.ingentaconnect.com/content/dav/aaua/2000/00000086/00000001/art00016

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. doi:10.1121/1.1345696

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. doi:10.1121/1.1408946

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. (K. J. Friston, Ed.)*PLoS computational biology*, *5*(7). doi: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. doi:10.1121/1.1907229

Crottaz-Herbette, S., Anagnoson, R. ., & Menon, V. (2004). Modality effects in verbal working memory: differential prefrontal and parietal responses to auditory and visual stimuli. *NeuroImage*, *21*(1), 340–351. doi:10.1016/j.neuroimage.2003.09.019

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. doi:10.1016/j.jneumeth.2003.10.009

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

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nat Rev Neurosci*, *2*(10), 704–716. doi: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. doi:10.1207/s15327108ijap1403

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), 2303–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11387378

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn Sci*, *9*(10), 474–480. doi:S1364-6613(05)00242-1 [pii] 10.1016/j.tics.2005.08.011

Giraud, A.-L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R. S. J., & Kleinschmidt, A. (2000). Representation of the Temporal Envelope of Sounds in the Human Brain. *J Neurophysiol*, *84*(3), 1588–1598. Retrieved from http://jn.physiology.org/content/84/3/1588.short

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

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. doi:10.1111/j.1469-8986.2011.01314.x

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. doi:10.1152/jn.00251.2010

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 : official publication of the American Electroencephalographic Society*, *19*(2), 113–24. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11997722

James, W. (1890). *The principles of psychology*. New York, NY: Henry Holt and Co.

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral cortex (New York, N.Y. : 1991)*, *18*(7), 1560–74. doi:10.1093/cercor/bhm187

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. doi:10.1523/JNEUROSCI.3631-09.2010

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2007). Informational Masking. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Auditory Perception of Sound Sources* (pp. 143–189).

Lakatos, P., Chen, C.-M., O’Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron*, *53*(2), 279–92. doi:10.1016/j.neuron.2006.12.011

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science*, *110*(2008), 110–3. doi:10.1126/science.1154735

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of neurophysiology*, *94*(3), 1904–11. doi:10.1152/jn.00263.2005

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *The European journal of neuroscience*, *31*(1), 189–93. doi:10.1111/j.1460-9568.2009.07055.x

Lambrecht, J., Spring, D. K., & Münte, T. F. (2011). The focus of attention at the virtual cocktail party--electrophysiological evidence. *Neuroscience letters*, *489*(1), 53–6. doi:10.1016/j.neulet.2010.11.066

Luo, H., Liu, Z., & Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS biology*, *8*(8). doi:10.1371/journal.pbio.1000445

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

Luo, H., Wang, Y., Poeppel, D., & Simon, J. Z. (2006). Concurrent encoding of frequency and amplitude modulation in human auditory cortex: MEG evidence. *Journal of neurophysiology*, *96*(5), 2712–23. doi:10.1152/jn.01256.2005

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*(04), 499–511. Retrieved from http://journals.cambridge.org/abstract_S0140525X98001265

Malsburg, C. Von der, & Schneider, W. (1986). A neural cocktail-party processor. *Biological cybernetics*, *54*, 29–40. Retrieved from http://link.springer.com/article/10.1007/BF00337113

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

Miller, G. A. (1947). The masking of speech. *Psychological bulletin*, *44*(2), 105–29. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20288932

Münte, T. F., Spring, D. K., Szycik, G. R., & Noesselt, T. (2010). Electrophysiological attention effects in a virtual cocktail-party setting. *Brain research*, *1307*, 78–88. doi:10.1016/j.brainres.2009.10.044

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. doi:10.3389/fnhum.2012.00250

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*, 1378–1387. doi:10.1093/cercor/bhs118

Pollack, I. (1975). Auditory informational masking. *The Journal of the Acoustical Society of America*, *57*(S1), S5. doi:10.1121/1.1995329

Ponjavic-Conte, K. D., Hambrook, D. A., Pavlovic, S., & Tata, M. S. (2013). Dynamics of distraction: competition among auditory streams modulates gain and disrupts inter-trial phase coherence in the human electroencephalogram. (M. J. Chacron, Ed.)*PloS one*, *8*(1), e53953. doi:10.1371/journal.pone.0053953

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. doi:10.1111/j.1460-9568.2012.08060.x

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

Seither-Preisler, A., Krumbholz, K., Patterson, R. D., Seither, S., & Lutkenhoner, B. (2004). Interaction between the neuromagnetic responses to sound energy onset and pitch onset suggests common generators. *European Journal of Neuroscience*, *19*(April), 3073–3080. doi:10.1111/j.1460-9568.2004.03423.x

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience*, *18*, 555–86. doi:10.1146/annurev.ne.18.030195.003011

Suppes, P., Han, B., Epelboim, J., & Lu, Z.-L. (1999). Invariance between subjects of brain wave representations of language. *Proceedings of the National Academy of Sciences*, *96*(22), 12953–12958. doi:10.1073/pnas.96.22.12953

Telenczuk, B., Nikulin, V., & Curio, G. (2010). Role of neuronal synchrony in the generation of evoked EEG/MEG responses. *Journal of neurophysiology*, *104*(6), 3557–3567. doi:10.1152/jn.00138.2010

Treisman, A. M. (1964). The Effect of Irrelevant Material on the Efficiency of Selective Listening. *American Journal of Psychology*, *77*, 533–546. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=14251963

Treisman, A. M. (1969). Strategies and models of selective attention. *Psychological Review*, *76*(3), 282–299. Retrieved from

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=4893203

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. doi:10.1523/JNEUROSCI.3675-12.2013

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., … Schroeder, C. E. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party." *Neuron*, *77*(5), 980–991. doi: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. doi:10.1016/j.bandl.2011.12.010