UNIVERSITY OF CALIFORNIA,
IRVINE

Computational Modeling of Speech Production and Aphasia

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Psychology

by

Grant M. Walker

Dissertation Committee:
Professor Gregory Hickok, Chair
Professor William Batchelder
Professor Mark Steyvers

2016

DEDICATION

"*potamoisi toisin autoisin embainousin hetera kai hetera hudata epirrei.*

On those stepping into rivers staying the same other and other waters flow. (Cleanthes from Arius Didymus from Eusebius)

. . .

If this interpretation is right, the message of the one river fragment . . . is not that all things are changing so that we cannot encounter them twice, but something much more subtle and profound. It is that some things stay the same only by changing. One kind of long-lasting material reality exists by virtue of constant turnover in its constituent matter. Here constancy and change are not opposed but inextricably connected. A human body could be understood in precisely the same way, as living and continuing by virtue of constant metabolism–as Aristotle for instance later understood it. On this reading, Heraclitus believes in flux, but not as destructive of constancy; rather it is, paradoxically, a necessary condition of constancy, at least in some cases (and arguably in all). In general, at least in some exemplary cases, high-level structures supervene on low-level material flux. The Platonic reading still has advocates (e.g. Tarán 1999), but it is no longer the only reading of Heraclitus advocated by scholars."

- Heraclitus, Stanford Encyclopedia of Philosophy
(http://plato.stanford.edu/entries/heraclitus/)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I thank my advisor, Dr. Gregory Hickok, for being extremely patient and reasonable, and for offering invaluable guidance and opportunities. I also thank my dissertation committee members, Dr. William Batchelder and Dr. Mark Steyvers, for their feedback which greatly improved this work.

Working with aphasic patients has been a privilege and an honor, and this dissertation would not have been possible without the many people who provide care and conduct clinical research. I thank my mentors and co-authors from my previous work at the University of Pennsylvania and the Moss Rehabilitation Research Institute in Philadelphia, including Dr. Myrna Schwartz, Dr. H. Branch Coslett, and Dr. Daniel Mirman. I thank Dr. Julius Fridriksson and Dr. Chris Rorden of the University of South Carolina for providing access to their amazingly well-curated data.

I thank Springer for permission to include Chapters Two and Three of my dissertation, which were originally published in Psychonomic Bulletin & Review. The co-author listed in these publications directed and supervised research which forms the basis for the dissertation.

Financial support was provided by the University of California, Irvine and the National Science Foundation Graduate Research Fellowship under Grant Number 8108915.

Emotional support and other discretionary financial support was provided by my family. My grandmother, Carolyn Ziffrin, has always encouraged me to be curious and has provided the means to explore the world. My mother and father, Judith Walker and Dr. James Walker, are great parents; there is no better compliment, though they are deserving of many. I especially thank my brother, Dr. Aaron Walker, and my sister-in-law, Dr. Yi-Hsien Walker, for encouraging me to pursue graduate studies in Southern California, and then making me feel at home in a new and wonderful place. This dissertation mostly exists because they thought it was a good idea for me to write one. Finally, I thank the American Society for the Prevention of Cruelty to Animals and my adopted dog, Maverick, for providing the perspective that I needed to tackle difficult problems.

CURRICULUM VITAE

Grant M. Walker

2006	B.A. in Cognitive Sciences and Communications, University of Pennsylvania
2006-08	Research Assistant, Language and Aphasia Lab, Moss Rehabilitation Research Institute
2008-09	Research Assistant Supervisor, Language and Aphasia Lab, Moss Rehabilitation Research Institute
2009-11	Clinical Research Coordinator, Laboratory for Cognition and Neural Stimulation, University of Pennsylvania School of Medicine
2011-12	Teaching Assistant, School of Social Sciences, University of California, Irvine
2012-15	Graduate Research Fellowship Program, National Science Foundation, University of California, Irvine
2014	M.S. in Cognitive Neuroscience, University of California, Irvine
2015-16	Teaching Assistant, School of Social Sciences, University of California, Irvine
2016	Ph.D. in Psychology, University of California, Irvine

FIELD OF STUDY

Computational Modeling, Speech Production, Aphasia

PUBLICATIONS

Walker, G. M., & Hickok G. (2015). Evaluating conceptual and quantitative models of speech production: How does SLAM fare? *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-015-0962-9

Walker, G. M., & Hickok G. (2015). Empirical and computational evidence converge in support of the Hierarchical State Feedback Control theory. *Language, Cognition and Neuroscience*. doi: 10.1080/23273798.2015.1096404

Walker, G. M., & Hickok G. (2015). Bridging computational approaches to speech production: The semantic-lexical-auditory-motor model (SLAM). *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-015-0903-7

Walker, G. M. & Schwartz, M. F. (2012). Short-form Philadelphia Naming Test: Rationale and empirical evaluation. *American Journal of Speech-Language Pathology, 21*(2), S140. doi: 10.1044/1058-0360(2012/11-0089)

Walker, G. M., Schwartz, M. F., Kimberg, D. Y., Faseyitan, O., Brecher, A., Dell, G. S, & Coslett, H. B. (2011). Support for anterior temporal involvement in semantic error production in aphasia: New evidence from VLSM. *Brain and Language, 117,* 110-122. doi: 10.1016/j.bandl.2010.09.008

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O., Dell, G. S., Mirman, D., & Coslett, H. B. (2011). A neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *Proceedings of the National Academy of Sciences, 108*(20), 8520-8524*.* doi: 10.1073/pnas.1014935108

Mirman D., Walker G. M., Graziano K. M. (2011). A tale of two semantic systems: taxonomic and thematic knowledge. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, editors Carlson L., Hoelscher C., Shipley T. F. (Austin, TX: Cognitive Science Society), 2211–2216.

