BOSTON UNIVERSITY
SCHOOL OF MEDICINE

Dissertation

SELECTIVE ATTENTION AND SPEECH PROCESSING IN THE CORTEX

by

SIDDHARTH RAJARAM
M.Math., University of Waterloo, 2007
B.A., Middlebury College, 2006

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
2014
Approved by

First Reader

Barbara G. Shinn-Cunningham, Ph.D.
Professor of Biomedical Engineering

Second Reader

Adrian KC Lee, Sc.D.
Assistant Professor of Speech and Hearing Sciences
University of Washington

Third Reader

Uri T. Eden, Ph.D.
Assistant Professor of Mathematics and Statistics
Acknowledgments

I owe several debts of gratitude to all the people who have supported me along this journey. First and foremost, I would like to thank Barbara Shinn-Cunningham for her mentorship and guidance throughout my time in graduate school. Her patience and generosity in allowing me to discover my interests by switching projects multiple times and exploring vastly different approaches, even in areas beyond her realm of expertise, has allowed me to grow as a scientist that would not have been possible under anyone else. Barb’s lively nature and enthusiasm for science is responsible for an amazing lab environment that I feel lucky and privileged to have been a part of for so many years. I will always look back fondly on my years in graduate school thanks to her incredibly supportive guidance.

For all her love and support over the past four years, I thank my girlfriend Diana Gergel. Her presence throughout this process, especially when the going got rough, has been critical in keeping me grounded and sane. Sharing my life with her outside of the lab has allowed me to put things in perspective and I consider myself lucky to have had her in my life throughout graduate school and going forward.

KC Lee has been a fantastic friend, mentor, collaborator, and committee member. He patiently introduced me to neuroimaging, which forms the basis of my entire dissertation; so for that I owe him a big thank you. I have greatly enjoyed his cheerful company over the years and he has played a huge role in my time in graduate school. I would also like to thank the rest of my committee: Edmund Lalor, Malcolm Slaney, and Uri Eden, for their time and input over the past few years. My discussions with Ed and Malcolm starting in Telluride have greatly influenced my modeling work and general understanding of speech perception. Uri’s inputs have helped me gain an insight into the statistical modeling process. I could not have wished for a more supportive committee with a better combination of expertise, that allowed me to achieve my research goals.

Lab members, current and past, have been immensely helpful by providing direct help
with experiments, useful input and feedback, as well as a nurturing environment for doing research. In particular, I would like to thank Erol Ozmeral, who helped me get started with work in the lab; Tim Streeter, Lenny Varghese, and Arash Fazl for much needed technical help at times; Hari Bharadwaj for all the statistics and signal processing help; Scott Bressler, Inyong Choi and Sam Mathias for stimulating discussions. I also owe much to collaborators over the years: Gin Best and Erick Gallun for all their help with my very first experiment in the lab, Steve Colburn for his support, Eric Larson for his help with MNE, and my Woods Hole collaborators Melissa Coleman, Eric Fortune, and Noah Cowan. I would also like to thank my friends in the old CNS department/GPN program for invigorating discussions and conversations over the years - particularly Jeff Markowitz and Misha Panko. Michio Painter’s friendship has been invaluable over the years, especially in ensuring I had a life outside the lab.

Finally, I would like to thank my family - my parents and sister Shalini, as well as my grandparents and extended family for their love and support over the years.
ABSTRACT

In noisy and complex environments, human listeners must segregate the mixture of sound sources arriving at their ears and selectively attend a single source, thereby solving a computationally difficult problem called the cocktail party problem. However, the neural mechanisms underlying these computations are still largely a mystery. Oscillatory synchronization of neuronal activity between cortical areas is thought to provide a crucial role in facilitating information transmission between spatially separated populations of neurons, enabling the formation of functional networks.

In this thesis, we seek to analyze and model the functional neuronal networks underlying attention to speech stimuli and find that the Frontal Eye Fields play a central ‘hub’ role in the auditory spatial attention network in a cocktail party experiment. We use magnetoencephalography (MEG) to measure neural signals with high temporal precision, while sampling from the whole cortex. However, several methodological issues arise when undertaking functional connectivity analysis with MEG data. Specifically, volume conduction of electrical and magnetic fields in the brain complicates interpretation of results. We compare several approaches through simulations, and analyze the trade-offs among various measures of neural phase-locking in the presence of volume conduction. We use these insights to study functional networks in a cocktail party experiment.

We then construct a linear dynamical system model of neural responses to ongoing
speech. Using this model, we are able to correctly predict which of two speakers is being attended by a listener. We then apply this model to data from a task where people were attending to stories with synchronous and scrambled videos of the speakers’ faces to explore how the presence of visual information modifies the underlying neuronal mechanisms of speech perception. This model allows us to probe neural processes as subjects listen to long stimuli, without the need for a trial-based experimental design. We model the neural activity with latent states, and model the neural noise spectrum and functional connectivity with multivariate autoregressive dynamics, along with impulse responses for external stimulus processing. We also develop a new regularized Expectation-Maximization (EM) algorithm to fit this model to electroencephalography (EEG) data.
## Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>i</td>
</tr>
<tr>
<td>Copyright</td>
<td>ii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>iv</td>
</tr>
<tr>
<td>Abstract</td>
<td>vi</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>x</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xii</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>xiii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2 Functional Networks in Auditory Selective Attention</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>4</td>
</tr>
<tr>
<td>2.2 Measures of Neural Functional Connectivity</td>
<td>7</td>
</tr>
<tr>
<td>2.2.1 Functional Connectivity in the Presence of Volume Conduction</td>
<td>10</td>
</tr>
<tr>
<td>2.2.2 Correction with Empty Room Measurements</td>
<td>11</td>
</tr>
<tr>
<td>2.2.3 Measures Accounting for Volume Conduction</td>
<td>13</td>
</tr>
<tr>
<td>2.3 Evaluating Connectivity Measures with Simulations</td>
<td>15</td>
</tr>
<tr>
<td>2.4 ERP subtraction</td>
<td>19</td>
</tr>
<tr>
<td>2.5 Space vs. Pitch Experiment</td>
<td>20</td>
</tr>
<tr>
<td>2.5.1 Methods</td>
<td>21</td>
</tr>
<tr>
<td>2.5.2 Functional Connectivity Analysis</td>
<td>23</td>
</tr>
<tr>
<td>2.5.3 Selection of ROIs and Frequency Bands for Functional Connectivity</td>
<td>25</td>
</tr>
</tbody>
</table>
3 A Dynamical System Model for Tracking Ongoing Speech

3.1 Introduction .................................................. 38
3.2 Methods ......................................................... 40
3.3 Linear Dynamical System Model .................................. 40
  3.3.1 Autoregressive Dynamics .................................. 40
  3.3.2 State Space Formulation .................................. 42
3.4 System Identification ............................................. 44
  3.4.1 Regularized EM Algorithm .................................. 45
  3.4.2 Uniqueness of Parameter Estimates ........................ 48
  3.4.3 Initialization using CCA .................................. 48
3.5 Model Validation on Data ....................................... 49
  3.5.1 Data Acquisition and Stimuli ............................. 49
  3.5.2 Data Preprocessing ....................................... 50
  3.5.3 Model Specification and Inference Parameters ............ 50
3.6 Results .......................................................... 51
3.7 Interpretation of Estimated Parameters .......................... 52
3.8 Conclusions and Extensions .................................... 58

4 Audio–Visual Speech ........................................... 60
4.1 Introduction .................................................... 60
4.2 Methods ........................................................ 63
  4.2.1 Stimulus and Task ......................................... 63
  4.2.2 Data Collection and Pre-Processing ....................... 64
  4.2.3 Model Parameter Estimation ............................... 65
  4.2.4 Analysis of Inferred Parameters .......................... 66
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3 Results</td>
<td>68</td>
</tr>
<tr>
<td>4.4 Discussion</td>
<td>72</td>
</tr>
<tr>
<td>4.5 Conclusions</td>
<td>73</td>
</tr>
<tr>
<td>5 Conclusions</td>
<td>75</td>
</tr>
<tr>
<td>5.1 Potential Applications</td>
<td>76</td>
</tr>
<tr>
<td>Bibliography</td>
<td>78</td>
</tr>
<tr>
<td>Curriculum Vitae</td>
<td>91</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Test Performance by Subject ........................................... 53

4.1 Synchronous and Asynchronous Test Performance by Subject .... 69
List of Figures

2.1 Mean alpha coherence maps . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
2.2 Illustration of the effects of inverse imaging point spread . . . . . . . . 16
2.3 Sensitivity of the various functional connectivity measures . . . . . . . 18
2.4 Functional network in the alpha band, post-stimulus onset . . . . . . . 31
2.5 Functional network in the beta band, post-stimulus onset . . . . . . . . 32
2.6 Functional network in the alpha band, pre-stimulus onset . . . . . . . . 33
2.7 Functional network in the beta band, pre-stimulus onset . . . . . . . . 34
3.1 Graphical structure of ARX model . . . . . . . . . . . . . . . . . . . . . . 42
3.2 Performance of a Kalman-filter based classifier . . . . . . . . . . . . . 52
3.3 Estimated state-space model parameters for all subjects . . . . . . . . 54–57
4.1 Estimated model parameters during audio-only and audio–visual listening . 69–71
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>Autoregressive</td>
</tr>
<tr>
<td>ARX</td>
<td>Autoregressive with eXogenous inputs</td>
</tr>
<tr>
<td>BEM</td>
<td>Boundary Element Model</td>
</tr>
<tr>
<td>BOLD</td>
<td>Blood Oxygenation Level Dependent</td>
</tr>
<tr>
<td>CCA</td>
<td>Canonical Correlation Analysis</td>
</tr>
<tr>
<td>dB</td>
<td>decibels</td>
</tr>
<tr>
<td>DCM</td>
<td>Dynamic Causal Modeling</td>
</tr>
<tr>
<td>dSPM</td>
<td>dynamic Statistical Parametric Mapping</td>
</tr>
<tr>
<td>EEG</td>
<td>Electroencephalography</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation-Maximization</td>
</tr>
<tr>
<td>EOG</td>
<td>Electro-oculography</td>
</tr>
<tr>
<td>ERF</td>
<td>Event Related Field</td>
</tr>
<tr>
<td>ERP</td>
<td>Event Related Potential</td>
</tr>
<tr>
<td>FEF</td>
<td>Frontal Eye Field</td>
</tr>
<tr>
<td>FLASH</td>
<td>Fast Low Angle Shot</td>
</tr>
<tr>
<td>fMRI</td>
<td>functional Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>HPI</td>
<td>Head Position Indicator</td>
</tr>
<tr>
<td>HRTF</td>
<td>Head-Related Transfer Function</td>
</tr>
<tr>
<td>iFEF</td>
<td>inferior Frontal Eye Field</td>
</tr>
<tr>
<td>IFJ</td>
<td>Inferior Frontal Junction</td>
</tr>
<tr>
<td>IIR</td>
<td>Infinite Impulse Response</td>
</tr>
<tr>
<td>infPCS</td>
<td>inferior Pre-Central Sulcus</td>
</tr>
<tr>
<td>IPS</td>
<td>Intra-Parietal Sulcus</td>
</tr>
<tr>
<td>LDS</td>
<td>Linear Dynamical System</td>
</tr>
<tr>
<td>LFP</td>
<td>Local Field Potential</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>LTI</td>
<td>Linear Time-Invariant</td>
</tr>
<tr>
<td>MEG</td>
<td>Magnetoencephalography</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>MNE</td>
<td>Minimum Norm Estimate</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>MVAR</td>
<td>Multivariate Autoregressive</td>
</tr>
<tr>
<td>MVARX</td>
<td>Multivariate Autoregressive with eXogenous inputs</td>
</tr>
<tr>
<td>NARX</td>
<td>Nonlinear Autoregressive with eXogenous inputs</td>
</tr>
<tr>
<td>NFIR</td>
<td>Nonlinear Finite Impulse Response</td>
</tr>
<tr>
<td>OLS</td>
<td>Ordinary Least Squares</td>
</tr>
<tr>
<td>PCS</td>
<td>Pre-Central Sulcus</td>
</tr>
<tr>
<td>PGC</td>
<td>Partial Granger Causality</td>
</tr>
<tr>
<td>PLI</td>
<td>Phase Lag Index</td>
</tr>
<tr>
<td>PLV</td>
<td>Phase Locking Value</td>
</tr>
<tr>
<td>PPC</td>
<td>Pairwise Phase Consistency</td>
</tr>
<tr>
<td>RMS</td>
<td>Root Mean Squared</td>
</tr>
<tr>
<td>ROI</td>
<td>Region of Interest</td>
</tr>
<tr>
<td>RTS</td>
<td>Rauch-Tung-Striebel</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SPL</td>
<td>Sound Pressure Level</td>
</tr>
<tr>
<td>SQUID</td>
<td>Superconducting Quantum Interference Device</td>
</tr>
<tr>
<td>SSP</td>
<td>Signal Space Projection</td>
</tr>
<tr>
<td>SSS</td>
<td>Signal Space Separation</td>
</tr>
<tr>
<td>STG</td>
<td>Superior Temporal Gyrus</td>
</tr>
<tr>
<td>STRF</td>
<td>Spectro-Temporal Response Function</td>
</tr>
<tr>
<td>STS</td>
<td>Superior Temporal Sulcus</td>
</tr>
<tr>
<td>TMS</td>
<td>Transcranial Magnetic Stimulation</td>
</tr>
<tr>
<td>WPLI</td>
<td>Weighted Phase Lag Index</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Human listeners must be able to parse complex auditory scenes with multiple sound sources into its constituent components, or *auditory objects* (Shinn-Cunningham, 2008). This parsing procedure is called *auditory scene analysis* (Bregman, 1994) and normal-hearing human listeners are able to perform this computationally challenging task with relative ease (Shinn-Cunningham & Best, 2008). The ability to solve this problem is particularly astute for ecologically relevant stimuli, most notably speech (for humans). Speech signals have a rich spectro-temporal structure that support the formation of rich hierarchical structures of objects: phonemes, syllables, words, sentences. This hierarchy allows us to optimally combine information from acoustic, lexical, semantic, syntactic and grammatical cues. This problem of unmixing speech signals and processing one speaker while ignoring or even suppressing the others was originally described as the *cocktail party problem* in Cherry (1953) and has garnered great interest, particularly in recent years.

Attention is known to play an important role in cocktail party situations, and auditory scene analysis in general. In particular, attention is not only necessary to select an object, but can effect the formation of the perceptual object itself. Despite widespread interest in the field, we are still in the early stages of uncovering the underlying neural mechanisms of attention, particularly in the auditory domain. Studies of attentive speech perception are also limited, as the vast majority of human neurophysiological studies are performed non-invasively.

The majority of neurophysiological studies in humans look at first-order effects, or changes in neural activity across neural regions, corresponding to a particular task. Such
analyses give us an incomplete picture of processing in the brain, as they only let us infer sets of regions acting independently, whereas second-order effects, i.e., communication amongst regions, allow cortical regions to perform dedicated functions and send appropriate priming cues to other regions further along the processing pathway. The central theme of this thesis is to characterize neural function during attentive processing of speech, with both first and second-order effects. To do so, we will first need to develop computational and statistical tools that enable such analyses.

Oscillations are known to play an important role in neural processing by modulating neuronal excitability in a temporally predictable way. This property is also beneficial for establishing neural communication across different, possibly spatially distant, cortical regions. Oscillatory activity and synchrony in specific frequency bands have been found to be implicated in several different cognitive tasks all across the cortex (Siegel et al., 2012; Buzsaki, 2004). Since we require fine temporal resolution to distinguish activity in different frequency bands, we use MEG and EEG to perform imaging, rather than fMRI.

The first study in this thesis, described in Chapter 2, analyzes functional connectivity elicited specifically by attention to a location in space, as compared to attention to another auditory feature (here, the pitch of the speaker). The data, and first-order analysis for this experiment have been previously reported in Lee et al. (2013). In that study, cortical activity was estimated from activity in the sensors using the minimum norm estimate (MNE) of neural activity (Hämäläinen et al., 1993). In order to effectively compare the first and second order effects, we use the same inverse imaging technique, but in doing so we encounter technical hurdles due to the spread of activation inherent with MNE inverse imaging. We compare statistical methods that account for this artifact and discuss other issues in measuring functional connectivity. Finally, we use the appropriate methods to study the functional network, or pairs of connected cortical regions, underlying auditory spatial attention. A preliminary version of the simulations and data analysis performed in this chapter has been presented in a conference paper, Rajaram et al. (2011).

While the analysis above provides a descriptive, or exploratory account of the data,
it does not describe a mechanism which gives rise to the observed data. In Chapter 3, we describe a neural model of processing ongoing speech that simultaneously models neural activity as an autoregressive model with exogenous inputs (ARX). This model offers the benefit of being able to model neural impulse responses with natural ongoing speech stimulus, instead of only looking at event related potentials (ERPs) to sound onsets. Additionally, the parameters of this model correspond to physiological quantities that can be interpreted directly, and constrained by prior information. The model is also expressed in a state-space notation, allowing potential use of Kalman filters for efficient online implementations. We also present a modified expectation maximization (EM) algorithm to fit this model, and fit the model to experimental data from Power et al. (2012). This chapter is an expanded version of the conference paper, Rajaram et al. (2013).

Finally, in Chapter 4, we present an application of this model to the analysis of audio-visual speech. We constrain our neural sources to correspond to auditory and visual cortices and study how the auditory impulse response changes while processing multi-modal speech, as compared to audio-only speech with meaningless visual input. We also use the fitted model to perform connectivity analysis to compare directed (causal) information flow between the modeled regions.
Chapter 2

Functional Networks in Auditory Selective Attention

2.1 Introduction

Oscillatory synchronization of neuronal activity between cortical areas is thought to provide a crucial role in facilitating information transmission between spatially separated populations of neurons (Siegel et al., 2012; Buzsaki, 2004). This mechanism enables the formation of functional networks, or transient collections of neural populations that communicate with each other, despite being distributed across distinct spatial regions of the cortex. These functional networks are typically associated with cognitive functions, such as selective attention (Siegel et al., 2008; Gross et al., 2004), working memory (Palva et al., 2010), conscious perception of stimuli (Hipp et al., 2011), the default or “resting state” of the brain (Buckner et al., 2008), and sensorimotor decision making (Heekeren et al., 2008). We note here that functional connectivity is distinct from anatomical connectivity (projection of axons from one region to another). Axonal projections provide the fixed (at least in the time scales of interest for cognitive tasks) substrate upon which functional networks are dynamically created and destroyed based on stimuli and task demands.

In addition to specificity of cortical areas, these functional networks tend to coordinate in a frequency-specific manner. A mechanistic explanation for this is provided by the communication-through-coherence hypothesis (Fries, 2005). This hypothesis states that since activated neurons naturally tend to oscillate due to their intrinsic electrical dy-
namics, phase-locked oscillations can rhythmically coordinate the opening and closing of temporal windows of excitability, as determined by the relative phases of the oscillations. This relative phase, or difference in phases, corresponds to the time taken by the action potential to propagate through the axon (or the total spike travelling time if there are intermediary neurons in the communication channel). The hypothesis that spike propagation times between regions determine the time-scales of coherent oscillations is consistent with the general findings that synchrony among nearby regions, or within the same region, typically occurs at higher frequencies, whereas distal communication tends to occur at lower frequency bands. However, we also note that recently there have been reports of long-range high-frequency synchronization (Gregoriou et al., 2009). There have also been several reports of cross-frequency (non-linear) synchronization of neural populations in the literature, enabled by neural processes such as entrainment of higher frequencies by lower frequencies (Lakatos et al., 2008), and nesting of oscillations (Kopell et al., 2011).

There have been several invasive electrophysiological studies in animals that explore coordination of oscillations across brain regions (Gregoriou et al., 2009; Lakatos et al., 2008; Gregoriou et al., 2012). These studies typically measure synchrony between the spikes in two regions, local field potentials (LFPs) in two regions, or the synchrony between the LFP in one region and spikes in another, a measure called the spike-field coherence (Canolty et al., 2010). While such invasive studies provide invaluable perspectives within the fine-scale neural dynamics, they have the disadvantage of being able to measure only a few (usually two) cortical regions at a time. Thus, we are often left to combine network connections from data obtained with different subjects and experimental designs, making inferences about networks extremely difficult.

Several functional connectivity studies in humans have used functional Magnetic Resonance Imaging (fMRI) (Newton et al., 2007; Fox et al., 2006). The blood oxygenation level dependent (BOLD) signal of fMRI measures changes in blood flow and blood oxygenation in the brain, which is known to be correlated with neural activity (Leopold et al., 2003); but the exact mapping of neural activation patterns to BOLD signals (called the hemodynamic
response function) is still largely unknown. In addition, functional connectivity with fMRI can only reveal correlations in overall power fluctuations at very slow (< 0.1 Hz) rates (or specifically gamma power, according to Nir et al. (2007)), rather than phase-locked activity across a range of different oscillation frequencies.

We therefore choose to explore auditory attentional networks using magnetoencephalography (MEG). A majority of past electroencephalography (EEG) and MEG studies have conducted connectivity analysis between pairs of sensors (Doesburg et al., 2008). However, because the sensor responses reflect a mixture of activity from underlying sources, these studies provide limited information about the cortical regions in the network that contribute to these responses. Therefore we first estimate neural source activity, and then analyze the connectivity structure amongst these source activity estimates. The distributed-source inverse imaging is performed with cortically constrained minimum norm estimation (MNE) (Hämäläinen et al., 1993) to estimate neural activity across source dipoles distributed over the entire cortical surface. We then analyze functional connectivity using these estimates. Such an analysis will allow us to compare and contrast connectivity results with the evoked responses. The analysis allows for rather straightforward interpretation of the MNE inverse solution, which makes minimal assumptions on the structure of the data. However, since volume conduction of electric and magnetic fields in the brain leads to a spatial spread of estimated source activity, we may detect connectivity where there is none. We mitigate the effect of such spurious connectivity by using an appropriate metric that is relatively insensitive to these volume conduction effects. Note that using alternate inverse solutions to obtain distributed source estimates involve the same issues, and the discussion in this chapter is applicable to these solutions as well.

One family of approaches for estimating functional connectivity operates in the time domain and involves modeling the multivariate time series of neural activation, typically using an autoregressive (AR) model, similar to the approach we use in Ch. 3. We may extract measures such as Granger causality, or its frequency-domain analogs such as partially directed coherence, or directed transfer function (Schoffelen & Gross, 2009). These
AR models may be fit to estimated neural source waveforms after inverse imaging, or directly fit to sensor space data and projected to source space (Michalareas et al., 2013) for exploratory analysis of connectivity. One significant disadvantage of this approach is that the multivariate AR model has a large number of parameters, so therefore requires vast amounts of data for inference. Another popular approach for directly modeling the time series is called *dynamic causal modeling* (DCM). DCM makes strong assumptions about the underlying biophysical mechanisms, using the idea of neural fields (David & Friston, 2003).

The approach we take to measure connectivity operates in the frequency domain, involving the distribution of the relative phase between neural sources. In this chapter, we explore the issues surrounding inference of neural source connectivity using frequency-domain measures. We begin by introducing various phase-based connectivity metrics and the problem of inferring connectivity in the presence of volume conducting sources in Section 2.2. In Section 2.3, we perform a simulation study to compare the robustness of these metrics to volume conduction artifacts. We discuss the issue of evoked-response subtraction with regards to our choice of statistical measure in Section 2.4. Finally, we apply our analysis to the inference of functional networks controlling spatial attention in a cocktail party listening task and discuss the resultant networks in Section 2.5.

### 2.2 Measures of Neural Functional Connectivity

The two most common phase-based measures for assessing oscillatory synchronization are *coherence* and *phase-locking value (PLV)* (Dobie & Wilson, 1993). Both these measures yield real-valued numbers between 0 and 1, where larger values imply more phase-locked synchrony across trials. It is important to observe that both these quantities are closer to 1, when the *relative phase*, or phase difference between the two sources, does not vary across trials. They are insensitive to the actual phase at which the synchronization occurs.

The *PLV* may be best understood as the magnitude of the normalized relative phase
vectors (or complex numbers), averaged across all trials. To express this mathematically, we define $Z_{1,i}$ and $Z_{2,i}$ as the complex-valued Fourier coefficients, in trial $i$, of the two signals we are comparing. Then, the cross-spectrum of these two signals in trial $i$ is given by $X_i = Z_{1,i} \cdot Z_{2,i}^*$, where $*$ represents the complex conjugate operator. The normalized relative phase vector for trial $i$ is defined as $\omega_i = X_i / ||X_i||$. Then, given $n$ trials of data, the PLV is defined as

$$PLV := \left| \frac{1}{n} \sum_{i=1}^{n} \omega_i \right| = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{|X_i|} \right| = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{Z_{1,i} \cdot Z_{2,i}^*}{|Z_{1,i}| \cdot |Z_{2,i}|} \right|. \quad (2.1)$$

Note that different approaches to computing PLV directly compute the phases of the sources using Hilbert transform techniques (Le van Quyen et al., 2001). Several approaches also exist to estimate Fourier coefficients, including wavelet methods (Le van Quyen et al., 2001) and multi-taper spectral analysis (Thomson, 1982). The phase properties of these estimators have also been characterized (Lepage et al., 2013).

The coherence can then be viewed as a weighted average of relative phase vectors; the PLV weights the phases all trials equally, while in the coherence measure, trials are weighted in proportion to the amplitude of the Fourier coefficients:

$$\text{Coh} := \left| \sum_{i=1}^{n} Z_{1,i} \cdot Z_{2,i}^* \right| / \sum_{i=1}^{n} |Z_{1,i}| \cdot |Z_{2,i}| = \left| \sum_{i=1}^{n} X_i \right| / \sum_{i=1}^{n} |X_i| = \left| \frac{1}{n} \sum_{i=1}^{n} v_i \omega_i \right|, \quad \text{where} \quad v_i = \frac{n |X_i|}{\sum_{i=1}^{n} |X_i|}. \quad (2.2)$$

We can classify all the phase-based metrics as either being either weighted (coherence-like) or unweighted (PLV-like) measures.

One of the problems with both these statistics is that sample coherence and PLV are biased estimates of the respective population statistics (Brillinger, 1981; Amjad et al., 1997). The complex-valued quantity within the outer absolute value operator in the definition of coherence in (2.2) is called coherency, and is an unbiased complex-valued estimator of phase-locking. But the use of complex statistics makes interpretation and hypothesis testing difficult, so we take the length of the resultant vector to get a simple index between 0
and 1. The sample coherence and PLV are functions of sample size; more trials in computing the sample statistics leads to smaller biases. Some attempts (Bokil et al., 2007) at mitigating the effects of these biases use analytic expressions to estimate the bias and correct appropriately. However, other studies (Maris et al., 2007) have pointed out that these estimates only hold for coherence (and PLV) values that are much larger than those falling in the range typically seen in neurophysiological data.

A new measure has been proposed (Vinck et al., 2010) that generalizes the PLV and is bias-free: the *pairwise phase consistency* (PPC). This statistic is derived from the observation that for any pair of distinct trials with relative phase angles $\theta_i$ and $\theta_j$, $\cos(\theta_i - \theta_j)$ is an unbiased estimator of the square of the population squared PLV. Note here that the relative phase angles $\theta_i$ and $\theta_j$ are the angles subtended by the relative phase vectors $\omega_i$ and $\omega_j$, respectively. Thus, if we take the mean of this measure over all pairs of trials, we still get an unbiased estimator of PLV-squared:

$$\text{PPC} := \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \cos(\theta_i - \theta_j). \quad (2.3)$$

Note that this is an unweighted measure, since it discards the magnitude information in the cross-spectrum. Also, this measure can take on values between $-1$ and $1$. Negative values are harder to interpret, and would arise in cases such as a bimodal distribution of relative phases.

The measures presented here are all insensitive to the relative phase at which synchronization occurs. However, we will show in the following section that due to volume conduction inherent in M/EEG measurements, along with our inverse imaging technique, we observe phase-locking between sources where none exists. We demonstrate this false-positive detection of synchrony and discuss various approaches to mitigate its effects in our estimates.
2.2.1 Functional Connectivity in the Presence of Volume Conduction

M/EEG scalp measurements have poor spatial resolution due to large volume conduction effects in the brain. A particular concern for EEG is the anisotropy caused by the dura, skull and scalp (Nunez & Srinivasan, 2005). Due to this loss in spatial information, and to a lesser extent the comparatively small number of M/EEG sensors (relative to possible dipolar neural sources), the M/EEG inverse problem is inherently ill-posed (Hämäläinen et al., 1993). A minimum-norm estimate of neural activity makes minimal assumptions on the structure of the data, but minimizes the \(l_2\)-norm of estimated activity. This produces smoothed, but not sparse, estimates of activity. Due to this smoothing, neural activity estimates might contain additive components from far away sources. This can cause us to infer that there is phase-locking between sources, where none actually exists.

As the frequencies of interest in M/EEG are well under 1 kHz, we can assume a quasi-stationary model of the Maxwell equations (Plonsey & Heppner, 1967; Hämäläinen et al., 1993). This implies that the effects of volume conduction are instantaneous and do not occur with a phase-shift (rather, any phase shift across sources is negligibly small and can be safely ignored). Volume conduction can, however, lead to a negative mixing coefficient, since the two lobes of a dipole field have opposite signs. This implies that any phase-locking that occurs due to volume conduction (or spread of MNE activations) must occur at phases of either 0 or \(\pi\). All the connectivity measures discussed here operate by discounting phase-locking at or close to these phases. It must be noted that these approaches are conservative, in that if neural sources actually do phase-lock at 0 or \(\pi\) phases, as is known to occur (Gollo et al., 2011), they will not be detected. However, if we use multiple frequency bins to estimate the cross-spectrum, then under the assumption of a fixed time-lag model, the phase at which signals lock should vary as a function of the frequency. Some frequency bins in the band may still detect phase-locking; this is a function of the time-lag, frequency band, and number of bins used in the analysis.
2.2.2 Correction with Empty Room Measurements

To illustrate the effects of point spread in MNE inverse imaging, we present raw coherence maps that we first reported in a preliminary study (Rajaram et al., 2011). The approach used here was to empirically estimate a baseline measure of coherence by calculating coherence with measurements made in an empty room (without a subject present), an approach that has been used for studies of resting-state coherence measurements (Ghuman et al., 2011). This baseline coherence estimate is then subtracted from the experimentally observed coherence with actively attending subjects. The coherence estimates for the experiment and the empty room are estimated using the same cortical dipoles and inverse solutions, with an additivity assumption that any increase in coherence during the task relative to the empty room baseline indexes true physiological phase-locking.

The data used here are a subset of those used in the experiments described in Section 2.5 using similar preprocessing steps. We measure coherence from a seed region of interest (ROI) to each dipole source on the cortical surface, producing coherence maps, as in Fig. 2.1. To select the sources corresponding to the seed ROI waveform, we average all trials, plot the dipole activity using dSPM (Dale et al., 2000) and threshold activation at $p < 0.05$, Bonferroni corrected. We select an ROI in each subject based on the largest contiguous cluster of significant activation around Heschl’s gyrus, constrained to have roughly equivalent numbers of sources. In order to obtain an estimate of the average activation in the ROI that is not corrupted by cancellation of the signal due to point spread across sulcal banks, we pick a reference dipole direction and flip the polarity of the dipoles within this ROI to align with this chosen reference direction. The mean waveform of all the flipped dipoles gives us the average activation in the ROI. This mean waveform is then used to compute the coherence with the waveforms of each vertex in the alpha band (8 – 12 Hz) on the cortical surface for the experimental and empty room conditions, individually for each subject. These coherences are Fisher z-transformed ($\arctanh^{-1}$) to stabilize the variance and make the distributions more normally distributed (Bokil et al., 2007). The task coher-
Figure 2.1: Mean alpha coherence maps (morphed on to an average brain) relative to individually defined left-hemisphere auditory cortex (Heschl’s gyrus, represented by the black box) ROIs. The coherence is the mean calculated over the three frequency bins in the 8 − 12 Hz frequency range. The epochs over which these maps are plotted are 500 ms following stimulus presentation in the pitch trials. (a) Fisher $z$-transformed (arctanh) coherence values of the trials and (b) Coherence values in (a) after subtracting $z$-transformed coherence values obtained from empty room measurements. The same processing was done on the empty room data and the neural data. An equal number of trials is used in both conditions, so that there is an equivalent amount of trial-count bias. The same inverse solution was used to control for artifacts from the point-spread of the inverse. Reproduced from Rajaram et al. (2011).

We note a very large spread of coherence around the seed region (Heschl’s gyrus) in Fig 2.1(b). The range of these coherence values (0.2 − 0.5) is much higher than coherence values typically observed in neuroimaging studies, strongly suggesting that the coherence is due to inverse imaging point spread. We therefore conclude that this approach does not
do a very good job of cancelling the spurious coherence. Additionally, this is an unstable estimate, as the variance of the empty-room coherence estimates can be large and is not accounted for with this analysis. Also, since the noise covariance structure of the empty room and the non-task related neural noise covariance structures are very different, the empty room may not be the most appropriate baseline coherence measure. Finally, such analysis could also raise questions in interpretation, such as in the case when empty room coherence is greater than the task coherence. For these reasons, we decided to adopt the more conservative approach of using metrics that discard all phase synchronization (including true physiological coherence) at 0 and \( \pi \) phases.