Mirman, D., Strauss, T. J., Brecher, A., Walker, G. M., Sobel, P., Dell, G. S., & Schwartz, M. F. (2010). A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive Neuropsychology, 26*(6), 495-504. doi: 10.1080/02643294.2011.574112

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Faseyitan, O., Brecher, A., Dell, G. S., & Coslett, H. B. (2009). Anterior temporal involvement in semantic word retrieval: voxel-based lesion-symptom mapping evidence from aphasia. *Brain, 132*, 3411-3427. doi: 10.1093/brain/awp284

ABSTRACT OF THE DISSERTATION

Computational Modeling of Speech Production and Aphasia

By

Grant M. Walker

Doctor of Philosophy in Psychology

University of California, Irvine, 2016

Professor Gregory Hickok, Chair

We investigated a new computational model of speech production, the Semantic-Lexical-Auditory-Motor Model (SLAM), which was designed to test a critical assumption of the Hierarchical State Feedback Control theory (Hickok, 2012): specifically, that speech production relies on the coordination of dual representations in auditory and motor cortices, with auditory representations serving as targets during speech planning. The computational details are based on the interactive, two-step lexical retrieval model of Foygel and Dell (2000); our novel architecture allows us to predict the consequences of different patterns of damage among the proposed speech representations. The additional model structure is expected to better explain conduction aphasia in particular. We analyzed archived picture naming data from 255 people with aphasia in Philadelphia, PA, in addition to new data from another 95 people with aphasia in Columbia, SC. We found that the SLAM model made adequate predictions generally, and it did improve the fit to data from conduction patients specifically and in the expected manner. We also analyzed neuroanatomical data in the form of lesion masks standardized to a template for 83 of the participants from the SC cohort. Although we were unable to replicate a study to localize

brain regions where damage leads to a significant increase in particular error types, a behavioral comparison of the different cohorts revealed the potential sampling variability that exists in the aphasia population. Next, we developed a Bayesian approach to estimating the parameters of the lexical network, providing a more comprehensive assessment of the model's quality. Additionally, we simulated word and nonword repetition tasks with our network, generating new predictions for a subset of 28 people with aphasia and unimpaired hearing. We found that the 4-parameter SLAM model could simultaneously find good fits for the frequencies of 6 naming and 3 nonword repetition response types while also correctly predicting a novel set of 6 word repetition response types. The conduction patients' data was best explained by strong lexical-auditory and weak auditory-motor connections. Our results demonstrate that the assumption of coordinated speech representations in auditory and motor cortices can lead to viable predictions of speech production behavior in aphasia.

Introduction

Aphasia is the medical term for an acquired language impairment due to neurological injury such as stroke. It is estimated that there are approximately 1 million people with stroke-induced aphasia in the United States, with about 83,000 new cases each year, and a prevalence in the developed world of about 0.1-0.4% (Code & Petharam, 2011). Clinical interventions primarily involve speech therapy, but, given the heterogeneity of the condition, it has been notoriously difficult to predict the efficacy of any particular type of aphasia treatment, for instance, using randomized controlled trials (Kelly, Brady, & Enderby, 2010). There is a considerable gap in the necessary knowledge about how the brain enables language processing, so that when the brain is damaged, clinicians are sometimes left without effective tools to diagnose and fix the language impairment. Over the past few decades, however, neuroimaging and cognitive measurement tools have continued to develop toward a point where they can be leveraged in the service of clinical intervention. The impetus for the present work is to develop a complete enough understanding of the relationships between neural systems and human language processing to design effective and reliable aphasia treatments, eventually including neural prosthetics to replace lost function. This ambitious goal will not be achieved in this dissertation; rather, we will settle for incremental progress, by attempting to identify and characterize important structure-function relationships that can aid clinicians and engineers in the future.

Indeed, we will maintain a rather limited focus on single word production. Our computational modeling approach is intended to test some critical assumptions about

speech production, by implementing these assumptions in a computer program and observing the consequences. We acknowledge from the outset that our models are wrong, insofar as they are descriptions of reality, but they still may be useful, particularly for clarifying the implications that follow from our theoretical positions. To that end, we try to adopt an inclusive perspective, bringing together different frameworks and methodologies to view these challenging problems in new ways.

References

Code, C., & Petheram, B. (2011). Delivering for aphasia. *International Journal of Speech-Language Pathology*, *13*(1), 3-10.

Kelly, H., Brady, M. C., & Enderby, P. (2010). Speech and language therapy for aphasia following stroke. Cochrane Database of Systematic Reviews, Issue 5. Art. No.: CD000425. DOI: 10.1002/14651858.CD000425.pub2.

CHAPTER 1: A Review of Computational Models of Speech Production and Aphasia

In this chapter, we review some of the computational modeling literature that inspired and guided our own efforts. We begin with preliminary comments about modeling speech and brains with artificial neural networks before discussing the literature on the interactive two-step lexical retrieval model developed by Dell and colleagues, which provides much of the theoretical and computational foundation for our model. We also review two other computational models that take a slightly different approach to understanding the effects of brain damage on speech production. At points, we intentionally belabor the details of the models, to convey the scope of what has been termed 'modeler degrees of freedom,' the sometimes very large number of choices that must be made to fully specify a model. A crucial part of the modeling process involves comparison of competing explanations, so identifying the similarities and differences between existing models can be an important first step in developing new models.

Speech and Brain Systems

Some preliminary comments about speech systems and brain systems are required to properly motivate the computational models that seek to explain their relationships. Speech and language are distinct but closely related systems. Language is the set of symbols and production rules that govern the generation and interpretation of continuous speech signals that are usually emitted from a vocal tract. Linguistic symbols and

production rules also govern other cognitive processes beyond speech, including reading, writing, signing, or self-guided attention. Speech signals have therefore typically been described at two broad, complementary levels of analysis: the lower-level, continuous, physical instantiations of speech (e.g., acoustic waveforms or vocal tract configurations) and the higher-level, hierarchically organized, symbolic sequences (e.g., phonemes, words, or phrases) that convey semantic messages. While it is clear that these levels must interface, the majority of research has been focused within one or the other domain (cf. Tzeng & Wang, 1984), and synthesis remains elusive. Another broad distinction can be drawn between the computational versus the psychological approaches to understanding speech production (Scharenborg, 2007). The former perspective is adopted in the field of natural language processing, which develops algorithms for computational devices to process speech signals in practical engineering applications, such as voice-recognition technology. On the other hand, the field of cognitive neuroscience attempts to discover the algorithms that brains use to process speech signals, so there is potential for crosstalk between these domains.

Neural systems, like speech systems, process continuous inputs and outputs in the service of hierarchically organized goals. Many simple processing units (neurons) are connected together to produce a network architecture that exhibits complex functional dynamics and provides a tight coupling between an organism's actions and perceptions. A distinction between levels of analysis exists within neuroscience as well: low-level analysis focuses on the processing units themselves and how they connect with neighbors (e.g., the dynamics of action potentials or neurotransmitter release) and high-level analysis focuses on the

network structures that influence behavior. Again, as in speech research, there is also a distinction in the motivations for understanding neural circuits: an engineering approach seeks to exploit the highly efficient parallel processing found in nervous systems by recreating their critical features in electronic circuits, while a neurobiological approach seeks to identify and manipulate the actual circuits found in brains.