### 2.2.3 Measures Accounting for Volume Conduction

The first measure introduced to discount the effects of activation spread was the imaginary coherence (Nolte et al., 2004). This measure simply takes the complex valued coherence, used to calculate the coherence, and projects it onto the imaginary axis, giving us a real-valued measure between \(-1\) and 1:

\[
\text{ImCoh} := \text{Im} \left( \frac{\sum_{i=1}^{n} Z_{1,i} \cdot Z_{2,i}^*}{\sum_{i=1}^{n} |Z_{1,i}| \cdot |Z_{2,i}|} \right) = \text{Im} \left( \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} |X_i|} \right),
\]

(2.4)

where the Im operator returns the imaginary value of a complex-valued number. We may also take the absolute value of this quantity to get a measure of phase locking from 0 to 1. We can see that this measure discounts cross-spectral angles at 0 and \( \pi \) phases, since the imaginary components at those phases are zero. However, the size of this metric varies with the phase at which synchronization occurs; positive relative phases have positive imaginary coherence values, and negative relative phases return negative values. Even in the noiseless condition, two signals perfectly phase locked with a phase difference different from \( \pm \pi/2 \) do not return imaginary coherence values of \( \pm 1 \), as one would want with a measure of phase-locking strength.

To overcome these issues, an unweighted (PLV-like) measure called the phase lag index
(PLI) was introduced (Stam et al., 2007) and defined as follows:

$$\text{PLI} := \left| \frac{1}{n} \sum_{i=1}^{n} \text{sign}(\theta_i) \right| = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\text{Im}(X_i)}{|\text{Im}(X_i)|} \right|.$$  \hspace{1cm} (2.5)

This measure is 1 for phase-locking at all non-zero phase lags in the noiseless case and was shown to be more sensitive in detecting true changes in phase-synchronization and more robust to detection of spurious synchronization due to volume conduction than imaginary coherence.

The weighted PLI (WPLI) measure was recently introduced (Vinck et al., 2011) as a metric with less sensitivity to additional, uncorrelated noise sources (that is, uncorrelated with the two sources being analyzed and uncorrelated with each other). It is also more robust to high noise levels for the detection of true phase locking. The WPLI weights each trial in the PLI proportionally to the magnitude of the imaginary component of the cross-spectral component of the trial. This is exactly analogous to the weighting of trials in coherence computation in (2.2), except restricted to the imaginary components. The WPLI is defined as:

$$\text{WPLI} := \left| \frac{\sum_{i=1}^{n} \text{Im}(X_i)}{\sum_{i=1}^{n} |\text{Im}(X_i)|} \right| = \sum_{i=1}^{n} v_i \cdot \text{sign}(\theta_i), \quad \text{where } v_i = \frac{n \cdot |\text{Im}(X_i)|}{\sum_{i=1}^{n} |\text{Im}(X_i)|}. \hspace{1cm} (2.6)$$

Note how this quantity differs from the imaginary coherence in (2.4), which performs the averaging and weighting with the complex cross-spectral terms and then computes the imaginary projection; the WPLI averages and weights the imaginary projections of each trial.

In the same work (Vinck et al., 2011), unbiased (or de-biased in the case of WPLI) estimators for the PLI-squared and the WPLI were also introduced, with a similar approach to the one taken in the design of the PPC measure in (2.3). The unbiased PLI-squared estimator is derived by noting that for two distinct trials $i$ and $j$, the quantity $\text{sign}(\text{Im}(X_i)) \cdot \text{sign}(\text{Im}(X_j))$ is an unbiased estimator of the PLI-squared. Averaging over all pairs of trials...
reduces the variance, giving us the definition:

\[
uPLI := \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \text{sign}(\text{Im}(X_i)) \cdot \text{sign}(\text{Im}(X_j)) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \frac{\text{Im}(X_i) \cdot \text{Im}(X_j)}{\text{Im}(X_i) \cdot \text{Im}(X_j)}.
\] (2.7)

The only guarantee made by this measure is that in expectation, it equals the population estimate of the squared PLI. Thus, these estimates can take on negative values.

A similar approach for the WPLI-squared comes by weighting the quantities in the PLI-squared definition:

\[
\text{debiased WPLI} := \frac{\sum_{1 \leq i < j \leq n} \text{Im}(X_i) \cdot \text{Im}(X_j)}{\sum_{1 \leq i < j \leq n} |\text{Im}(X_i) \cdot \text{Im}(X_j)|}.
\] (2.8)

This estimator is only asymptotically consistent, i.e., there is still a bias that depends on the number of trials (albeit reduced compared to that of the WPLI estimator), but this is shown in simulations (Vinck et al., 2011) to get negligibly small even for a relatively small number (< 30) trials. Again, this estimator can take on negative values despite being a consistent estimator of a squared quantity.

### 2.3 Evaluating Connectivity Measures with Simulations

In this section we compare the various phase-based synchrony measures, and compare their sensitivity in detecting synchrony in the presence of varying levels of volume conducting artifact. We show that the debiased-WPLI\(^2\) is the most sensitive measure of synchrony, while being very robust to point spread synchrony artifacts.

To illustrate the effects of inverse imaging point spread causing the spurious detection of functional connectivity, and the ability of imaginary-component based measures to be robust to such effects, we simulated two sources with varying strengths of point spread. We ran 100,000 Monte carlo simulations of 80 trials each (roughly the median number of trials we had across subjects and conditions in the data in Section 2.5), We simulated two sources with phases \(\theta_{1,i}\) and \(\theta_{2,i}\) respectively, in trial \(i\), drawn i.i.d. from a uniform distribution of
phases in the range $[-\pi, \pi]$. We assume that we observe the true phase of the first source, but that the second source includes energy due to inverse imaging point spread artifacts. Adding the artifact to only one signal simulates point spread without loss of generality, as we are systematically varying the correlation between the two signals; adding point spread artifact to the first signal would only modify the value of the true correlation between the two signals, not the shape of the resulting curve. The observed phase of the second source in trial $i$ is given by $\hat{\theta}_{2,i} = \theta_{2,i} + \alpha \cdot \theta_{1,i}$, where alpha indexes the strength of the point spread between the two sources.

![Effects of Volume Conduction on Connectivity Measures](image)

Figure 2.2: Illustration of the effects of inverse imaging point spread in causing spurious detection of functional connectivity, and robustness of imaginary-component based measures. Each simulation consisted of two sources, both with uniformly distributed true phases. We observe the second source with volume conduction artifacts indexed by $\alpha$, which is plotted on the x-axis in dB, measuring power of the volume conduction component relative to the power true phase component of the second source. The squared functional connectivity measures are plotted on the y-axis. Solid lines represent the mean of the measures across 100,000 Monte carlo simulations of 80 trials each and dashed lines represent $\pm 1$ standard deviation. Standard deviations are not plotted for the imaginary-component based measures for clarity.

Fig 2.2 plots the squared connectivity measures (this allows for a direct comparison since PPC, debiased WPLI-squared and unbiased PLI-squared are estimates of squared measures) as functions of $\alpha$ between the two observed signals $\theta_{1,i}$ and $\hat{\theta}_{2,j}$. We note that the imaginary-component based measures are roughly constant as functions of $\alpha$, while the
PLV, coherence and PPC tend to 1, as $\alpha$ gets larger. Previous studies (Stam et al., 2007; Vinck et al., 2011) note that imaginary-component estimates should theoretically decrease as point spread increases, however we find this decrease to be negligible relative to the standard deviation of the estimates across runs, given the levels of noise and point spread typically observed.

We know that the value of imaginary coherence for phase locking at all relative phases different than $\pi$ and $-\pi$ is strictly less than 1 even in the noiseless case (Vinck et al., 2011). We therefore exclude it from further consideration. But it is also true that the other imaginary-component based measures vary in their sensitivity as a function of relative phase. We investigated this sensitivity by performing a series of simulations, analyzing sensitivity as a function of signal-to-noise ratio at a range of relative phases.

In these simulations, the true phase of the first source in trial $i$, denoted $\theta_{1,i}$, is fixed to 0 without loss of generality. The observed phase of the first source, $\hat{\theta}_{1,i}$ in trial $i$, includes an additive noise component with SNR $\alpha$ and random phase drawn from a uniform i.i.d. distribution. The true phase of the second source, which we denote as $\theta_{2,j}$ in trial $j$, is defined as $\theta_{2,j} = \theta_{1,j} + \nu_j$, where the relative phase $\nu_j$ is drawn i.i.d. from a circular von Mises distribution with mean $\mu$ and concentration parameter $\kappa$. In our simulation in Fig 2.3, we set $\kappa = 5$; each subplot displays the sensitivity curves for different values of $\mu$. The observed phase of the second source $\hat{\theta}_{2,j}$ in trial $j$ is obtained by including an additive noise component with SNR $\alpha$ and random phase drawn from a uniform i.i.d. distribution. The results of these simulations (100,000 Monte carlo runs with 80 trials each) for relative phases $\pi/2, \pi/4, \pi/6$, and $\pi/12$ are plotted in Fig. 2.3.
Figure 2.3: Results of simulations measuring sensitivity of the various functional connectivity measures. One source is simulated as a signal with constant phase that is observed with additive noise at uniformly random phases. The phase of the signal in the second source is von Mises distributed with mean equal to the phase of the first source’s noise-free (constant) signal. This second source is also observed with independent additive noise at the same levels as the first source. SNR levels are varied and plotted on the x-axes. The concentration parameter for the von Mises distribution was fixed at \( \kappa = 5 \), and the means were set at (I) \( \pi/2 \), (II) \( \pi/4 \), (III) \( \pi/6 \), and (IV) \( \pi/12 \). In each subplot, (a) plots the average value of the measure across all Monte Carlo runs and (b) displays the z-score (mean divided by standard deviation) of each measure, as functions of SNR.

We see from the sensitivity analysis in Fig 2.3 that the sensitivities of the unbiased estimators are identical (although consistently slightly less sensitive) to those of their biased counterparts, but the biases do get large for the unbiased estimators at noise levels below 0
The two WPLI measures have very favorable sensitivity to noise as compared to all the other measures at relative phases above $\pi/4$. Even at a relative phase of $\pi/6$, they perform favorably compared to other imaginary-component based measures, but are less sensitive than the coherence and PLV. At the smallest relative phase, the two WPLI measures are roughly the same as the two PLI measures and significantly worse than coherence and PLV. Thus, we prefer the debiased WPLI-squared measure for its ability to reject connectivity induced by point spread, its good statistical power, and its negligible bias for small to medium-sized pools of trials.

### 2.4 ERP subtraction

Another important issue to consider in measuring functional connectivity involves the subtraction of the evoked response prior to analyzing connectivity. This is analogous to subtracting the means from variables prior to calculating their covariance. If we assume signals to have fixed event related potentials (ERPs) in response to a stimulus, then two sources with separate ERPs occurring at different phases will appear as being synchronized, whether or not there is any communication between them. The statistics of the relative phase, in this scenario, are driven by the difference in ERP shapes, rather than by trial-by-trial consistency. Thus, assuming this model, it is recommended that one estimates the ERP and subtracts this quantity from each trial prior to estimating synchrony between regions.

But this approach has been criticized (Truccolo et al., 2002; Wang et al., 2008) as it does not account for trial-by-trial variability in the amplitude and latency (which affects the phase) of ERPs. This approach may lead to apparent phase-locking amongst regions where there is none. However, we claim that if we are using an imaginary-component based measure, this problem is also mitigated. To illustrate this, we model two sources with phases $\theta_{1,i}$ and $\theta_{2,i}$ in trial $i$ that are drawn from independent von Mises distributions with distinct means. The relative phase distribution of these sources will also be von
Mises distributed with mean equal to the difference of the means; as seen in the previous sections, these sources may be detected as connected. Now, consider what happens when we subtract the ERPs prior to looking at the synchrony, noting that the phase distributions of $\theta_{1,i}$ and $\theta_{2,i}$ will be von Mises distributed with zero-mean. Given this, the difference in the phases of the two sources will also have zero mean. As discussed in the previous section, the imaginary component-based measures will then have an expected value of 0 (or some bias term corresponding to the number of trials, for the biased measures). This is, therefore, another reason to use the debiased WPLI-squared metric, along with ERP subtracted trials.

2.5 Space vs. Pitch Experiment

We analyze the data from a study conducted to assess neural activation in selective attention tasks, comparing attention to spatial and non-spatial auditory features (Lee et al., 2013). In this section we analyze the functional connectivity between regions while subjects are performing the task and look at which pairs of regions show increased communication when attending to spatial, relative to non-spatial, auditory features. In the activation analysis, it was found that neural activation was greater in the left frontal eye fields (FEF) in the pre-stimulus preparatory period, when subjects were preparing to attend to a spatial location, rather than the pitch of a speaker. Also, the left superior temporal gyrus (STG), the region containing the primary auditory cortex, showed greater activation in trials where subjects were attending the pitch, rather than the spatial location of a speaker. The same pre-processing and inverse imaging steps were used for the current analysis, and are described below. We refer the reader to the study (Lee et al., 2013) for further details.
2.5.1 Methods

2.5.1.1 Task and Stimuli

Sixteen normal-hearing (defined as having thresholds with 20 dB of normal-hearing thresholds in each ear) listeners participated, giving informed consent approved by Massachusetts General Hospital and Boston University. The ages of the participants ranged from 18 to 35 years. Only 2 subjects were female; we do not expect any significant effects of sex in this study. Diffuse white noise (where each ear receives the same signal but with opposite polarity, providing inconsistent interaural spatial cues and thus a ‘diffuse’ spatial percept) was presented at 60 dB SPL throughout the experiment to mask environmental sounds. The signal-to-noise ratio (SNR) of the speech tokens was 20 dB, a comfortably audible level for all listeners.

In each trial, two distinct auditory tokens with different randomly assigned spatial locations and pitches were presented simultaneously. The speech tokens used in this experiment were monosyllabic spoken digits (‘one’ through ‘four’) taken from the TIDIGIT corpus (Leonard, 1984), each cropped to be approximately 400 ms in duration. The pitch of each token was monotonized and set to 3 semitones above or below 100 Hz (the nominal pitch of the original speech) using PRAAT software (Boersma & Weenink, 2009). Each token was then convolved with a head-related transfer function (HRTF) sourced from Shinn-Cunningham et al. (2005) to simulate a source either 30° left or 30° right of midline.

Each subject performed four runs of roughly 6 minutes duration. Subjects were instructed to maintain visual fixation at a dot in the center of the visual display. Each run consisted of 72 trials, giving us a total of 288 trials per subject, evenly divided across conditions. In each trial, subjects were visually cued to attend to either one of the spatial locations (left/right) or one of the pitches (low/high) in a randomly intermingled and counter-balanced order. The visual cue was presented for 300 ms, followed by the fixation dot during a 700 ms preparatory period prior to stimulus onset. Subject responses were recorded by a 4-button response box, controlled with the right hand. To minimize motor
artifact, subjects were told to respond only at the end of the trial, when cued by a visual ring (appearing one second after the onset of the auditory mixture).

2.5.1.2 Data Acquisition

MEG data (306 channels - two planar gradiometers and one magnetometer at each of 102 sites) were recorded at a 600 Hz sampling rate (bandpass 0.1 – 200 Hz) simultaneously with two bipolar electro-oculogram (EOG) electrode pairs measuring horizontal eye movements and blinks. The data were recorded inside a magnetically shielded room using a dc-SQUID Neuromag VectorView system (Elekta-Neuromag). Four head-position indicator coils were placed on the head of the subject to monitor the position and orientation of the head relative to the MEG sensor array. Four head position indicator (HPI) coils were used to monitor head position relative to the sensors at the beginning of each run (Liu et al., 2010). MRI scans were performed with a 1.5 T Avanto scanner (Siemens Medical Solutions, Erlangen, Germany).

2.5.1.3 Data Preprocessing

All raw MEG data were first processed using the signal-space separation (SSS) method (Taulu et al., 2005) to remove biological artifacts and other environmental magnetic sources originating outside the head. In addition, we eliminated the subspace containing the heart-beat artifact using the signal-space projection (SSP) method. Only those trials with correct behavioral responses were included in further processing. Trials were also rejected if there were eye movements or blinks (if the bipolar EOG electrode pairs showed a voltage difference greater than 200 µV), or if the peak-to-peak MEG amplitude exceeded 10 pT in magnetometers or 3 pT/cm in gradiometers.

2.5.1.4 Source Imaging

Individualized three-layer boundary element models (BEMs) were constructed for each subject using T1-weighted and two FLASH (flip angle of 5 and 30 degrees) MRI scans.
The cortical, inner and outer skull surfaces were reconstructed using the Freesurfer software suite (http://surfer.nmr.mgh.harvard.edu/). Source estimation was performed using cortical-surface-constrained (Fischl et al., 1999), l2-norm-based minimum norm estimation using the MNE software suite (http://www.nmr.mgh.harvard.edu/mne). We averaged the forward solutions from each run (incorporating the head positions) before calculating the inverse operator, allowing the use of a single inverse projection for each subject. A grid spacing of 7 mm was used for dipole placement, yielding roughly 3000 sources per hemisphere. Dipole orientations were constrained to be normal to the cortical surface. The noise covariance matrix was estimated from the 200 ms baseline periods prior to the onset of each trial (before the visual cue). Anatomical parcellation yielding boundaries for vertex selection within ROIs was performed by Freesurfer (Fischl et al., 2004).

2.5.2 Functional Connectivity Analysis

To assess the connectivity between a pair of ROIs, we extracted epochs (either 500 ms preceding or following the auditory stimulus onset) corresponding to each trial, for every dipole source that lies within a specified anatomical label. The ERP of each source was subtracted from each epoch at that source, as discussed in Section 2.4. All the sources within a given ROI were sorted in order of ERP power and the 10 sources with maximum ERP power were selected to represent the label activity. Debiased WPLI-squared measures were computed for all pairs of sources between two regions using multi-taper spectral estimation; for alpha connectivity, we used one frequency bin at 10 Hz, with one Slepian taper; for beta connectivity, we used 6 frequency bins in the range 16 – 26 Hz with three Slepian tapers. These bins were selected so as to ensure minimal bias in the alpha (8 – 12 Hz) and beta (13 – 30 Hz) bands, respectively. These estimates were then averaged to yield an average label-to-label connectivity measure for each subject. After computing these estimates for each subject, we performed a one-sided paired t-test (synchrony in Space trials > synchrony in Pitch trials) across subjects, and obtained t-statistics for each pair of labels we considered.
The resulting t-statistic weighted graph was thresholded at the t-value corresponding to a p-value of 0.1. This value was chosen as a threshold so as to reject noisy or unreliable connections. We do not claim that every edge in the resulting graph represents statistically significant connectivity, as we have not accounted for multiple comparisons. Now, if we treat each label as a vertex and significant t-values as weighted edges, we have a weighted graph representing the functional network (van Wijk et al., 2010; Breakspear et al., 2010) underlying spatial attention (with attention to pitch as the baseline measure). In order to test the hypothesis that a cortical region serves a central role or a ‘hub’ in the network, we can compute the eigenvector centrality, which measures the influence of a node in the network (Kolaczyk, 2009). It is defined recursively by attaching a higher weight to connections with larger centrality. Specifically, for a given vertex \( v \), the eigenvector centrality of \( c_v \) in the graph is defined as:

\[
c_v = \frac{1}{\lambda} \sum_{u \in N(v)} c_u,
\]

where \( \lambda \) is a constant (actually the largest eigenvalue of the adjacency matrix) and \( N(v) \) denotes the neighborhood of \( v \) or the set of all vertices that are adjacent to \( v \). As suggested by the name, this measure is computed simply by looking at the eigenvector of the largest eigenvalue of the adjacency matrix of the graph. By the Perron–Frobenius theorem, all the entries of this eigenvector are non-negative and the vector is scaled to have unit norm.

In order to perform a hypothesis test on the significance of the centrality measure, we permute the weighted edges randomly on the graph using the Erdős-Rényi model of random graphs (Kolaczyk, 2009), which involves uniformly random permutations. By performing repeated simulations we are able to generate an empirical distribution of the eigenvector centrality at a given vertex (actually, any vertex) and reject the hypothesis with a given \( p \) value if the proportion of simulated measures that were smaller than than the observed measure was less than \( 1 - p \).
2.5.3 Selection of ROIs and Frequency Bands for Functional Connectivity

We analyzed the alpha (8 – 12 Hz) and beta (13 – 30 Hz) frequency bands because of the numerous reports of the importance of their roles in attention and other cognitive processes. One of the proposed mechanisms through which alpha oscillations aid selective attention is by suppressing sensory representations; increase in alpha power often corresponds to better suppression of irrelevant stimuli (Kelly et al., 2006; Snyder & Foxe, 2010; Foxe & Snyder, 2011; Weisz et al., 2013). A mechanistic explanation of this phenomenon is that top-down modulation causes a phase reset of the alpha oscillations, leading to an apparent increase in alpha power in the evoked response (Klimesch et al., 2007). Indeed, several reports indicate that the strength of alpha synchronization indexes performance in various cognitive tasks (Sauseng et al., 2005; Klimesch et al., 2007; Palva & Palva, 2007; Palva et al., 2010). It must be noted here that the vast majority of the evidence about the role of alpha oscillations and synchrony come from studies in the visual domain, but there is evidence for a similar role for alpha in the auditory domain as well (Kerlin et al., 2010; Weisz et al., 2011; Thorpe et al., 2011; Obleser & Weisz, 2012; Obleser et al., 2012a).

Similarly, modulations of beta oscillations and cortical synchrony have been implicated in several visual attention tasks (Tallon-Baudry et al., 1999; Siegel et al., 2008; Gregoriou et al., 2012) as well as auditory attention tasks (Shahin et al., 2009). Beta band activity is also increasingly implicated in studies of higher order linguistic functions, such as semantic memory, syntactic binding and discrimination of word categories (Weiss & Mueller, 2012). While other bands, particularly theta (4 – 8 Hz) and gamma (> 30 Hz) have also been implicated in attentional processing (Lakatos et al., 2008; Siegel et al., 2008; Gregoriou et al., 2009, 2012; Lakatos et al., 2013), we decided not to analyze them for this experiment; due to the short duration of our stimuli and pre-stimulus epochs, we are able to observe very few cycles of theta oscillations within an epoch, leading to poor spectral estimates. The poor SNR due to the 1/f spectral distribution of EEG (Nunez & Srinivasan, 2005) also led to poor spectral estimates in the gamma band, prompting our decision to leave it
out of the analysis.

In order to construct our functional connectivity networks, we first assemble a list of candidate regions of interest based on studies examining neural activation through invasive electrophysiology, fMRI and M/EEG. Regions were defined anatomically using a parcelation based on individual subjects’ sulcal-gyral cortical folding patterns (Fischl et al., 2004). Previous studies have performed such analyses using all possible regions from the parcellation (Palva et al., 2010), but we choose to perform a manual pruning as a first step, so as to allow easier visualization and interpretation of the inferred networks. We list the regions (both hemispheres of each label were used in the analysis) and list some known results of their involvement in attentional processing of sensory stimuli.

2.5.3.1 Primary Auditory Cortex

We defined the primary auditory cortex to be the region corresponding to the superior temporal gyrus, Heschl’s gyrus and planum temporale. This is a natural candidate ROI while studying auditory attention networks, due to its central role in auditory processing in the cortex. A recent pair of studies (Weisz et al., 2013; Müller & Weisz, 2012) have shown that alpha power increases in auditory cortex in the pre-stimulus period during an auditory selective spatial attention tasks, and this increase is right hemispheric dominant. Additionally, they also showed that the right auditory cortex also shows increased functional connectivity with right IPS (in the alpha band) and right FEF (in the theta band, which we do not analyze here).

2.5.3.2 Superior Temporal Sulcus (STS)

The STS is a region that is associated with general auditory processing, for both spatial and non-spatial features (Ahveninen et al., 2006; Hill & Miller, 2010). In the evoked response analysis of our data (Lee et al., 2013), it was found that left STS activity was greater when listeners direct attention to target pitch rather than target location, and only in the pre-stimulus period.
2.5.3.3 Inferior Pre-Central Sulcus (infPCS)

The inferior pre-central sulcus contains the inferior frontal eye fields (iFEFs) and the inferior frontal junction (IFJ). The iFEF is known to be activated during saccade tasks (Derrfuss et al., 2012), including the Go-No Go functional localizer for FEF (Lee et al., 2013) in our dataset. The IFJ is known to be involved in task switching (Derrfuss et al., 2009) and is proposed as the site where the dorsal (top-down) and ventral (bottom-up) attention networks converge (Asplund et al., 2010). Additionally, a combined fMRI-EEG study showed that IFJ is involved in visual attention to color and motion; moreover, across subjects, the functional connectivity with sensory cortex (V4) in the alpha band predicted the degree of attentional modulation (Zanto et al., 2010). Finally, strong fMRI activation is found over large portions of inferior pre-central sulcus bilaterally in an audiospatial working memory task (Tark & Curtis, 2009), suggesting it plays a role in auditory spatial attention. Given this evidence, we hypothesize that IFJ is involved in a supramodal attention network, and that similar strong alpha connectivity is predicted with auditory processing areas as well.

2.5.3.4 Intra-Parietal Sulcus (IPS)

The IPS is known to be involved in visuospatial attention (Buschman & Miller, 2007; Siegel et al., 2008; Asplund et al., 2010) and to have maps of visual space; recent studies have shown that it is also involved in auditory spatial attention (Salmi et al., 2009; Cohen et al., 2005), but uses a different coding scheme, without spatial maps (Kong et al., 2012).

2.5.3.5 Precuneus

The precuneus is located in the medial wall of the posterior parietal cortex. It is thought to be a supra-modal attention region and is implicated in shifting attention between objects. This has been demonstrated with fMRI in vision (Nagahama et al., 1999; Le et al., 1998) and in audition (Pugh et al., 1996). Another study found (Shomstein & Yantis, 2006)
that only the right precuneus showed a difference when comparing spatial and non-spatial auditory attention tasks. The precuneus also shows a contralateral gain modulation in its activity for auditory spatial attention shifts, but not for visuospatial attention shifts (Krumbholz et al., 2009). Additionally, the precuneus shows bilateral activation while preparing to attend, prior to stimulus onset for an auditory spatial attention task (Mayer et al., 2006; Wu et al., 2007). However, it must be noted that all these studies are conducted with fMRI, making it challenging to predict how MEG oscillatory activity will be affected.

2.5.3.6 Frontal Eye Fields (FEF)

In humans, the area traditionally known as the frontal eye fields, or FEFs (as opposed to the inferior frontal eye fields, which we discussed above in Section 2.5.3.3) are contained in the superior precentral sulci (Rosano et al., 2002), which were the anatomical parcellations we used to define this region. Traditionally, the FEFs are thought to control saccades and direct visuospatial attention (Corbetta et al., 2008; Gregoriou et al., 2009), but recent studies show that auditory spatial attention also elicits FEF activation (Mayer et al., 2006; Salmi et al., 2009), even in extra-retinal space (Tark & Curtis, 2009).

We also know from invasive studies in monkeys that FEF beta band spike-field coherence increases when the animal is performing a visuospatial attention task (Gregoriou et al., 2012). There are also reports visuospatial attention results in causal broadband control signals (that include the beta, but exclude the alpha band) directed from FEF to the visual region V4 (Gregoriou et al., 2009), as well as increased beta band LFP coherence between FEF and the intraparietal sulcus when the animal is attending to a spatial location (Buschman & Miller, 2007). Other human MEG studies (Siegel et al., 2008) reveal significant beta activation in FEF in the pre-stimulus and post-stimulus periods. Hemispheric differences in the roles of FEF have also been reported, with the left FEF being implicated in top-down attention and right FEF being implicated in both top-down and bottom-up attention (Corbetta et al., 2008). However, these differences have been largely ignored in a majority of the literature and are thus poorly understood, especially with regards to
Due to its strong recruitment in visuospatial, as well as audiospatial attention, along with numerous studies showing beta band functional connectivity with sensory areas, we hypothesize that FEF plays a central role in directing spatial attention. Specifically, given the results from animal studies, we expect strong connectivity to emerge in the beta band.

2.5.4 Results

We construct and plot functional connectivity network graphs for the pre- and post-stimulus onset epochs for alpha and beta bands in Figs. 2.4 - 2.7. In the post-stimulus onset alpha band network, shown in Fig. 2.4, we found that left inferior PCS showed significant eigenvector centrality of 0.621 ($p < 0.05$) with reliable synchrony with contralateral STS ($p < 0.01$, mean debiased WPLI$^2$ in Space trials: $3.13 \cdot 10^{-3} \pm 7.64 \cdot 10^{-3}$ std. dev., mean debiased WPLI$^2$ in Pitch trials: $-1.72 \cdot 10^{-3} \pm 6.63 \cdot 10^{-3}$ std. dev. across subjects) and ipsilateral IPS ($p < 0.01$, mean debiased WPLI$^2$ in Space trials: $4.61 \cdot 10^{-3} \pm 7.02 \cdot 10^{-3}$ std. dev., mean debiased WPLI$^2$ in Pitch trials: $-1.53 \cdot 10^{-3} \pm 6.60 \cdot 10^{-3}$ std. dev. across subjects). It must be noted that these $p$ values for the individual connections are not corrected for multiple comparisons. The right inferior PCS did not show a similarly significant connectivity patterns for spatial versus non-spatial attention.

The spatial attention network in the pre-stimulus epochs in the beta band show that the right FEF has significant eigenvector centrality of 0.61 ($p < 0.05$) for spatial attention with strong connectivity to ipsilateral STS ($p < 0.01$, mean debiased WPLI$^2$ in Space trials: $1.25 \cdot 10^{-3} \pm 2.69 \cdot 10^{-3}$ std. dev., mean debiased WPLI$^2$ in Pitch trials: $-1.04 \cdot 10^{-3} \pm 1.82 \cdot 10^{-3}$ std. dev. across subjects) and auditory cortex ($p < 0.05$, mean debiased WPLI$^2$ in Space trials: $0.51 \cdot 10^{-3} \pm 2.34 \cdot 10^{-3}$ std. dev., mean debiased WPLI$^2$ in Pitch trials: $-1.60 \cdot 10^{-3} \pm 2.51 \cdot 10^{-3}$ std. dev. across subjects). Note that despite having only two strong connections, both STS and auditory cortex have eigenvector centralities (0.485 and 0.476, respectively), boosting the centrality of right FEF.

The post-stimulus onset beta band network, shown in Fig. 2.5, and pre-stimulus alpha
band network, shown in Fig. 2.6, did not find reliable synchrony between ROIs and were not analyzed further.
Figure 2.4: Functional network underlying auditory spatial attention, relative to auditory non-spatial attention, in the alpha band (8 – 12 Hz) in the 0.5 s epoch following stimulus onset. Debiased WPLI-squared measures were computed for all pairs of sources, for each subject, using one frequency bin at 10 Hz, with one Slepian taper, so as to ensure minimal bias in the alpha band. These estimates were then averaged to yield an average label-to-label connectivity measure for each subject, and a one-tailed paired t-test (Space > Pitch) was performed. Each edge (corresponding to a pair of ROIs) was weighted by its t-statistic and edges were thresholded at the t-value corresponding to a p-value of 0.1 and plotted. Locations of ROIs are pictured under the labels. All left-hemisphere labels are to the left of the vertical midline, while all right-hemisphere labels are to the right. The same regions share the same color code in the two hemispheres and are located in polar opposite locations.
Figure 2.5: Functional network underlying auditory spatial attention, relative to auditory non-spatial attention, in the beta band (13 – 30 Hz) in the 0.5 s epoch following stimulus onset. Debiased WPLI-squared measures were computed for all pairs of vertices, for each subject, using all frequency bins in the range 16 – 26 Hz, with three Slepian tapers, so as to ensure minimal bias in the beta band. The layout of the figure is identical to that of Fig. 2.4.
Figure 2.6: Functional network underlying auditory spatial attention, relative to auditory non-spatial attention, in the alpha band (8 – 12 Hz) in the 0.5 s epoch preceding stimulus onset. The layout of the figure is identical to that of Fig. 2.4.
34

Figure 2.7: Functional network underlying auditory spatial attention, relative to auditory non-spatial attention, in the beta band (13–30 Hz) in the 0.5 s epoch preceding stimulus onset. The layout of the figure is identical to that of Fig. 2.4.

2.5.5 Discussion

We tested both FEFs as hubs in both beta band networks, as they are known to show a high degree of connectivity in that band from previous studies (Buschman & Miller, 2007; Siegel et al., 2008; Gregoriou et al., 2009, 2012). We found right FEF centrality to be significant ($p < 0.05$) only in the pre-stimulus beta band spatial network. We may interpret a high centrality score as an ‘importance’ score and the score of a node
increases if it is connected to other nodes with high scores. We may alternatively think of it as representing the frequency with which a node is visited during a random walk on the underlying graph. The previous studies found beta band synchrony between FEF and higher-order visual areas, such as V4 (primary visual cortex was not recorded). As a parallel, in this experiment the FEF shows increased synchrony with the higher-order auditory region, STS.

This is consistent with recent human transcranial magnetic stimulation (TMS) study (Chanes et al., 2013) that shows that a sequence of rhythmically timed pulses delivered to the right FEF in the beta frequency (assumed to elicit neural activity in the beta range) prior to viewing a stimulus enhances perceptual sensitivity in a visual detection task. Non-rhythmic pulse trains had no effect on the sensitivity. A preliminary follow-up to the study (Valero-Cabrè et al., 2013) reports that this selective enhancement for beta band rhythmic TMS activation was not present in left FEF. Additionally, these studies suggest beta band synchrony between FEF and visual cortex are of importance in visual perception; this mirrors with our findings of FEF and auditory cortex connectivity. These results strongly suggest that FEF plays an important role in controlling cross-modal spatial attention and does so in very similar ways in the auditory and visual modalities.