Artificial Neural Networks

In humans, speech systems are instantiated in neural systems, so it is perhaps unsurprising that they share some organizational principles. A broad class of computational models known as artificial neural networks (ANN), inspired by neural connectivity, have been used to model both neural and linguistic structures. ANNs have many applications for pattern recognition and classification across a diversity of fields, from recognizing chromosomal abnormalities to suggesting a movie you might like (Cheng & Titterington, 1994). The vast literature regarding ANNs developed from the seminal work by McCulloch and Pitts (1943) on the Perceptron, and the standard, practical tutorial on contemporary ANN modeling was presented by Rumelhart, McClelland, and the PDP research group (1986). ANNs represent a powerful class of statistical model that can capture complex nonlinear relationships in noisy environments. However, speech and language models that are posed within an ANN framework have no guarantee of relating to real neural networks in any obvious way. While the computational models reviewed here all fall under the umbrella term of ANN, they vary with respect to their level of abstraction; therefore, an attempt is often made to highlight the connection between useful model components and brain systems.

ANNs consist of units (nodes, neurons) that can take a value (activation level), and directed connections between units (links, synapses) that can also take a value (weight, strength). An activation function determines the value of a unit based on the activation levels and connection weights of all of its sending units. Units are often categorized in terms of the modeler's access to them: input units have activation parameters that are directly set by the modeler, hidden units have activation parameters that depend only on other units in the model, and output units have activation parameters that are meant to be interpreted by the modeler. Connections may be unidirectional, bidirectional, or recurrent. They may also be excitatory or inhibitory. Sometimes weights are set by the modeler to instantiate a particular theoretical perspective on network organization. Alternatively, a learning rule may determine the weights of connections based on an optimization of model performance over training examples. The majority of ANNs developed for speech can be classified as rate-coding models (Dayan & Abbott, 2001). This type of ANN model assumes that a unit takes continuous values akin to a neuron's firing rate or a brain region's energy consumption. While an analogy can be made with units being neurons or gray matter regions and connections being synapses or white matter pathways, in practice, adherence to this analogy falls on a spectrum. Some modelers choose units to represent real neurons or brain regions with activation functions and learning rules that mimic neural behavior, whereas others choose units to represent mental representations and design their networks accordingly. In the case of a mental network, representations may be identified by a single unit (locally) or by a distributed pattern over multiple units. Even if a network model is designed around mental representations though, its components may still be related to neural data via an appropriate linking hypothesis.

The Interactive Two-step Lexical Retrieval Model

A spreading activation framework for sentence production was proposed by Dell (1986), and further developments focused on the lexical retrieval component implemented by the DSMSG model (Dell, Schwartz, Martin, Saffran, & Gagnon, 1997), designated by an acronym of the authors' names. The DSMSG model is motivated by facts about speech errors observed in healthy and aphasic speakers, and instantiates an interactive two-step retrieval process. The model has three layers of units that represent different levels of mental representation: semantic, lexical, and phonological. Because units represent mental constructs, the relationship between activation levels and neural data is tenuous and requires additional assumptions to interpret. Units have a linear activation function with decay that is perturbed by inherent and activation-based noise, and negative activation is prevented from spreading. Bidirectional, excitatory connections between layers are specified according to the appropriate linguistic relationships, and weights are sought that minimize the discrepancy between model outputs and results from real lexical retrieval processes.

The primary lexical retrieval task that the model simulates is picture naming. In the first step of the process, distributed semantic input representations are mapped onto hidden lexical representations which also resonate with their constituent output phonological representations. After a fixed number of time steps, the most activated lexical node receives a boost, followed by a further fixed interval of activation propagation. Finally, the most activated phonological units (coding for both identity and syllable position) are

selected for output. The output is categorized with respect to the target as: Correct (CAT→CAT), Semantic (CAT→DOG), Formal (CAT→HAT), Mixed (CAT→RAT), Unrelated (CAT→FOG), or Nonword (CAT→ROG). Errors are possible because the activation function includes noise and decay. After a number of time steps, the decay term has driven activation levels back down toward the stochastic resting level, and an alternative unit may be more active by chance at selection time. Errors that occur during the first selection step are inherently lexical in nature (Semantic, Formal, Mixed, or Unrelated), whereas errors in the second step are typically Nonword or Formal. Although the simulated lexical neighborhood is quite small, it captures many of the statistical properties of error opportunities in English.

The model accounts for some important facts about speech errors in healthy speakers. It can produce highly accurate naming, and when errors occur, they tend to be semantically related. Consistent with empirical observations, pure formal errors do not occur when the model is parameterized to match healthy speakers, supporting the separation of stages in the model (Dell et al., 1997). The influence of phonology on retrieval is detectable through the mixed-error effect however, which manifests as more formally related semantic errors than would be expected by chance. The model explains this effect as a result of the bottom-up links that allow a semantic competitor to receive additional activation through shared phonology with the target. The model also accounts for the lexical-error effect, in which speech errors are more likely to result in words than non-words. If incorrect phonemes that form a word become activated, they will receive additional support through resonance with the corresponding lexical node, whereas incorrect phonemes that do not form a word

will not receive activation from the lexical layer. While the model was initially developed to account for errors in healthy speakers, it has also been used to interpret speech produced by aphasic patients.

The DSMSG model assumed that damage affected the entire network as a whole, but further modeling studies suggested that localized damage may be a better approach. Foygel and Dell (2000) designed a new version of the model to capture patterns in data from aphasic speakers: the model's maximum weight values are set to match healthy speakers' performance, and then the lexical-semantic connections (s-weight) and/or the lexical-phonological connections (p-weight) are reduced (lesioned) to alter the model's output. We call this the SP model, referring to the semantic and phonological weight parameters. To match a particular patient's data with the model, many simulations are run at different parameter settings, and the one that produces the most similar data is selected as the best match (Dell et al., 2004; Foygel & Dell, 2000). This approach was able to account for 94.5% of the variance in data from an unselected group of 94 chronic aphasic speakers (Schwartz et al., 2006). Much of the model's success is attributed to its instantiation of the continuity hypothesis, which posits that aphasic deficits lie on a continuum between healthy speech and random linguistic behavior. The two-stages of lexical retrieval are also important for explaining the observed error distributions. The mixed-error effect is absent in some patients with aphasia, and the model is able to account for this through a reduction of the bottom-up connections that generate the effect in healthy speakers. The fact that Formal errors have two sources in the model, occurring at either lexical or phonological selection, receives support from the grammatical category constraint on lexical errors. That is, if a

Formal error occurs at the lexical selection stage, it will be a noun, whereas Formal errors that occur at the phonological selection stage needn't adhere to this constraint. The model fails to explain a small proportion of the sample, and these failures are attributed to unrealistic simplifying assumptions, in particular, to cognitive components not represented in the model, e.g., visual-semantic or phonological-articulatory processes.