In addition to the involvement of the FEF, we hypothesized that the IFJ in the inferior PCS would show connectivity with regions that performed higher order auditory processing, based on similar synchrony with extrastriate cortex during visual tasks (Zanto et al., 2010). Indeed inferior PCS showed significant ($p < 0.05$) centrality within the post-stimulus alpha audiospatial specific functional network. By contrast the label in the right hemisphere showed no such centrality; this strong hemispheric asymmetry has not been reported in activation studies and raises interesting possible questions for future work.

It is important to note here that lack of significance of an edge in the networks we have constructed does not imply a lack of importance of the region, for audiospatial attention tasks. Rather we are evaluating the relative importance of the region for spatial versus non-spatial attention tasks.
2.6 Conclusion

Oscillatory synchrony is an important mechanism for communication between cortical regions. This communication creates dynamic neural assemblies that work in concert to jointly process an incoming stimulus or top-down command. While invasive measurements provide far richer and cleaner estimates than MEG and EEG, this approach can only measure from a few sources at a time. MEG and EEG allow us to sample from the entire brain at once, but pose myriad challenges in the interpretation of the data. The effects of volume conduction on M/EEG data cause significant point spread between estimates of neural activity. Here, we compared various techniques that account for this problem. In Sec. 2.3, we conducted a simulation study and found that the weighted phase locking index (Eq. 2.6) was least affected by point spread and was the most sensitive to detecting synchrony through noise, although this sensitivity is dependent on the phase of synchrony. We also noted that the unbiased estimators, that account for bias due to finite trial count, are very similar to their biased counterparts in terms of sensitivity to noise. We also noted the need to subtract event-related potentials prior to computing functional connectivity and how imaginary-component based measures are robust to the problems raised by trial-by-trial variability in ERPs.

However, one must also be aware of potential problems with these measures. While WPLI values can go as high as 1 in the noiseless case, for all relative phases different from 0 and $\pi$, with the addition of noise, these curves asymptote at different values for different relative phases. Also, at lower SNRs, connectivity in one condition may seem stronger due to a favorable relative phase versus another condition with the same underlying level of synchrony, at a different relative phase.

With this in mind, we used the debiased WPLI$^2$ measure (Eq. 2.8) to analyze auditory spatial functional connectivity networks in the alpha and beta bands, for when a subject is preparing to attend, as well as when the subject is attending to the sound.

We performed a graph theoretic analysis of centrality and found that the right frontal
eye fields serve as a hub for the beta-band network that arises when a listener is preparing to attend to a spatial location. Interestingly, in the evoked response analysis of the same data (Lee et al., 2013), we saw an increase only in the left FEF activity for spatial trials, relative to the pitch trials, both for the pre-stimulus and post-stimulus epochs. Significant signal amplitude differences between “attend location” and “attend pitch” trials were not seen in the right FEF in this analysis. Thus, we note that signal amplitude (evoked responses) and functional connectivity measures can provide independent, and sometimes even conflicting perspectives on functional networks, even in the same dataset. Taking the evoked response and functional connectivity results of this experiment together, we propose that FEF controls spatial attention (although it known to be involved in non-spatial attention control as well) across sensory modalities. The beta band is clearly implicated in processing within the region (Gregoriou et al., 2012), and for communication with sensory cortices, particularly in the right hemisphere.

We also found that the left inferior post-central sulcus serves as a hub for the alpha-band network that arises when a listener is actively attending a spatial location. Interpretation of this result is made harder as there are two functional subregions that are important for spatial attention within this anatomical region: the inferior frontal junction and the inferior eye field (Derrfuss et al., 2012). Distinguishing between the activity of these two regions with MEG inverse imaging estimates is not possible due to the poor spatial resolution inherent to the measurements. However, previous reports of IFJ alpha band synchrony with visual extrastriate cortex, combined with the fact that we see strong alpha band synchrony with auditory regions, suggests to us that the underlying neural substrate is likely the IFJ. Little is known about the role that IFJ plays in the spatial attention network, however our results strongly suggest an important role for this region in the deployment of spatial attention.
Chapter 3

A Dynamical System Model for Tracking Ongoing Speech

3.1 Introduction

The ability of human listeners to selectively attend to a single speaker in noisy environments, or the ‘cocktail party problem’ (Cherry, 1953) is critically important for speech comprehension and social interactions. However the neural mechanisms underlying attentive processing of ongoing speech are not well understood. To understand the effects of attention in modulating the neural representation of speech, it is important to identify what acoustic features are extracted from the stimuli arriving at the listeners’ ears.

The majority of human electrophysiological studies of speech processing use short utterances as stimuli (Sanders et al., 2002; Lee et al., 2013) and study the evoked response as an event-related potential (ERP) with EEG, or an event-related field (ERF) with MEG. Such short speech utterances do not contain rhythm, intonation, or prosody cues present in longer segments of natural coarticulated speech. Full sentences also contain syntactic and semantic linguistic cues that can greatly aid in speech comprehension.

There have been several studies conducted with sentence-length stimuli using EEG (Kerlin et al., 2010) and MEG (Ahissar et al., 2001; Luo et al., 2010). However, such studies analyze the neural response as ERPs/ERFs, forming neuronal response patterns by averaging responses to multiple presentations of the same closed set of sentences. So while accounting for prosodic and linguistic cues, such studies do not provide mechanistic
explanations of the underlying neuronal processing. Furthermore, the averaged patterns cannot be generalized to new sentences not in the original closed set, limiting the utility of the inferred patterns.

Recently, several MEG (Ding & Simon, 2012; Zion Golumbic et al., 2013b) and EEG (Lalor & Foxe, 2010; Power et al., 2012) studies have taken the approach of modeling neural responses as linear systems responding to the broadband temporal envelope of speech, where the gain of this response is modulated by selective attention. Similar effects of selectively attending to continuous speech have also been shown using invasive electrocorticography measurements (Mesgarani & Chang, 2012; Zion Golumbic et al., 2013a) by modeling the neuronal responses as spectro-temporal response functions (STRFs), which are linear functions of spectro-temporal acoustic cues.

All these M/EEG speech tracking models fit system response functions in sensor space. However, this is clearly inefficient from a statistical perspective, assuming the number of neural sources contributing to this response (i.e., the degrees of freedom in the spatial domain) is much less than the number of sensors. Based on this insight, we present a modeling framework based on the state-space representation of a linear dynamical system (LDS). Our model is also based on linear time-invariant (LTI) responses to the broadband envelopes of the input speech streams with different impulse responses for the attended and unattended speech. Additionally, the model captures spectral characteristics of the noise in the neural sources (i.e. neural activity unrelated to the speech envelope processing) and neural functional connectivity using a multivariate autoregressive (MVAR) model. Note that without an epoched structure to the data, we may not assume a phase distribution and infer connectivity using the techniques discussed in Ch. 2.

Our model describes the activities of the neural substrates as latent states, with a much lower dimensionality than that of the EEG sensor space, and projects these activities onto the scalp, to be recorded with additive noise by the EEG (or MEG) sensors. Similar state-space approaches have been recently proposed for modeling the inverse imaging problem of inferring cortical sources from electro-magnetic fields measured non-invasively with
M/EEG. In particular, the model in (Lamus et al., 2012) also accounts for connectivity among neural sources using autoregressive (AR) dynamics. Ours additionally includes a model of linear time-invariant responses to external stimuli. We use a regularized expectation maximization (EM) (Dempster et al., 1977) algorithm for parameter identification, extending similar work in (Schneider, 2001).

The parameters of our state-space model correspond to the impulse responses of the system, the AR coefficients (for noise spectra and functional connectivity parameters), the forward model or neural source-EEG projection matrix, as well as the state and measurement noise covariance matrices. Our formulation allows us to very easily add priors or constraints to these parameters in order to make stronger assumptions in our analysis, as we will see in Ch. 4. Another benefit of a state-space model is it allows for efficient online state estimation using a Kalman filter, or an efficient Kalman smoother or Rauch-Tung-Striebel forward-backward algorithm (Gelb, 1974), which is non-causal but can be used in an online setting with a sliding temporal window. An efficient online estimation algorithm is important if we hope to use such a system in a brain-computer interface, or to control an intelligent hearing prosthetic for the hearing-impaired.

In this chapter, we present the model and the parameter estimation procedure. We test the fitted models on new data to predict attended speakers and analyze the prediction performance of the model.

3.2 Methods

3.3 Linear Dynamical System Model

3.3.1 Autoregressive Dynamics

We formulate our model so that the latent state-space $p$-dimensional vector $x(t)$ corresponds to the activities of the underlying $p$ neural sources at time $t$. This latent variable has multivariate autoregressive dynamics with exogenous inputs (MVARX model) (Ljung, 1998). This model assumes that the activity at time $t$ of any neural source (i.e., any com-
ponent of the latent state vector $x(t)$ is a linear function of past states and past inputs (with fixed constant-width time windows), stated as:

$$x(t) = A_1 x(t-1) + \ldots + A_{na} x(t-na) + B_1 u(t-1) + \ldots + B_{nb} u(t-nb) + \epsilon(t),$$

with $\epsilon(t) \sim \mathcal{N}(0, Q)$. (3.1)

The autoregressive parameters $A_i$ are $p$-by-$p$ dimensional matrices and the $m$ exogenous inputs at time $t$ are in the $m$-dimensional exogenous input vector $u(t)$. This exogenous input vector is also referred to as the control signal in control systems engineering (Anderson & Moore, 1979). The impulse response to the inputs is captured in the $p$-by-$m$ matrices $B_j$, also called control matrices. The window width parameters $n_a$ and $n_b$ are positive integers. The additive driving noise $\epsilon(t)$ is assumed to be zero-mean normally distributed with covariance $Q$.

From this formulation, we can see that for the $k$th neural source (or the $k$th component $x_k(t)$ of the latent state vector), the noise spectrum is given by the $k$th columns of $A_1, \ldots, A_{na}$ and all other terms in the matrices $A_i$ measure functional connectivity amongst sources at lag $i$. Similarly, the $k$th columns of $B_1, \ldots, B_{nb}$ measure the impulse response to the $k$th input stream, $u_k(t)$. 
Figure 3.1: Graphical structure of ARX model illustrating the dependency between variables at a single time point \( x(t) \), a \( p \)-dimensional vector corresponding to the latent state at time \( t \). The exogenous inputs to the system are given by the \( m \)-dimensional vector \( u(t) \) and the driving white noise is shown by the \( p \)-dimensional vector \( \epsilon(t) \). Note that if \( n_a > 1 \), this is not a Markov process since \( x(t) \) is not independent of \( x(t - 2) \) given \( x(t - 1) \).

### 3.3.2 State Space Formulation

We can express the MVARX model described above with state space dynamics by employing an augmented state vector (Gelb, 1974) \( \tilde{x}(t) \) that includes \( n_a - 1 \) past states along with the current state \( x(t) \), defined as:

\[
\tilde{x}(t) = \begin{bmatrix} x(t) \\ x(t - 1) \\ \vdots \\ x(t - n_a + 1) \end{bmatrix}, \tag{3.2}
\]

and an augmented state transition matrix \( \tilde{A} \) that includes all the autoregressive matrices \( A_i \) and keeps a running history of past states in the augmented state vector \( \tilde{x} \), as follows:
\[
\tilde{A} = \begin{bmatrix}
A_1 & \cdots & \cdots & A_n \\
I & 0 & 0 & 0 \\
0 & \ddots & 0 & 0 \\
0 & 0 & I & 0 \\
\end{bmatrix}.
\]  

(3.3)

Similarly, we create an augmented input vector \( \tilde{u}(t) \) by stacking past inputs row-wise from appropriate time points:

\[
\tilde{u}(t) = \begin{bmatrix}
u(t) \\
\vdots \\
u(t - n_b + 1)
\end{bmatrix}.
\]

(3.4)

and an augmented impulse response (or control matrix) \( \tilde{B} \) by stacking the matrices \( B_j \) column-wise with zero entries for the past states, as follows:

\[
\tilde{B} = \begin{bmatrix}
B_1 & \cdots & B_{n_b} \\
0 & \cdots & 0 \\
\vdots & \vdots & \vdots 
\end{bmatrix}.
\]

(3.5)

The state noise only affects the current state in the augmented state vector (the first \( p \) components of \( \tilde{x}(t) \)) and thus has a zero-mean and covariance

\[
\tilde{Q} = \begin{bmatrix}
Q & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \cdots & \vdots 
\end{bmatrix}.
\]

(3.6)

It is easy to verify that in the augmented state \( \tilde{x}(t + 1) = \tilde{A}\tilde{x}(t) + \tilde{B}\tilde{u}(t) + \tilde{\epsilon}(t) \) with \( \tilde{\epsilon}(t) \sim \mathcal{N}(0, Q) \) is exactly equivalent to the original equation in (3.1). We also note that in this augmented space, the latent variables define a Markov process since \( \tilde{x}(t) \) is independent of \( \tilde{x}(1), \ldots, \tilde{x}(t - 2) \), given \( \tilde{x}(t - 1) \).

The latent neural states \( x(t) \) are projected onto the EEG (or MEG) montage by the
channels-by-\( p \) lead field matrix \( C \). The observations are assumed to have additive zero-mean Gaussian noise with diagonal covariance matrix \( R \) (the diagonal assumption is only made to reduce the degrees of freedom in our statistical learning procedure and may be easily dropped). If our observed EEG signal at time \( t \) is given by \( y(t) \), our observation model is \( y(t) = Cx(t) + \nu(t) \), where \( \nu(t) \sim \mathcal{N}(0,R) \). In the augmented state space, only the states corresponding to the current time point are projected. This is because we make a quasi-static approximation to Maxwell’s equations due to the relatively slow (< 1 kHz) dynamics of M/EEG signals (Hämäläinen et al., 1993). Additionally, we assume the underlying neural sources are fixed and respond in an ongoing and continuous manner to the incoming speech. This yields the extended measurement matrix:

\[
\tilde{C} = \begin{bmatrix} C & 0 & \cdots & 0 \end{bmatrix}.
\] (3.7)

Thus, our final state space formulation is:

\[
\begin{align*}
\tilde{x}(t+1) & = \tilde{A}\tilde{x}(t) + \tilde{B}\tilde{u}(t) + \tilde{\epsilon}(t), \quad \tilde{\epsilon}(t) \sim \mathcal{N}(0,\tilde{Q}) \\
y(t) & = \tilde{C}\tilde{x}(t) + \nu(t), \quad \nu(t) \sim \mathcal{N}(0,R), R \text{ diagonal.}
\end{align*}
\] (3.8) (3.9)

This is now the standard formulation of a state-space linear dynamical system with an exogenous input (or control signal), since the latent states form a Markov process in the augmented state space. Thus, we are able to perform Kalman filtering and smoothing on this augmented system.

### 3.4 System Identification

Kalman filtering or smoothing allows us to estimate values of the latent states \( \tilde{x}(t) \), given the parameters of the system. Similarly, if we are able to directly observe the latent states, then estimating the parameters involves simple maximum likelihood estimation. However, in this situation, we know neither the model parameters, nor the latent states. To simul-
taneously estimate the two, we need to perform expectation maximization (EM). In this section we describe our implementation of the EM algorithm which computes regularized estimates of parameters in the M-step.

### 3.4.1 Regularized EM Algorithm

We denote the parameters of the system as \( \theta = \{A_1, \ldots, A_n, B_1, \ldots, B_n, Q, C, R\} \) and estimate them using a variant of the EM algorithm (Dempster et al., 1977). Note that this parameter set is also sufficient to uniquely identify the augmented system defined above in (3.8-3.9). The EM algorithm for LDS was first presented in (Shumway & Stoffer, 1982). We use a novel regularized variant of EM for LDS, based on work from (Schneider, 2001), where the estimation of regression coefficients in the M-step is done with ridge regression (Hastie et al., 2009). This technique was used in (Schneider, 2001) to deal with imputation of missing values, but we use the same idea to regularize estimates of our model parameters.