Given the success of the SP model in the picture naming domain, it was further developed to simulate another task with a bearing on word production: word repetition. Dell et al. (1997) assumed there exist dissociable networks for word form production and recognition, based on the general lack of correlation between these deficits in aphasia. Thus, they predicted that word repetition would essentially consist of the second step of naming, that is, mapping lexical to phonological representations, assuming auditory representations are intact and can be accurately mapped to the lexical layer. This assumption enables simulation of word repetition without introducing any new network parameters, in particular, auditory representations. They found preliminary support for this approach by simulating 10 out of 11 unselected aphasic patients' repetition performance with satisfactory accuracy, using the second step of the parameterized DSMSG naming model. Other researchers, however, proposed that auditory representations might activate phonological output directly, bypassing the lexical layer, and therefore repetition could be better explained by a dual-route model (Hanley et al., 2002; Hillis & Caramazza, 1991). A single node was therefore added to the network representing auditory input that connects directly to the phonological output units (Figure 1.1). After fitting the naming

model parameters, the non-lexical connection weights (nl-weight) are estimated using a non-word repetition task.

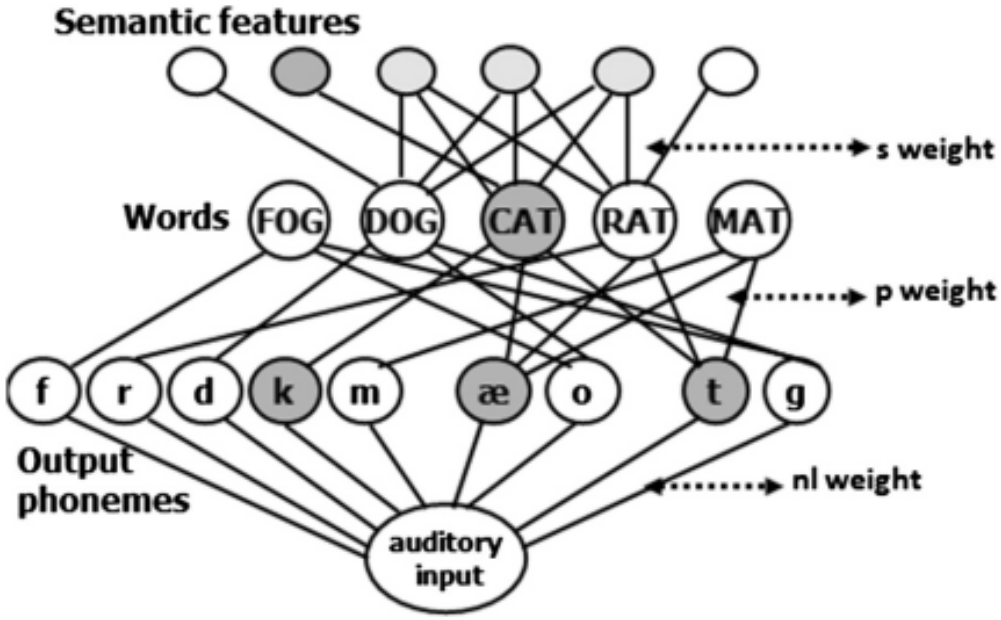

Figure 1.1 The architecture of the dual-route interactive two-step model (Dell et al., 2013).

The dual-route model has received some empirical support, but it comes with some caveats. Overall, the data so far suggest that a non-lexical route plays a role in word repetition, but it remains unclear exactly when this route becomes important for predicting aphasic word repetition scores. A basic problem for the single lexical-route model is that it necessarily predicts a strong relationship between naming and repetition ability, but some patients retain repetition skills in the presence of a strong naming deficit (Hanley et al., 2002). Hanley et al. (2004) demonstrated an advantage for the dual-route over the lexical-route model in predicting responses from 2 patients who both had poor naming but different repetition abilities; but, Baron et al. (2008) reported 6 more patients that exhibit the opposite modeling advantage, supporting the lexical-route model. When Dell et al. (2007) conducted a similar comparison on an unselected group of 30 aphasic patients, they

found an advantage for the dual-route model in only 4 patients (13% of the sample). Abel et al. (2009) adapted the tests for a set of German aphasic patients, and found a marginally significant advantage for the dual-route over the lexical-route when comparing RMSD values ($p=.082$). The above results were summarized by Nozari et al. (2010): taken together, models were directly compared for 53 patients, and only 15 showed a strong advantage for one model over the other, with 9 supporting the dual-route and 6 supporting the single lexical-route. Based on an examination of frequency effects in naming and repetition for 59 patients, Nozari et al. (2010) similarly find that the dual-route and lexical-route models are largely indistinguishable, but argue for the use of a dual-route model, because it can capture the interaction between non-word repetition ability and non-word error rates in word repetition versus picture naming. In the dual-route model, the non-lexical route is only used in repetition and not naming, so an improvement in non-word repetition will confer a benefit depending on the task. Additionally, a model with only a single lexical-route has no way of simulating non-word repetition. The authors conclude that the existence of a non-lexical mapping from auditory to motor representations should not be in dispute, but how and when individuals recruit this mapping for different tasks requires further clarification.

Recall that neural damage is simulated in the model by reducing the connection weights between representational layers. The model's account of neural damage, therefore, is that it specifically impairs the flow of information between levels of linguistic representation, while remaining silent on how the representations themselves are instantiated in neural systems. If a relationship can be found between damage to a particular neural region and a

model parameter, then it can be inferred that the region is critically important for mediating between certain types of representations. It is important to recognize, however, that in natural speech processing there are many simultaneously interacting levels of linguistic representation, and this model is only a partial implementation of a larger theoretical framework. Consequently, these simplifications may introduce unintended correlations between a model parameter and other cognitive components that are not explicitly modeled.

Dell et al. (2013) used a technique called voxel-based lesion parameter mapping to investigate the association between model parameters and localized brain damage. The technique was adapted from its original formulation by Bates et al. (2003). First, model parameter fits were obtained for 103 aphasic patients, all with left-hemisphere stroke, but unselected for any other criteria besides correctly naming at least one of 175 pictures. A measure of model fit quality was reported for picture naming (average RMSD=.023), but not for the dual-route model fit of word repetition. The s-weights were found to be uncorrelated with p-weights (r=.08), while p-weights were moderately correlated with nl-weights (r=.46). The p-weights and nl-weights, however, both made independent, significant contributions to a multiple regression model predicting word repetition accuracy. Regarding ancillary language tests, p-weights were correlated with degree of apraxia but not auditory discrimination, while nl-weights showed the reverse correlation pattern. The p-weights (estimated first, from picture naming) and nl-weights (estimated second, from non-word repetition) thus appear to have a shared component, while also

having separate contributions from articulatory-motor and auditory-phonological processes, respectively.

High-resolution structural brain scans (MRI or CT) were collected to localize regions of neural damage. Lesions were segmented and registered to a common template, so that in each voxel there were sets of patients with and without damage to the voxel. In each voxel then, these sets were compared for a difference in mean parameter values with a t-test, and a False Discovery Rate threshold (q=.05) was set to identify voxels in which damage predicts a significant decrement in the parameter of interest. The s-weight regions almost totally dissociated from the p-weight and nl-weight regions (Phi=.02), while the latter regions had a moderate degree of overlap (Phi=.34). Regions that were identified with the s-weight parameter included anterior temporal lobe, prefrontal cortex, and temporal-parietal-occipital junction. The p-weight localized to the supramarginal gyrus, pre- and post-central gyri, and insula; the nl-weight highlighted superior temporal gyrus and area Spt, overlapping with p-weight regions in SMG and post-central gyrus (Figure 1.2).

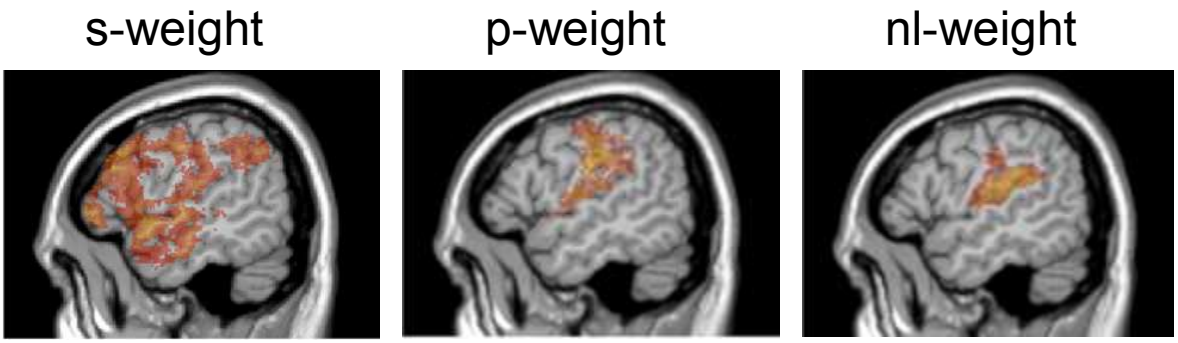

Figure 1.2. Lesion sites associated with a decrement in a model parameter, as identified by voxel-based lesion parameter mapping in 103 aphasic patients. The color scale represents significant t-values from low/red to high/yellow (Dell et al., 2013).

In light of the behavioral and neurological evidence, the authors suggested a reinterpretation of the model parameters: The s-weight was expanded to incorporate executive-attention mechanisms and semantic representations themselves, in addition to its originally conceived post-semantic lexical functions. The p-weight was extended to cover post-phonological articulatory mechanisms, in addition to phonological representations which it now shares with the nl-weight. This shared representation was hypothesized to be specifically phonological in nature, as opposed to representing larger syllabic units, leading to phonological slips when damaged rather than complete omissions.

Although the model is simplistic, it has provided a formal framework for interpreting speech error data, and relates a number of linguistic cognitive functions to neural structures. While we take the models of Dell and colleagues as a starting point for the modeling work presented in this dissertation, we turn now to examine two other ways of modeling word production for comparison.

The Lichtheim 2 Model

Ueno et al. (2011) presented a model named Lichtheim 2 that explicitly combines computational and neuroanatomical insights from speech production. The model uses hidden layers and backpropagation to learn network weights (Rumelhart, Durbin, Golden, & Chauvin, 1996), but it instantiates an architecture inspired by the dual routes identified in the neuropsychology literature (Hickok & Poeppel, 2004; Nozari et al., 2010). Layers of units represent specific brain regions, and connections represent the efferent and afferent white matter pathways that link those regions (Figure 1.3). The dorsal route is simulated as

a primary auditory ↔ inferior SMG ↔ insular-motor cortex pathway, underpinned by the arcuate fasciculus. It is assumed that this pathway is important for auditory-motor integration, and phonology in particular. The iSMG layer has recurrent connections, acting as a memory buffer. The ventral route is simulated as a primary auditory ↔ middle STG ↔ anterior STG ↔ triangularis-opercularis ↔ insular-motor cortex pathway, underpinned by the middle longitudinal fasciculus and the extreme capsule. This pathway plays a major role in extracting the meaning of words, via access to semantic representations in vATL. Although the ventral pathway does not have directly recurrent connections, aSTG performs a similar memory buffering function through its connections with vATL.