EM works by iteratively estimating the expected values of the latent state vectors \( x(t) \), given some values of the parameters \( \theta \) (the E-step), then fixing \( x(t) \) to estimate the optimal parameter values \( \theta \) (the M-step). Given some fixed parameter values for \( \theta \), the E-step requires us to estimate \( \{\hat{x}(1), \cdots, \hat{x}(T)\} = E[\{x(1), \cdots, x(T)\} | \{y(1), \cdots, y(T)\}] \), where \( T \) is the total number of samples in the training sequence. It is important to note that these are not the quantities estimated by the Kalman filter, which are \( E[x(t) | \{y(1), \cdots, y(t)\}] \) for all \( t = 1, \ldots, T \). The key distinction is that the filter only uses past measurements in its estimation of \( x(t) \), whereas we use future measurements as well. We obtain the estimates \( \{\hat{x}(1), \cdots, \hat{x}(T)\} \) by the Kalman smoother (or fixed interval smoother) (Gelb, 1974). Here, we used the efficient Rauch-Tung-Striebel (RTS) forward-backward algorithm (Rauch et al., 1965). The ‘forward’ pass of the algorithm is exactly the Kalman filter. The RTS algorithm is a specific case of the more general sum-product message-passing algorithm on Bayesian networks (Koller & Friedman, 2009). We do not derive the RTS algorithms here, but refer the reader to several appropriate sources: (Gelb, 1974; Anderson & Moore, 1979; Rauch et al., 1965).
For the M-step, we take the latent state vector sequence \( \{ \hat{x}(1), \cdots, \hat{x}(T) \} \), estimated in the previous step, and find the most likely set of parameters \( \hat{\theta} \) that produced it. This is done in the max likelihood framework by computing the full loglikelihood of the parameters, differentiating with respect to the parameters of interest, setting to zero and solving (Ghahramani & Hinton, 1996; Shumway & Stoffer, 1982). The max likelihood (unregularized) estimates of \( A_i \)s and \( B_j \)s are given by linear regression or ordinary least squares (OLS). To prevent overfitting, we modify the max likelihood estimation procedure by using ridge regression (Hastie et al., 2009) instead of OLS. We simultaneously estimate the state transition matrices \( A_i \), and control matrices \( B_j \) as shown below:

\[
AB_{ML} = \begin{bmatrix} A_1 & \cdots & A_{n_a} & B_1 & \cdots & B_{n_b} \end{bmatrix} = D F^{-1}, \quad \text{where (3.10)}
\]

\[
D = E [x(t) \cdot [\hat{x}^T(t) \ \ \hat{u}^T(t)]] \quad \text{and (3.11)}
\]

\[
F = E \left[ [\hat{x}(t) \ \ \hat{u}(t)] \cdot [\hat{x}(t) \ \ \hat{u}(t)]^T \right]. \quad \text{(3.12)}
\]

We apply the Moore-Penrose psuedoinverse when \( F \) is not invertible. We require the expected state covariance matrices \( E[\hat{x}(t) \cdot \hat{x}^T(t)] \) in (3.12) for this computation. But this is computed by the Kalman smoother in the E-step. Also, the expected covariance matrix \( E[x(t) \cdot \hat{x}^T(t)] \) in (3.11) is given by the submatrix formed by the first \( p \) rows of the full expected state covariance matrix. Since the inputs \( \hat{u}(t) \) are known to us, we can compute cross-terms in the covariance matrices by simply taking the outer product of expected state vectors \( \hat{x}(t) \) from the E-step with the appropriate known input vectors \( \hat{u}(t) \).

To regularize this estimate, we simply replace \( F \), which acts as the covariance estimate in the least squares, from (3.12) to \( \hat{F} \) as defined below:

\[
\hat{F} = F + r_{xu} \cdot \begin{bmatrix} \max_i (\text{var}(\hat{x}_i(t))) \cdot I_{n_a \times p} & 0 \\ 0 & \max_j (\text{var}(\hat{u}_j(t))) \cdot I_{n_b \times m} \end{bmatrix} . \quad \text{(3.13)}
\]
Scaling the penalty matrix (here, the identity matrix) relative to the input and state variances makes the procedure robust to differences in scale between these two terms. With this substitution, we get our regularized estimates of $A_i$s and $B_j$s:

$$\hat{A}\hat{B} = \begin{bmatrix} \hat{A}_1 & \ldots & \hat{A}_{n_a} & \hat{B}_1 & \ldots & \hat{B}_{n_b} \end{bmatrix} = D\hat{F}^{-1}. \tag{3.14}$$

Next, we obtain the max likelihood estimate of $Q$ as shown below:

$$\hat{Q} = E \left[ \left( \tilde{x}(t) - \tilde{A}\tilde{x}(t-1) - \tilde{B}\tilde{u}(t-1) \right) \cdot \tilde{x}^T(t) \right] \tag{3.15}$$

In our implementation, we use the regularized estimates of $\tilde{A}$ and $\tilde{B}$ that we get from (3.14). We assume $Q$ is a diagonal matrix, so we restrict the extended state covariance matrix $\hat{Q}$ to be diagonal.

The max likelihood estimates for $C$ are then given by:

$$\hat{C} = E \left[ y(t) \cdot x^T(t) \right] \cdot E \left[ x(t) \cdot x^T(t) \right]. \tag{3.16}$$

Finally, we use a regularized estimate of the measurement covariance matrix $R$, by first defining the max likelihood estimate $R_{ML}$ and scaling the penalty matrix, similar to what we did in (3.13):

$$\hat{R} = R_{ML} + r_R \cdot \max \left( \text{diag}(R_{ML}) \right) \cdot I_{\# \text{channels}}, \quad \text{where} \tag{3.17}$$

$$R_{ML} = E \left[ (y(t) - Cx(t)) \cdot y^T(t) \right]. \tag{3.18}$$

Once again, the regularization parameter scales with the measurements levels, to be robust to any scaling that may occur in the signals or preprocessing. We choose to constrain $R$ to be diagonal, despite using regularization to further reduce overfitting; however this
procedure works without any modifications if we estimate the entire $R$ matrix.

### 3.4.2 Uniqueness of Parameter Estimates

It is important to note here that the optimal solution for the LDS model is not uniquely defined. Namely, we can find equivalent solutions by similarity transformations. Given values for $\theta$ and $x(t)$ and any $p$-by-$p$ invertible matrix $K$, we can define an equivalent solution, i.e., one that produces the same measurements $y(t)$ with the same likelihood, by transforming the given solution as follows:

\[
\begin{align*}
  x(t) \rightarrow K x(t), \quad & A_i \rightarrow K A_i K^{-1}, \quad B_j \rightarrow K B_j, \quad Q \rightarrow K Q K^T, \quad C \rightarrow C K^{-1}.
\end{align*}
\]

(3.19)

We can easily verify that by making the substitutions above into (3.1), the resulting equation would be the same as left-multiplying both sides of (3.1) by $K$. Thus, all the states $x(t)$ are transformed to $K x(t)$ and with the new measurement matrix, the observations are $C K^{-1} K x(t) = C x(t) = y(t)$, the same as in the original model formulation. Constraining the model such that the state noise covariance matrix $Q$ is diagonal circumvents this rotational invariance problem. It also allows for a simpler interpretation of the autoregressive connectivity parameters between the neural sources, as we are implicitly assuming independent neural noise within each source. Note however, that this restriction does not prevent the fitting algorithm from swapping components of $x$ or scaling components by arbitrary constants.

### 3.4.3 Initialization using CCA

The performance of the EM algorithm for the LDS model strongly depends on the initialization, as the likelihood function for the model in (3.8-3.9) is not convex. We initialize values of the observation matrix $C$ through canonical correlation analysis (CCA) (Hotelling, 1936). CCA is performed between a 1-dimensional input signal (component of the input vector $u(t)$) and the $\#$channels-dimensional time-lagged EEG signal. CCA yields
the weights on each channel that maximize the correlation between the corresponding EEG subspace and the input signal. If the input signals are speech envelopes, time lags between \(160 - 240\, \text{ms}\) are suggested (guided by the analysis in (Power et al., 2012) and (Ding & Simon, 2012)) and the lag with the maximum canonical correlation is selected. The weights on the EEG channels yielded by the CCA with this maximal lag are normalized and used as columns of \(C\), with each of the \(p\) columns corresponding to a neural source (component of state vector \(x(t)\)).

To obtain initial values for autoregressive matrices \(A_i\) and impulse responses \(B_j\), we used the \(p\) EEG subspaces corresponding to the columns of the initial \(C\), making the assumption that these subspace signals were exactly equal to the latent state values. We estimated parameters \(\hat{A}\) and \(\hat{B}\) using ridge regression with the same regularization parameters \(r_{xu}\) as in (3.13). The input to the MVARX model is the \(m\)-dimensional input vector \(u(t)\). Using these parameters, we compute the max likelihood estimate of \(Q\) and ridge regression estimate of \(R\), again using the same regularization parameters \(r_r\) as in (3.17) to initialize the EM algorithm.

### 3.5 Model Validation on Data

#### 3.5.1 Data Acquisition and Stimuli

We used EEG data gathered for a previous study (Power et al., 2012). Subjects with no reported history of hearing loss or neurological disorders were recorded with a BioSemi ActiveTwo system at 130 electrode locations, with a sampling rate of \(512\, \text{Hz}\). Audiobooks of two works of fiction (“Journey to the Centre of the Earth” and “20000 Leagues Under the Sea”) were presented, one in each ear. The story–ear pairing was fixed for each subject for the duration of the experiment, but randomized amongst the different subjects. The books were divided into 30 blocks, each approximately 1 min in length. RMS intensities of the stories were normalized to be equal in each block and silent gaps were truncated, so as to be at most \(0.5\, \text{s}\) in length. Subjects were also given multiple choice questions at the
end of each block on the attended and unattended stories. Behavioral performance was significantly above chance for the attended story, but not for the unattended story. Please see (Power et al., 2012) for further details.

We analyzed a subset of 15 randomly chosen subjects out of the 40 collected. Envelopes for the two stories were computed by passing the stories through a cochlear (gammatone) filter bank, computing narrowband envelopes using the Hilbert transform, and lowpass filtering the summed narrowband envelopes with a cutoff of 15 Hz. We use a filter bank to compute envelopes so that frequency modulations, which are known to elicit EEG responses (Obleser et al., 2012b), may be accounted for as they are encoded as amplitude modulations in a filter bank (Ghitza et al., 2012).

3.5.2 Data Preprocessing

Bad channels were manually rejected and replaced with the average of neighboring channels, in order to maintain consistency. The signals were downsampled to 64 Hz and re-referenced to the average of all the channels. We used the Signal Space Projection method (Uusitalo & Ilmoniemi, 1997) to project out the subspace corresponding to eye blinks. This is done by looking at the signals from the frontal EEG electrodes, which typically contain large EOG artifacts, and manually selecting a few 20ms epochs corresponding to eye blinks. Principal Component Analysis was then performed on these selected epochs and the component with the largest variance (corresponding to blink artifacts) was projected out. Only the 32 channels corresponding to the standard 10−20 system were then used in this analysis, as we found that using more channels did not significantly change the results or improve classification performance.

3.5.3 Model Specification and Inference Parameters

The model parameter values for defining the MVARX dynamics are \( n_a = n_b = 16 \), indicating that the autoregressive window and impulse response are 250ms long, consistent with typical evoked responses in EEG analysis. We assume there are two inputs to the
system, i.e. \( m = 2 \), corresponding to the broadband envelopes of the two stories presented. The dimensionality of the latent neural state space, is set to \( p = 2 \): one is initialized by CCA with the attended speech envelope and the other by CCA with the unattended speech envelope. The regularization hyperparameters used in this study were \( r_{xu} = r_r = 0.6 \) and were manually chosen so that learned parameter estimates across cross-validation runs were roughly similar, as dissimilar estimates strongly indicate overfitting. The EM algorithm was run for 20 iterations, as we observed that this was more than sufficient for convergence with all subjects.

## 3.6 Results

We ran our generalized EM algorithm on fifteen subjects. For each subject, we trained and tested the model using leave-one-out cross validation (i.e., each block was set aside as the test block and the remaining blocks were used to perform system identification). To test for convergence in the EM algorithm, we manually set a threshold of change in latent variable estimates over consecutive iterations. We noted that 20 iterations were typically sufficient to meet this threshold.

For testing, we ran two Kalman filters (Gelb, 1974), one with the attended and unattended speech in the same ordering of components in \( u(t) \) as the training data, and another with them swapped. We computed the log-likelihoods of the data given the model and the inputs, and selected the one with the larger log-likelihood to be the one with the correct ordering of inputs (thereby giving us the attended stream). This was done over test blocks of varying durations, ranging from 5s to 1min long with 5s increments. We analyzed performance as a function of the test block duration. Mean performance across all subjects is plotted in Fig 3.2.

On average, we were able to decode the attended stream significantly better than chance (binomial test, \( p < 0.05 \)), even with the test blocks that were as short as five seconds. The mean performance with one minute long test blocks was \( 84.6\% \pm 11.7\% \). Six of the fifteen
subjects were at or above 90% correct for one minute long test blocks. Only one subject did not perform significantly better than chance, even with one-minute long test blocks. As expected, we saw a decrease of performance gradually as we shortened the test block durations. Test performance for individual subjects is listed in Table 3.1.

![Graph of Test Performance of the State-Space Model](image)

Figure 3.2: Performance of a Kalman-filter based classifier identifying the attended and unattended speech streams. Leave-one-out cross-validation was performed, setting aside one block as the test set with all the other blocks used to identify system parameters. Varying durations of time windows from the test block were used to predict the attended stream based on the difference in the log-likelihoods of two competing models (corresponding to the choice of two possible speech streams to attend). The solid line denotes the mean performance across subjects ($N = 15$) and the shaded area corresponds to ± one standard deviation of the mean. The dashed line denotes chance performance (with $p < 0.05$), assuming a binomial distribution.

### 3.7 Interpretation of Estimated Parameters

To extract the impulse response from the fitted parameters, we note that in addition to the matrices $B_j$, the exogenous inputs $u(t)$ also affect $x(t)$ via the autoregressive terms $A_i$. 
Table 3.1: Test Performance by Subject

<table>
<thead>
<tr>
<th>Subj</th>
<th>60s</th>
<th>50s</th>
<th>40s</th>
<th>30s</th>
<th>20s</th>
<th>10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>100%</td>
<td>100%</td>
<td>96.7%</td>
<td>96.7%</td>
<td>93.3%</td>
<td>80.0%</td>
</tr>
<tr>
<td>AP</td>
<td>96.6%</td>
<td>96.6%</td>
<td>96.6%</td>
<td>93.1%</td>
<td>86.2%</td>
<td>86.2%</td>
</tr>
<tr>
<td>EL</td>
<td>93.3%</td>
<td>93.3%</td>
<td>90.0%</td>
<td>86.7%</td>
<td>93.3%</td>
<td>83.3%</td>
</tr>
<tr>
<td>EK</td>
<td>93.3%</td>
<td>90.0%</td>
<td>86.7%</td>
<td>86.7%</td>
<td>83.3%</td>
<td>73.3%</td>
</tr>
<tr>
<td>BWM</td>
<td>93.3%</td>
<td>83.3%</td>
<td>86.7%</td>
<td>86.7%</td>
<td>76.7%</td>
<td>66.7%</td>
</tr>
<tr>
<td>AS</td>
<td>93.3%</td>
<td>83.3%</td>
<td>86.7%</td>
<td>83.3%</td>
<td>76.7%</td>
<td>66.7%</td>
</tr>
<tr>
<td>DCM</td>
<td>89.7%</td>
<td>82.8%</td>
<td>79.3%</td>
<td>72.4%</td>
<td>72.4%</td>
<td>72.4%</td>
</tr>
<tr>
<td>CN</td>
<td>86.7%</td>
<td>90.0%</td>
<td>93.3%</td>
<td>86.7%</td>
<td>66.7%</td>
<td>56.7%</td>
</tr>
<tr>
<td>FW</td>
<td>83.3%</td>
<td>80.0%</td>
<td>83.3%</td>
<td>70.0%</td>
<td>63.3%</td>
<td>63.3%</td>
</tr>
<tr>
<td>BR</td>
<td>83.3%</td>
<td>80.0%</td>
<td>80.0%</td>
<td>76.7%</td>
<td>80.0%</td>
<td>70.0%</td>
</tr>
<tr>
<td>CM</td>
<td>80.0%</td>
<td>76.7%</td>
<td>83.3%</td>
<td>73.3%</td>
<td>76.7%</td>
<td>50.0%</td>
</tr>
<tr>
<td>CB</td>
<td>79.3%</td>
<td>82.8%</td>
<td>69.0%</td>
<td>65.5%</td>
<td>69.0%</td>
<td>62.1%</td>
</tr>
<tr>
<td>BM</td>
<td>70.0%</td>
<td>66.7%</td>
<td>56.7%</td>
<td>80.0%</td>
<td>66.7%</td>
<td>63.3%</td>
</tr>
<tr>
<td>DC</td>
<td>66.7%</td>
<td>70.0%</td>
<td>66.7%</td>
<td>56.7%</td>
<td>56.7%</td>
<td>66.7%</td>
</tr>
<tr>
<td>DW</td>
<td>60%</td>
<td>63.3%</td>
<td>73.3%</td>
<td>56.7%</td>
<td>60.0%</td>
<td>53.3%</td>
</tr>
</tbody>
</table>

The final impulse response of the system is then given by the following matrix polynomials:

\[
A(z^{-1}) = 1 + A_1 z^{-1} + \ldots + A_n z^{-n_a}, \quad \text{and} \quad (3.20)
\]

\[
B(z^{-1}) = B_1 z^{-1} + \ldots + B_n z^{-n_b}. \quad (3.21)
\]

where \( z^{-1} \) is the unit-time-shift operator. We can then express the MVARX model (3.1) in terms of this \( z^{-1} \) operator (i.e. take the \( z \)-transform of the system), to get the expression:

\[
A(z^{-1})x(t) = B(z^{-1})u(t) + \epsilon(t) \quad (3.22)
\]

\[
\Rightarrow x(t) = \frac{B(z^{-1})}{A(z^{-1})} u(t) + \frac{1}{A(z^{-1})} \epsilon(t). \quad (3.23)
\]

Thus, we see that the impulse response can be written in pole-filter form as \( \frac{B(z^{-1})}{A(z^{-1})} \), the noise spectrum is \( \frac{1}{A(z^{-1})} \). An inverse \( z \)-transform can be used to estimate the impulse response in the time domain. These systems generally act as infinite impulse response systems.
(IIR) filters, except in the degenerate case when \( A(z^{-1}) = 1 \). The impulse responses for the attended and unattended stories for both neural sources, along with the lead fields (columns of the \( C \) matrix) for the corresponding sources, are plotted for each subject in Fig 3.3.
Subject BWM

Subject CB

Subject CM

Subject CN

Subject DC

Subject DCM

(e) BWM parameter estimates

(f) CB parameter estimates

(g) CM parameter estimates

(h) CN parameter estimates

(i) DC parameter estimates

(j) DCM parameter estimates
Figure 3.3: Lead fields of the two neural sources are plotted in A and B, while the impulse responses of the of these sources to the attended story (in blue) and unattended story (in red) are plotted below in C and D, respectively. Each subplot corresponds to the parameters estimated from the data of one subject. Parameters were obtained by averaging parameter estimates from the cross-validation runs. Due to the scale indeterminacy of the state space model formulation, the plots are left as unitless quantities. Note the different scales used in subfigure (3.3i), as the estimates were unusually large, possibly due to large outliers in the data.