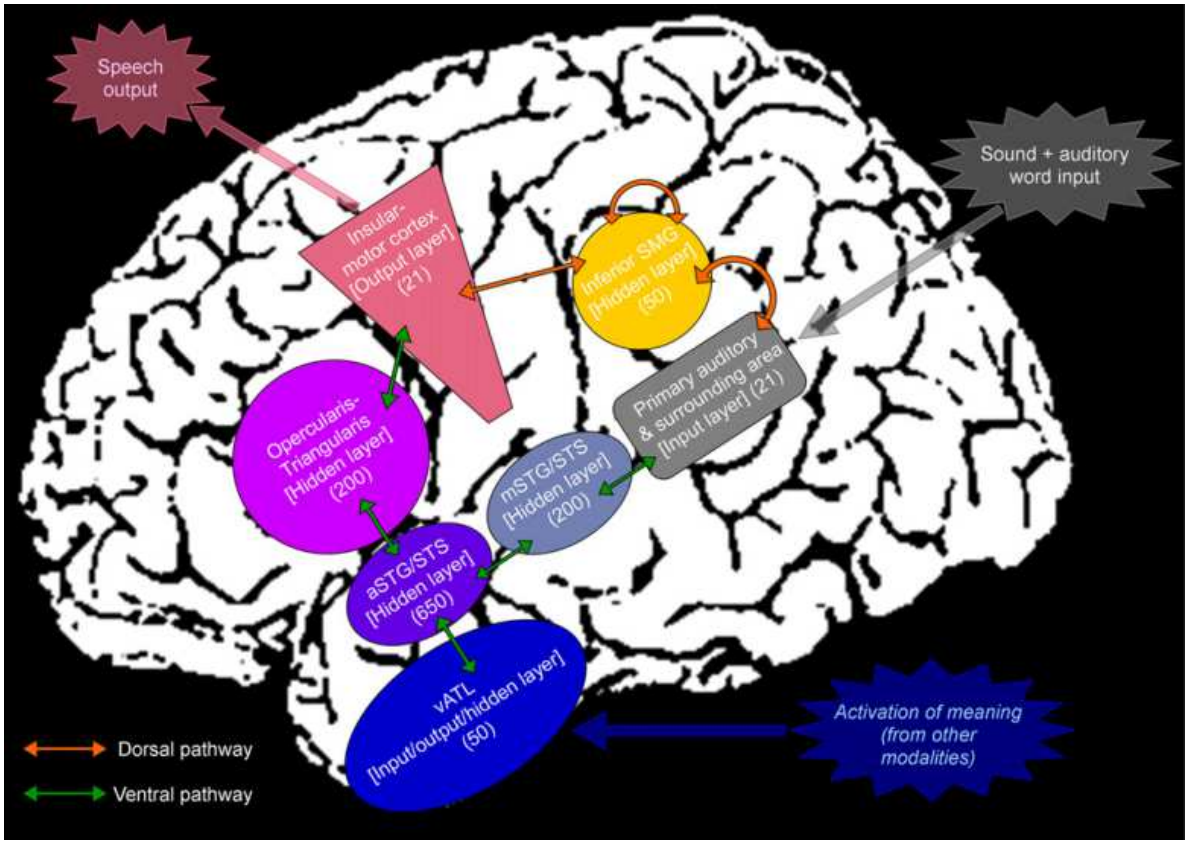

Figure 1.3. Brain localization of Lichtheim 2 components (Ueno et al., 2010).

The model was trained to perform repetition, comprehension, and naming of Japanese trimora words. Each mora had a phonetic representation distributed over 20 units, with one additional unit representing the pitch accents. The three moras were presented in succession over three time steps. The same auditory input representations were used for motor output representations. Semantic representations were defined so that words clustered into categories of overlapping distributed representations without any relation to auditory-motor representations. During comprehension, the input vectors were clamped to the auditory layer, and semantic targets were applied to the vATL layer. During naming, the developing semantic representation was clamped to vATL and targets applied to the insular-motor layer. During repetition, the auditory inputs were clamped while the motor output layer was required to be silent, followed by immediate recall of the trimora sequence. Output on a trial was scored correct if all units in the output layer were on the correct side of 0.5. A zero-error radius of 0.1 was used, meaning that no error was backpropagated if the difference between the output and target was 0.1 or less. All units had a trainable bias unit, and this connection was initialized to -1 to avoid strong bias activation early in learning. All units had a sigmoid activation function, and all weights were updated after each trial. One epoch of training included 3 comprehensions, 2 repetitions, and 1 naming of each word, in random order, and training concluded after 200 epochs (2.05 million words). After training, it was found that the model's internal representations varied depending on location and task, with the dorsal-route layers exhibiting more phonological than semantic similarities in their representations, and the ventral-route layers generally showing the opposite pattern. The aSTG representations

displayed semantic similarities during comprehension and phonological similarities during naming.

Specific regions were individually lesioned to test the effects of damage on task performance. Because stroke usually damages both cortex and white matter pathways, lesions were simulated by randomly removing incoming connections and adding Gaussian noise to activations. Each region was lesioned 50 times at 15 different severity levels, and performance was averaged to provide stable estimates. The performance of the damaged models captured many of the qualitative aspects of patient performance. In particular, damage to aSTG leads to a preponderance of semantic errors in naming, as judged by the confusability of pairs of near semantic neighbors. Further aphasic speech behavior can be captured by allowing the model to recover, that is, retraining the network to optimize its performance with its reduced computational abilities. For instance, the ventral pathway may be able to compensate for repetition tasks, but only for words and not non-words. While Lichtheim 2 captures many of the qualitative features of impairment patterns that have been reported in the literature, it does not address the role of auditory feedback, the influence of processing rates, or the effects of attention, nor does it account for the naming error patterns of individual patients like the SP model.

The WEAVER++ Model

A far-ranging theory of speech production was presented by Levelt et al. (1999) to account for chronometric investigations of normal speech. The theory is computationally instantiated in a model called WEAVER++, an acronym for Word-form Encoding by

Activation and Verification. The primary tasks in chronometric studies of word production are interference paradigms, in which a semantically or phonologically related distractor is presented at various stimulus onset asynchronies with the target. If the subject is processing a relevant type of information when the distractor is presented, increased response latencies are expected. Alternatively, if the distractor is presented too early or too late, no effect, or even facilitation, will occur. During these tasks, overt production errors are rare, and the model is therefore designed to not make errors. Paradoxically, binding-by-timing models such as SP or Lichtheim 2 that parsimoniously capture error patterns, have difficulty accounting for the timing of lexical access, while binding-by-checking, used to account for timing in WEAVER++, requires additional assumptions to handle checking failures that lead to errors (Levelt et al., 1999). Although WEAVER++ uses spreading activation to retrieve information from its predefined networks, this model takes a rather different approach to the selection of units and sequencing of outputs than the models reviewed above.

The WEAVER++ model integrates a spreading-activation based network with a parallel object-oriented production system, merging ANNs with classical logic-based cognitive models. Activation levels are used to trigger local production rules that select units, and only those units propagate their activation on to the next processing level. Simulation of naming begins with selection of a lexical concept from the conceptual layer. Semantic concepts are represented by nodes and their relationships are represented by labeled links. In practice, this requires a highly specific semantic neighborhood, and the model often begins simulations with the lexical concept node already set as the goal for production.

Next, at the lemma layer, the model selects the syntactic representation corresponding to the lexical concept. Activation spreads from the lexical unit throughout the lemma layer according to a linear activation rule with decay, until the correct lemma unit reaches an activation threshold and is selected. Incorrect units that reach threshold will not be selected, due to the production rules that use the labels on incoming top-down connections to verify that the unit is part of the intended utterance. The activation of incorrect units, however, will compete with the target unit and delay selection. This competition is instantiated through the use of a mathematical decision rule based on the hazard rate (Luce, 1986). Essentially, the probability of a unit being selected at a given time step is based on its activation normalized with respect to the other units in the layer. After selecting a lemma, its corresponding morphemes are activated serially, which in turn activate their phonemes, with labeled connections used to verify sequential positions. Finally, a similar competitive selection occurs within the syllabary layer, triggering the production of the corresponding motor syllable program. The binding procedure ensures that phonological representations are built online in a left-to-right, suspend-resume fashion. There is no stochastic noise in the model, and after fitting a few parameters regarding the spreading rate and duration of basic events, the expected value of the selection time has a closed-form solution. The authors suggest that internal and external feedback loops exist via the comprehension system, and a fair amount of evidence is marshaled to support this claim (Levelt et al., 1999; Roelofs, 2003a). The instantiated computational models only proceed up to syllable selection however, so the proposed feedback routes are generally not implemented.

Many chronometric phenomena can be accounted for by the model. The model produces the relevant patterns of facilitation and inhibition mediated by place of overlap and relatedness that are observed during picture-word interference experiments, using both monosyllabic and disyllabic words (Roelofs, 2000; Roelofs, 1997). The model also exhibits implicit priming, such that advance knowledge of the first syllable in a disyllabic word produces facilitation, while information about the second syllable does not. Another well-studied interference paradigm is the Stroop task, in which participants are presented with a word in colored text, and they are instructed to either read the word or name the color. The word may be either a color word or a neutral word, and if it is a color word, it may be congruent or incongruent with the text color. The WEAVER++ model successfully simulated 16 classic data sets from a review by MacLeod (1991), including congruency, incongruency, semantic-gradient, time-course, multiple-task, and pathological effects (Roelofs, 2003b). This task requires a great deal of attentional control, and the WEAVER++ model associates executive control and attention with the production rule system, which can both gate the appropriate input channels and maintain goal states throughout production.

Linking model components to neural structures, Roelofs (2011) summarizes evidence supporting the position that anterior cingulate cortex (ACC) plays a regulatory role in lexical production. Roelofs and Hagoort (2002) used WEAVER++ to simulate activity in this region during Stroop task performance. During each time step that required input gating or goal maintenance, a unit representing ACC received the model's standard input value, and this unit's activation was converted to a hemodynamic response with a gamma function. Thus, the time required to resolve selection competition, which is regulated by the model's

production rules, is related to the neural activity of ACC. Using a similar approach, Roelofs (2003a) used the activations in word-form perception and word-form production networks to generate hemodynamic responses for ventral and dorsal Wernicke's area, respectively.

Summary

It should be clear from the models reviewed above that there has been considerable effort invested in describing the mechanisms supporting speech production generally and identifying the neural substrates that implement them. The WEAVER++ model has addressed a wide experimental literature regarding the timing of speech production, the Lichtheim 2 model has addressed general speech production patterns in aphasia and recovery effects, and the SP model has addressed many of the effects observed in speech error data. Although the details may vary, much agreement can be found regarding the levels of representation and brain regions involved. Multiple levels of representation must be coordinated for successful production, and brain lesions causing impairment at specific levels of the representational system lead to characteristic changes in speech production behavior. Our own modeling efforts share these same core principles.

References

Abel, S., Huber, W., & Dell, G. S. (2009). Connectionist diagnosis of lexical disorders in aphasia. *Aphasiology*, *23*(11), 1353-1378.

Baron, R., Hanley, J. R., Dell, G. S., & Kay, J. (2008). Testing single- and dual-route computational models of auditory repetition with new data from six aphasic patients. *Aphasiology*, *22*(1), 62-76.

Bates, E., Wilson, S. M., Saygin, A. P., Dick, F., Sereno, M. I., Knight, R. T., & Dronkers, N. F. (2003). Voxel-based lesion–symptom mapping. *Nature neuroscience*, *6*(5), 448-450.

Cheng, B., & Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical science*, 2-30.

Dayan, P., & Abbott, L. F. (2001). Theoretical neuroscience: Computational and mathematical modeling of neural systems. Cambridge, MA: The MIT Press.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, *93*(3), 283.

Dell, G. S., Lawler, E. N., Harris, H. D., & Gordon, J. K. (2004). Models of errors of omission in aphasic naming. *Cognitive Neuropsychology*, *21*(2-4), 125-145.

Dell, G. S., Martin, N., & Schwartz, M. F. (2007). A case-series test of the interactive two-step model of lexical access: Predicting word repetition from picture naming. *Journal of Memory and Language*, *56*(4), 490-520.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological review*, *104*(4), 801.

Dell, G. S., Schwartz, M. F., Nozari, N., Faseyitan, O., & Branch Coslett, H. (2013). Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition*, *128*(3), 380-396.

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, *43*(2), 182-216.

Hanley, J. R., Dell, G. S., Kay, J., & Baron, R. (2004). Evidence for the involvement of a nonlexical route in the repetition of familiar words: A comparison of single and dual route models of auditory repetition. *Cognitive Neuropsychology*, *21*(2-4), 147-158.

Hanley, J. R., Kay, J., & Edwards, M. (2002). Imageability effects, phonological errors, and the relationship between auditory repetition and picture naming: Implications for models of auditory repetition. *Cognitive Neuropsychology*, *19*(3), 193-206.

Hillis, A., & Caramazza, A. (1991). Mechanisms for accessing lexical representations for output: evidence from a category specific semantic deficit. *Brain and Language, 40*, 106–144.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1), 67-99.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in spoken word production. *Brain and Behavioral Sciences, 22*, 1-38.

Luce, R.D. (1986). Response Times: Their Role in Inferring Elementary Mental Organization. Oxford University Press, New York.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological bulletin*, *109*(2), 163.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115-133.

Nozari, N., Kittredge, A. K., Dell, G. S., & Schwartz, M. F. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of memory and language*, *63*(4), 541-559.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*(3), 249-284.