First we consider subjects where the model was able to predict the attended stimulus well (> 90%): subjects DM (Fig 3.3k), AP (Fig 3.3a), EL (Fig 3.3n), EK (Fig 3.3m), BWM (Fig 3.3e), AS (Fig 3.3b). We note that for these subjects the impulse responses look distinct for the attended and unattended envelopes. Typically, the impulse response to the attended stimulus has a larger gain than the impulse response to the unattended stimulus, consistent with some kind of attention-modulated gain control mechanism in the neural coding of speech.

Subjects that performed poorly either had impulse responses that looked very similar for the attended and unattended envelopes, e.g. subject DW (Fig 3.3l), or that were degenerate, e.g. subject DC (Fig 3.3i - note the change in scales for this subplot). One possible explanation for the latter is the presence of a large number of outliers in the data.

Note also, that the lead fields for subject AS (Fig 3.3b) are essentially negative scalar multiples of each other (the cross-covariance of the two lead fields has a values of $<-0.99$).
This indicates that we have effectively estimated only one neural source for the subject. Despite this, the model is able to correctly predict the attended stimulus 93.3% of the time with one-minute long blocks.

We refrain from interpreting the lead fields because in order to make inferences about the underlying cortical substrates, we require anatomical data (such as MRI scans) of the subjects. Without such anatomical data to help constrain the inverse solutions, we cannot adequately model the electrical field propagation of the EEG measurements.

3.8 Conclusions and Extensions

In this chapter, we presented a state-space model of a linear dynamical system to fit the neural response to the envelopes of ongoing continuous speech streams presented in a cocktail party setting. The model assumes multiple neural sources with separate impulse responses to attended and unattended speech envelopes and multivariate autoregressive modeling of neural noise and functional connectivity. We use the model to predict which of two simultaneously presented speech sources was selectively attended to by a listener. Classification for six out of fifteen subjects was > 90% with one minute long blocks of speech. The estimation procedure uses Kalman filters, allowing for efficient online estimation. The success of this approach suggests that this kind of analysis could be applied to create auditory attention-driven brain-computer interfaces or intelligent assistive listening devices.

The general state-space framework presented here can easily be extended in multiple ways. These could include incorporating nonlinear impulse responses (NFIR models) or nonlinear autoregressive effects (NARX models) (Ljung, 1998). A possibility for improving the model inference is to use robust inference techniques (Wager et al., 2005), which would make the model inference more robust to outliers. This can also be done by using a Student t-distributed noise model instead of Gaussian noise. The heavy-tailed Student t-distribution is not as sensitive to large outliers and thus is better equipped to deal with
the EEG noise distributions.

We may also perform full Bayesian learning with a state-space model, which we use with variational Kalman filtering and smoothing algorithms (Beal, 2003). We could also very naturally incorporate the Student t-distributed noise assumptions in such models, similar to the approaches used in (Christmas & Everson, 2011; Wu et al., 2011). Such Bayesian models would need to be learned with the Variational Bayes algorithm (the high dimensional learning in this model would prohibit the use of sampling based learning such as Markov Chain Monte Carlo methods). An additional benefit of Bayesian learning would be that we would not have the need for explicit regularization, as it would incorporate estimates of parameter confidence in the fitting procedure.
Chapter 4

Audio–Visual Speech

4.1 Introduction

In daily face-to-face interactions, speech information is communicated through multiple modalities; in addition to auditory signals, visual and even tactile signals (Gick & Derrick, 2009) can convey speech information. Visual cues are known to provide robust cues that can, particularly for some articulatory gestures, be more informative than the corresponding auditory cues. A good example is the famous McGurk effect (McGurk & MacDonald, 1976). Speech-reading cues have long been known to provide substantial (up to 15 dB) benefit to understanding speech in noise (Sumby & Pollack, 1954). Dynamic cues are known to provide a substantial benefit over comparable static cues, such as photos of speakers making the appropriate articulatory gestures at the correct times (Rosenblum & Saldaña, 1996, 1998).

Cues such as the area of the mouth aperture and the maximum labial distance are known to correlate with envelopes of the second and third formant regions (Grant & Seitz, 2000; Chandrasekaran et al., 2009). However it must be noted that the area of mouth aperture is not likely to provide speechreading benefit by itself, as studies have shown that simulating envelope cues as an annulus with changing diameter does not provide a cross-modal benefit in either humans or monkeys (Bernstein et al., 2004; Ghazanfar et al., 2005). By using movement signals from more distal points on the face, one can reconstruct the broadband speech envelope with high accuracy (Munhall & Vatikiotis-Bateson, 1998; Slaney & Covell, 2000), so visual information provided by dynamic movements of the
face provide envelope cues from the entire speech spectrum. It should be noted that psychophysical results show varying levels of speechreading benefits for speech filtered into frequency bands corresponding to formants (Grant et al., 1991). This suggests that speechreading cues are not equally informative across the speech spectrum and may be redundant with some acoustic frequency bands.

The neural mechanisms through which visual speechreading cues enhance speech perception still remain largely unknown, partly due to the inherent limitations involved in conducting invasive electrophysiological studies in humans. Numerous studies point to cross-modal regions, particularly the superior temporal sulcus (STS), as being the site at which visual and auditory signals converge, since it is selective to both auditory and visual speech stimuli (Puce et al., 1998; Beauchamp et al., 2004; Pelphrey et al., 2005; Thompson et al., 2007). This characterization is consistent with the role STS is thought to play in auditory speech processing, which is to convert acoustic information to pre-lexical phonemic categories (Hickok & Poeppel, 2007; Kumar et al., 2007; Obleser & Eisner, 2009). While the STS is known to be selective for dynamic face stimuli, it contains some subdivisions that are activated only by auditory speech, and others by visual-only speech (talking faces); some (but not all) of the visual-only subregions show increased activity during audio–visual speech, while the auditory subregions uniformly do not (Beauchamp et al., 2004; von Kriegstein et al., 2008).

Recently, however, several studies have shown cross-modal effects earlier in the cortical hierarchy, namely in the primary sensory cortices. For instance, primary auditory cortex is known to be activated during speechreading, even when there is no auditory signal present (Calvert et al., 1997; Hall et al., 2005; Pekkola et al., 2005; Besle et al., 2008). Visual stimuli are also known to modulate activity in auditory cortex, especially in the case of audio–visual speech stimuli. A proposed mechanism for audio–visual speech integration argues that since visual cues temporally precede corresponding auditory cues (Chandrasekaran et al., 2009), visual regions send predictive signals to prime auditory regions preparing them to expect specific phonemic or place-of-articulation cues. (Lakatos et al., 2007; Okada
et al., 2013). Phase resetting of oscillations has been proposed as a coding mechanism for these expectations (Schroeder et al., 2008).

It is also the case that auditory-only cues can elicit or modulate activity in the visual subregions of STS. It has been shown by von Kriegstein et al. (2008), using fMRI, that a brief period of audio–visual exposure to a speaker helps improve the speaker’s intelligibility in audio-only speech conditions. Across subjects, these intelligibility benefits were also shown to correlate with increased activation in visually selective subregions of STS. Another fMRI study (Pekkola et al., 2006) found that STS was bilaterally activated by the presentation of audio–visual speech, regardless of whether it was attended or not. Significant attentional modulation was found in planum temporale in the left hemisphere, a secondary auditory cortical region that is earlier in the auditory processing cortical hierarchy than STS and is strongly implicated in processing speech.

Primary auditory cortex (located in the superior temporal gyrus) was also found to be involved in audiovisual speech processing (relative to audio only), but exhibited large individual differences across subjects (Hall et al., 2005). The activity levels (in left STG) were strongly correlated with the speechreading benefit obtained by subjects, showing that early sensory integration of modalities is both possible and beneficial for processing of audio–visual speech.

In this chapter, we deploy the state-space model, introduced in the previous chapter, to model the neural dynamics of audio–visual speech processing. In particular, we are interested in revealing the dynamics and interactions of early sensory cortices in attentional processing of audiovisual speech. To this end, we fix the lead fields of our latent sources as the fields defined by auditory and visual N100 ERP peaks, corresponding to primary auditory and visual cortices, respectively. These ERPs were collected for each subject in blocks separate from the audio–visual speech responses.

The estimated autoregressive and impulse response matrices, with these lead field constraints, reveal a speed-up in processing of attended audio–visual speech streams for 7 of the 8 subjects tested. Here, processing speed is defined as latency of the peak of the impulse
response. However, we did not find any reliable pattern for the gains of attended and unattended streams. We also found strong causal interactions, with visual cortex modulating the activity in auditory cortex, with no causal interactions in the opposite direction.

4.2 Methods

4.2.1 Stimulus and Task

Eight normal-hearing (defined as having thresholds within 20 dB of normal-hearing thresholds in each ear) listeners who were native speakers of American English, participated. Each gave informed consent approved by Boston University. The ages of the participants ranged from 18 to 25 years. Six subjects were female. Subjects were seated in front of a monitor at a distance of 60 cm and two videos of speakers telling stories were presented on a monitor, 10 degrees apart. The videos were front-facing, well-lit views of speakers’ faces, with clearly visible orofacial movements, telling personal stories at events organized by *The Moth Radio Hour* (Atlantic Public Media), and graciously donated to us for the purposes of this experiment. Subjects were free to move their eyes, but were instructed to look at the attended speaker in all conditions.

The audio of the two stories were set to have constant and equal levels across time, by normalizing the RMS power of each story to a fixed level within a running 3 s window. This normalization produced no audible artifacts or distortions. Normalized audio tracks of the two stories were summed and presented diotically with diotic speech-shaped noise at an SNR of $-10$ dB. The noise was presented at a level of roughly 60 dB SPL, but subjects were allowed to change the overall level of the mixture at the beginning of the experiment to a comfortable, but clearly audible level. Four pairs of stories were presented in blocks lasting between 10 and 12 minutes. Two blocks have two male speakers, one had two female speakers, and one had one male and one female speaker. Subjects were instructed to listen to either the left or the right speaker in a given block, randomly assigned. They attended both left and right for two blocks each. They were asked to briefly recount the
attended and unattended stories at the end of the blocks. All subjects could summarize
the attended story and reported not knowing the contents of the unattended story in each
block.

Each block was divided into four equal-length contiguous sections. The first and third
sections were presented in the synchronous condition and the other two, asynchronous. The
videos in the two asynchronous blocks were swapped with each other, without any change
in audio signals. So in these sections, subjects were presented with a moving face of the
speaker, but the visual cues contained no speechreading information about the attended
audio story. Subjects continued to attend the same speaker in all sections of a block.

Subjects were seated in front of a monitor at a distance of 60 cm in a sound-proof
booth, and the two videos subtended horizontal angles of roughly ±10 degrees from mid-
line. Videos were rendered offline with MATLAB (The Mathworks Inc.) and displayed
with the JRiver Media Center Software (JRiver Inc.) to ensure reliable audio–visual syn-
chronization. Audio was presented through a MOTU UltraLite-mk3 Hybrid (MOTU Inc.)
at a sampling rate of 48 kHz using ASIO drivers. Audio signals were delivered over tube
phones to ER-1 (Etymotic Inc.) insert earphones.

4.2.2 Data Collection and Pre-Processing

EEG signals were recorded using a 64 channel cap with a BioSemi ActiveTwo active
electrode system at a sampling rate of 2048 Hz. Bad channels were manually rejected and
replaced by a signal interpolated from the average of neighboring clean channels, in order
to maintain consistency. Signals were then re-referenced to the average of the 64 EEG
channels. We used the Signal Space Projection method (Uusitalo & Ilmoniemi, 1997) to
project out the subspace corresponding to eye blinks. After this, only the 32 channels
corresponding to the standard 10–20 cap were used for further analysis. Signals were
then bandpass filtered between 1 and 15 Hz and downsampled to 64 Hz offline.
4.2.3 Model Parameter Estimation

Envelopes for the two stories were computed by passing the stories through a cochlear (gammatone) filter bank, computing narrowband envelopes using the Hilbert transform, and lowpass filtering the summed narrowband envelopes with a cutoff of 15 Hz. We used these envelopes to train dynamical systems models (as presented in the previous chapter) for both synchronous and asynchronous data separately. The model parameter values for defining the MVARX dynamics are $n_a = n_b = 16$, indicating that the autoregressive window and impulse response are 250 ms long. The broadband envelopes of the two speakers specified the two inputs to the system (thus, $m = 2$).

The dimensionality of the latent neural state space was set to $p = 2$. Rather than learning the lead fields with CCA as we did in Sec. 3.4.3, we fixed the columns of lead field matrix $C$ to be proportional to the N100 peaks of visual and auditory ERPs elicited by flash and tone onsets, respectively. The N100 ERPs were normalized by the standard deviation of baseline activity to account for variable noise levels across channels. The lead field vectors were also normalized to have unit norm. For all but one subject (S06), these N100 data were gathered in separate blocks after the experimental trials. Lead fields for subject S06 were obtained by averaging lead fields of all other subjects.

The remaining parameters of the model ($A_i$, $B_j$, $Q$, and $R$) were trained using the modified regularized EM algorithm, as in Sec. 3.4.1, with regularization parameters $r_{xu} = r_r = 0.7$. The algorithm was run for 10 iterations to obtain estimates of the model parameters after the algorithm converged. We divided the 8 (2 per blocks, for 4 blocks) synchronous and asynchronous sections into 22 one-minute-long epochs, each. Leave-one-out cross-validation was performed on these blocks to assess performance in correctly identifying the attended speaker.
4.2.4 Analysis of Inferred Parameters

The parameter estimates across cross-validation runs were used to generate jackknifed estimates (Efron, 1982) of these parameters separately in the synchronous and asynchronous conditions. Using these estimates, we can test for significant differences across subjects in the dynamics of audio-only (asynchronous) and audio–visual (synchronous) speech processing. We are specifically interested in comparing the latencies and gains of the impulse responses to the attended speaker from the auditory ROI, as well as testing the hypothesis that visual regions communicate articulatory information to auditory regions in order to help predict upcoming phonemes.

4.2.4.1 Latencies and Gains of Impulse Responses

We calculated impulse responses for the attended and unattended speakers in each condition, as in the previous section, by inverse z-transforming the pole-zero representation $\frac{B(z^{-1})}{A(z^{-1})}$. We then interpolated these impulse responses using cubic splines to get better (sub-millisecond) temporal resolution in our latency and gain estimates. All our auditory impulse responses showed characteristic negative deflections between 50 and 150 ms, similar to an N100 evoked response. We therefore extracted the minimum value of the impulse responses and corresponding time points at which they occur, in the 50 – 150 ms interval in each cross-validation run. The average of these values and times provided jackknifed estimates of impulse response gains and latencies for each subject.

4.2.4.2 Causal Influence of Auditory Cortex from Visual Cortex

We tested the hypothesis that visual regions causally influence processing in auditory regions during audio–visual speech perception. To do this, we compute the partial Granger causality (Guo et al., 2008) from the estimated visual cortex to the estimated auditory cortex activity. These estimates are provided by the corresponding rows of the latent state space activity estimate, $x(t)$. Partial Granger causality (PGC) estimates the causal in-
fluence between two components of a multivariate autoregressive model by accounting for indirect influences through observed (the exogenous inputs, in our case) and potentially unobserved components of the autoregressive model. The classical Granger causality measure (Geweke, 1984) only accounts for effects of the observed components. Since we only model two ROIs in this experiment, classical Granger causality cannot account for spurious connectivity caused by the activity of other regions involved in audio–visual speech perception, such as parietal cortex or STS; therefore, we adopted the more conservative PGC measure.

Given two neural signals \( x_1(t) \) and \( x_2(t) \), we say that \( x_2 \) Granger causes \( x_1 \) if the past of \( x_2 \) helps predict \( x_1 \), conditioned on the information contained in the past of \( x_1 \) alone. Specifically, we fit a bivariate autoregressive model to \( x_1 \) and \( x_2 \) as follows:

\[
\begin{align*}
    x_1(t) & = a_{1,1} \cdot x_1(t-1) + \ldots + a_{1,n_a} \cdot x_1(t-n_a) + \\
             & \quad a_{2,1} \cdot x_2(t-1) + \ldots + a_{2,n_b} \cdot x_2(t-n_a) + \epsilon_1(t), \\
    x_2(t) & = b_{1,1} \cdot x_1(t-1) + \ldots + b_{1,n_a} \cdot x_1(t-n_a) + \\
             & \quad b_{2,1} \cdot x_2(t-1) + \ldots + b_{2,n_b} \cdot x_2(t-n_a) + \epsilon_2(t).
\end{align*}
\]

Note that we have omitted the exogenous input terms here for simplicity, but they may be included, without loss of generality, to the procedure that follows.

For classical Granger causality, we fit the full model described above and get the error variance for the \( x_1 \) term as \( \sigma_1 = \text{var} (\epsilon_1(t)) \). We then set the terms \( a_{2,1}, \ldots, a_{2,n_b} \) to zero and re-fit this restricted model. The new error variance term for \( x_1 \) in this restricted model is given by \( s_1 \). The classical Granger causality measure is defined as:

\[
GC_{2 \rightarrow 1} = \log \left( \frac{\sigma_1}{s_1} \right).
\]

Partial Granger causality is derived from a similar intuition as partial correlation and partial coherence (Guo et al., 2008). To compute this, we simply replace the variance terms...
in Eq. (4.1) by the partial correlations to get:

\[
P_{GC_{2\rightarrow1}} = \log \left( \frac{|\sigma_{1} - \sigma_{1,2} \cdot \sigma_{2}^{-1} \cdot \sigma_{2,1}|}{|s_{1} - s_{1,2} \cdot s_{2}^{-1} \cdot s_{2,1}|} \right),
\]

(4.2)

where \(\sigma_{2} = \text{var}(\epsilon_{2}(t))\), \(\sigma_{1,2} = \text{cov}(\epsilon_{1}(t), \epsilon_{2}(t))\) are covariance terms in the full model and \(s_{2}, s_{1,2}, s_{2,1}\) are the corresponding covariance terms in the restricted model. Partial Granger causality has shown to eliminate the influence of exogenous inputs and variables that are not modeled in the data (Guo et al., 2008; Roelstraete & Rosseel, 2012), both of which may be present in our model.

We applied this PGC measure by fitting autoregressive models to the estimated neural latent space time series. It is important to note that we do not re-estimate the neural latent space time series when we fit the restricted model. Thus this procedure does not propagate the uncertainty present in the measurement model, implicitly assuming that the neural latent space activity estimated during the EM fit was error-free. We computed the PGC measures for each subject separately in the synchronous and asynchronous conditions, and tested for differences on these measures across subjects using a paired t-test.

### 4.3 Results

We first test the model performance in correctly identifying the attended speaker using cross-validation separately for the synchronous and asynchronous conditions. As in the previous chapter, two Kalman filters were run on the test blocks (each one minute long), each with one of the speakers being attended. Log likelihoods of the two filters were compared to determine which was the more likely attended speaker. Performance for each of the subjects in both conditions are reported in Table 4.1. No consistent difference was noted in the synchronous and asynchronous condition performance. We perform a one-tailed binomial test on all 44 trials (combined across both conditions) to confirm that performance on test blocks did not occur by chance and report the \(p\)-values in Table 4.1 as well. The model was able to identify the attended speaker more consistently than would
be expected by chance ($p < 0.05$) for all but one (S03) of the eight subjects.

Table 4.1: Synchronous and Asynchronous Test Performance by Subject

<table>
<thead>
<tr>
<th></th>
<th>Synchronous blocks</th>
<th>Asynchronous blocks</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>81.8%</td>
<td>86.4%</td>
<td>$&lt; 10^{-6}$</td>
</tr>
<tr>
<td>S02</td>
<td>90.9%</td>
<td>86.4%</td>
<td>$&lt; 10^{-8}$</td>
</tr>
<tr>
<td>S03</td>
<td>54.6%</td>
<td>59.1%</td>
<td>0.146</td>
</tr>
<tr>
<td>S04</td>
<td>63.6%</td>
<td>63.6%</td>
<td>0.024</td>
</tr>
<tr>
<td>S05</td>
<td>54.6%</td>
<td>72.7%</td>
<td>0.024</td>
</tr>
<tr>
<td>S06</td>
<td>72.7%</td>
<td>77.3%</td>
<td>$&lt; 10^{-3}$</td>
</tr>
<tr>
<td>S07</td>
<td>63.6%</td>
<td>81.8%</td>
<td>$&lt; 10^{-3}$</td>
</tr>
<tr>
<td>S08</td>
<td>59.1%</td>
<td>72.7%</td>
<td>0.011</td>
</tr>
</tbody>
</table>

We now turn our attention to the estimated model parameters across the two conditions. We plot the impulse responses for the attended and unattended stream for each subject from the auditory and visual cortex ROIs in Fig 4.1. We note a stereotyped impulse response at the auditory cortex, roughly following the shape of a N100-P200 ERP complex. Note that subject S03, for whom our classifier performed at chance levels, the impulse response seems to contain an additional component (Fig. 4.1c), possibly due to overfitting to noise.
Figure 4.1: Lead fields of the auditory and visual cortical sources (as defined by N100 localizers) are plotted in A and B, respectively. The impulse responses of these sources to the attended speech (in blue) and unattended speech (in red) during the synchronous (audio–visual) condition for the auditory source is plotted in C, and for the visual source, in D. Similarly, the asynchronous (audio-only) condition impulse responses are plotted in E and F. Each subplot corresponds to the parameters estimated from the data of one subject and each trace corresponds to a single cross-validation run. Note that each subplot uses a different scale. The lead fields for S06 in Fig. 4.1f were computed as the average of corresponding lead fields of all other subjects, as the ERP blocks were not recorded for that subject.

We extract the latencies and gains for the early negative deflections (in the 50 − 150 ms range) in the auditory cortex impulse responses and compared them for audio–visual and audio-only processing. Across all subjects, the latencies were significantly shorter during audio–visual speech processing (one-tailed paired t-tests, $p < 0.05$), with an average latency decrease of $13.6 \pm 6.6$ ms (std. error). No significant differences were found in impulse response gain across the two conditions.

We also tested for task-related functional connectivity differences between visual and auditory cortex signals using partial Granger causality. Particularly, we tested for an increase in causal flow of information from visual to auditory cortex in the synchronous condition, as compared to the asynchronous condition. Across subjects, we found a significant increase in partial Granger causality in the synchronous condition (one-tailed paired
t-test, \( p < 0.05 \), confirming our hypothesis. Across subjects, the average PGC for the synchronous condition was 0.141 ± 0.014 (one standard deviation) and the average PGC in the asynchronous condition was 0.111 ± 0.013 (standard deviation).

4.4 Discussion

A popular hypothesized mechanism for explaining audio–visual speech integration states that the visual processing system sends predictive signals to auditory processing regions to prime the expectation of phonemic or articulation cues (Lakatos et al., 2007; Okada et al., 2013). This is supported by the basic stimulus property of visual articulation cues preceding voice onset (Chandrasekaran et al., 2009). Several prior studies have employed indirect approaches to measure this connectivity by measuring the degree of phase locking and phase patterns in cortical responses (Luo et al., 2010) and by looking at the evoked responses to monosyllabic tokens presented in audio-only and audio–visual conditions (van Wassenhove et al., 2005). Interestingly, the latter study shows that the reduction in ERF latencies is strongly correlated with performance in identification of the phoneme by visual speech alone, a reasonable index for the amount of visual information that can be extracted by listeners for a given phoneme. However, this result only establishes that there is some change in neural dynamics during audio–visual speech processing and does not directly measure the transmission of information. By measuring directed, causal interactions between auditory and visual regions, we demonstrate here for the first time that visual regions communicate to early auditory cortex in audio–visual speech processing.

As noted earlier in Sec. 4.2.4.2, our estimate of partial Granger causality assumes efficient estimation of the neural latent states during the expectation maximization parameter estimation, and computes PGC assuming those estimates are the true neural states. As Granger causality is essentially a model comparison, it would be of interest to re-fit the entire state-space model with the restricted parameter set to compute the measure. However, since the latent neural states are not directly observed, the exact computation of the
Granger causality measure is not as clearly defined; specifically it is not clear if one should use the EEG measurement error or the neural latent state noise covariance terms in the computation. We reserve these questions for future studies.

We also ask how this added visual information modifies speech processing in auditory cortical regions. The previous studies here provide an inconsistent image: the MEG study of Luo et al. (2010) found no increase in oscillatory gain with speechreading cues, which would be expected if a simple gain control mechanism were at work. However, a more recent MEG study (Zion Golumbic et al., 2013b) estimating impulse responses to ongoing speech stimuli (similar to our work) in a selective attention paradigm, found a gain increase in the impulse responses of attended audio–visual speech, but no latency effect. This contradicts our expectations from the monosyllabic token-evoked response study of van Wassenhove et al. (2005), which found a decrease in latency and a decrease in gain during audio–visual speech processing. We thus analyzed latencies and gains of auditory cortex impulse responses in the synchronized (audio–visual processing) and asynchronized (audio-only processing) and found no significant difference in impulse response gains, but a significant decrease in latency with audio–visual speech. These differences (from the results of Zion Golumbic et al. (2013b)) may be partially due to differences in EEG and MEG, particularly since EEG mixes the signals from both auditory cortices, making it impossible to see any hemispheric differences that may arise.

4.5 Conclusions

In this chapter, we demonstrated the benefit of the linear dynamical system modeling framework introduced in the previous chapter, by applying it to a study of audio–visual speech processing. By gathering surrogate data, we are able to restrict our model to our regions of interest without placing additional burdens on noisy EEG data. Using this, we were able to directly demonstrate causal information transfer of information from visual processing regions to auditory processing regions. We propose that this information
transfer corresponds to the transmission of visual articulatory cues to auditory regions in cross-modal speech processing.

We also found evidence of faster processing of auditory speech information in the presence of corresponding speechreading cues, but no evidence of more robust coding in the form of increased gain. This suggests that the robust coding of audio–visual speech (which is well established by psychophysical studies) may be due to reduction in uncertainty in speech processing, i.e. decreased variability in the impulse response. As we treat the impulse response as a fixed deterministic quantity in our model, we cannot directly test this hypothesis and suggests a Bayesian extension to the model, to be able to learn the variability in the model parameters themselves.
Chapter 5

Conclusions

In the previous three chapters, we developed statistical and computational tools and used these tools to probe the neural mechanisms for attentive processing of speech. In Ch. 2, we discussed issues in inferring oscillatory synchrony and ROI based functional connectivity networks with MEG data and inverse imaging. We found a central role for right frontal eye fields (FEFs) in the beta band network as listeners were preparing to attend to sounds. This supports the numerous studies stating the importance of beta band activity in the FEFs, and also the recent studies showing the special role of beta band activity in the right FEF during the preparatory period (Chanes et al., 2013; Valero-Cabré et al., 2013). We also found a central role for the inferior pre-central sulcus (PCS) in the alpha band network during stimulus presentation. Due to lack of a substantial literature of the functional role and connectivity properties of this region, we hope to see further studies to confirm and further probe our findings.

The study of functional networks, describing connectivity in addition to activation is gaining traction in the neuroscience literature. However much of the recent work on functional networks focuses on broader topological properties of the networks (Bolaños et al., 2013), particularly the small world property, at the expense of looking at specific regions and their connectivity structures. In a sense, this discards much of the first-order information we know about the roles of various cortical regions. We stress the importance of jointly looking at first and second-order properties of functional networks, similar to our work here and the studies reviewed in Palva & Palva (2011), to gain a fuller mechanistic perspective on the computational roles of physiological neural networks.
While trial-based experiments allow us to get better estimates of neural responses to specific tokens by averaging over multiple trials, they do not simulate the ecologically realistic scenario of continuous ongoing speech streams. To address this, we proposed a linear dynamical system model in Ch. 3 that characterizes neural dynamics to attending speech stimuli in this ongoing setting. This model is designed to account for first and second order effects as described above. The parameters describing the first-order effects capture the impulse responses to speech, and neural noise spectra of sources separately. The second-order effects are described by functional connectivity terms in the multivariate autoregressive formulation. The model also includes a component to specify the neural substrates, which may be inferred from the data, or defined by the modeler with functional localizers, or anatomical constraints.

In Ch. 4, we used this model to study the neural dynamics of audio–visual speech processing. Specifically, we demonstrated that the presence of informative visual cues speeds up attentive speech processing in auditory cortex, relative to audio-only speech. We also demonstrated an increase in causal influence from visual to auditory cortex in the presence of visual cues. This is the generally accepted mechanistic view of audio–visual speech processing: visual areas process articulatory information and prime the auditory system to expect certain phonemes over others. We directly observed evidence for this causal flow of information for the first time.

5.1 Potential Applications

In addition to the scientific applications of our statistical tools and models in interpreting experimental data, our models and results may have potential implications in designs for the next generation of intelligent hearing prostheses. Hearing-impaired listeners are known to have trouble in attending a specific speaker in a cocktail party setting (Shinn-Cunningham & Best, 2008). But even listeners with normal hearing thresholds may display large inter-subject variability in attention based tasks (Ruggles et al., 2011); one of the
major explanatory factors in this variability is age (Ruggles et al., 2012).

Thus, we envision an EEG-based controller based on the model developed in Chapters 3 and 4 to automatically detect which of multiple sound sources is being attended by a listener and artificially amplify it to enable easier listening. Such a system has already been proposed with a visually guided controller by Kidd et al. (2013). This would require sound source segregation solutions in addition to the EEG model. Since the model is expressed in the state-space formulation, it is computationally feasible to perform online estimation using a Kalman filter on an embedded device.

In addition to possible uses in hearing prostheses, such a model may be used to identify pathological speech processing. Subjects with otherwise intact hearing, who have difficulties in solving the cocktail party problem, such as subjects with traumatic brain injuries may benefit (Turkstra & Kennedy, 2008) from biofeedback-based attention rehabilitative systems, potentially using our model.


Kong, Lingqiang, Michalka, Samantha W., Rosen, Maya L., Sheremata, Summer L., Swisher, Jascha D., Shinn-Cunningham, Barbara G., & Somers, David C. 2012. Auditory Spatial Attention Representations in the Human Cerebral Cortex. *Cerebral Cortex*.


Ruggles, Dorea, Bharadwaj, Hari, & Shinn-Cunningham, Barbara G. 2011. Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. Proceedings of the National Academy of Sciences of the United States of America, Aug.


Weisz, Nathan, Müller, Nadia, Jatzev, Sabine, & Bertrand, Olivier. 2013. Oscillatory Alpha Modulations in Right Auditory Regions Reflect the Validity of Acoustic Cues in an Auditory Spatial Attention Task. *Cerebral Cortex,* May, bht113–.


Curriculum Vitae

Contact

Siddharth Rajaram
Center for Computational Neuroscience and Neural Technology,
Boston University, 677 Beacon Street, Boston, MA 02215

e-mail: srajaram@bu.edu

Year of Birth

1984

Education

Boston University, Ph.D. candidate, September 2007 – present.

University of Waterloo, M.Math., Combinatorics & Optimization,
Faculty of Mathematics, September 2006 – August 2007.


Research Experience

Auditory Neuroscience Laboratory, Boston University

Graduate Research Assistant, May 2008 – present.

Faculty of Mathematics University of Waterloo


Teaching Experience

Dept. of Cognitive and Neural Systems, Boston University

Graduate Teaching Fellow, Quantitative Studies of the Auditory System,
Fall 2008, Fall 2009.

Faculty of Mathematics University of Waterloo

Teaching Assistant, Calculus I for Engineers, Fall 2006.


Teaching Assistant, Calculus II for Engineers, Spring 2007.

Publications


Conference Papers


Conference Abstracts


Awards
**Travel Award Recipient and Finalist in the Student Paper Competition** at the 6th International IEEE/EMBS Conference on Neural Engineering in San Diego, CA (2013).

**Second Prize in the Student Paper Competition** at the 8th International Conference on Bioelectromagnetism in Banff, Canada (2011).

**IBRO Fellowship** to attend the Neural Systems & Behavior Course at the Marine Biological Laboratory at Woods Hole, MA (2010).

**Travel Award** to attend the Inverse Problems in Imaging Workshop in Montreal, Canada (2009).

**Boston University Arts and Sciences Dean’s Fellowship** (2007 – 2008).