Roelofs, A. (2000). WEAVER++ and other computational models of lemma retrieval and word-form encoding. In L. Wheeldon (Ed.), *Aspects of language production*, (pp. 71-114). Philadelphia, PA: Psychology Press.

Roelofs, A. (2003a). Modeling the relation between the production and recognition of spoken word forms. *Phonetics and phonology in language comprehension and production: Differences and similarities*, 115-158.

Roelofs, A. (2003b). Goal-referenced selection of verbal action: modeling attentional control in the Stroop task. *Psychological review*, *110*(1), 88.

Roelofs, A. (2011). Modeling the attentional control of vocal utterances: From Wernicke to WEAVER++. In J. Guendouzi, F. Loncke, & M. J. Williams (Eds.), *The Handbook of Psycholinguistic and Cognitive Processes: Perspectives in Communication Disorders* (pp. 189-207). Hove, UK: Psychology Press.

Roelofs, A., & Hagoort, P. (2002). Control of language use: Cognitive modeling of the hemodynamics of Stroop task performance. *Cognitive Brain Research*,*15*(1), 85-97.

Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA: MIT Press.

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1996). Back-propagation: The basic theory. In P. Smolensky, M. C. Mozer, & D. E. Rumelhart (Eds.), *Mathematical perspectives on neural networks* (pp. 533–566). Hillsdale, NJ: Erlbaum.

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*,*49*(5), 336-347.

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and language*, *54*(2), 228-264.

Tzeng, O. J., & Wang, W. S. (1984). Search for a common neurocognitive mechanism for language and movements. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *246*(6), R904-R911.

CHAPTER 2: Bridging Computational Approaches to Speech Production: The Semantic-Lexical Auditory Motor Model (SLAM)

Speech production has been studied from several theoretical perspectives including psycholinguistic, motor control, and neuroscience, often with little interaction between the approaches. Recent work, however, has suggested that integration may be productive, particularly with respect to applying computational principles from motor control, such as the combined use of forward and inverse models, to higher-level linguistic processes (Hickok, 2012, 2014a, 2014b). Here we explore this possibility in more detail by modifying Foygel and Dell's (2000) highly successful, psycholinguistic, computational model of speech production, using a motor control inspired architecture, and assess whether the new model provides a better fit to data and in a theoretically interpretable way.

We first present the theoretical foundations for this work by (i) describing the motivations behind Foygel and Dell's (2000) Semantic-Phonological model (SP), (ii) briefly summarizing the motor control approach, (iii) highlighting some principles from our recent conceptual attempt to integrate the approaches, and (iv) describing our modification of SP using a fundamental principle from motor control theory to create our new Semantic-Lexical-Auditory-Motor model (SLAM). We then present the computational details of both the SP and SLAM models, along with simulations comparing SP with SLAM. To preview the outcome of these simulations, we find that SLAM outperforms SP, particularly with respect to a theoretically predictable subcategory of aphasic patients. We conclude with a discussion of how the new model relates to some other extant models of word production.

The Semantic-Phonological (SP) model

SP has its roots in Dell's (1986) theory of retrieval in sentence production, which was developed to account for speech errors, or slips of the tongue, found in large collections of natural speech. To this end, the theory integrated psychological and linguistic concepts: from psychology it adopted the notion of computational simultaneity in which multiple internal representations compete for selection prior to production, and from linguistics it incorporated the hierarchical levels of representation, as well as the separation between the stored lexical knowledge and the applied generative rules at each level.

Dell et al. (1997) proposed a computational model that limited the focus to single word production, but extended the theoretical scope to include explanations of speech errors in the context of aphasia. The basic idea was that the pattern of aphasic speech errors reflects the output of a damaged speech production system, which could be modeled by adjusting parameters in the normal model to fit aphasia data. The model's architecture consisted of a 3-layer network with semantic, lexical, and phonological units, and the connections among units were selected by the experimenters to approximate the structure of a typical lexical neighborhood (Figure 2.1). Word production was modeled as a spreading activation process, with noise and decay of activation over time. Damage was implemented by altering the parameters that control the flow of activation between representational levels. Simulations were then used to identify parameter values that generated similar frequencies of error types as those made by aphasic patients.

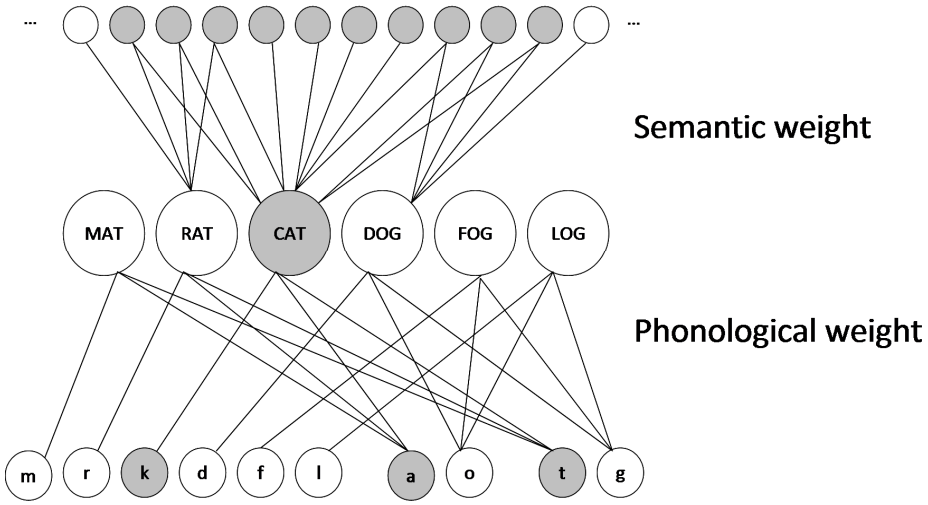

Figure 1.1. The SP model architecture.

Due to the computationally intensive nature of the simulation method, however, comprehensive explorations were effectively limited to only two parameters at a time. Nevertheless, in a series of papers beginning with Foygel and Dell (2000), two free parameters in the model were identified that account for an impressive variety of data derived from a picture naming task, including clinical diagnostic information (Abel et al., 2009), lexical frequency effects (Kittredge et al., 2008), characteristic error patterns associated with different types of aphasia (Schwartz et al., 2006), characteristic patterns of recovery (Schwartz & Brecher, 2000), and interactive error effects (Foygel & Dell, 2000). These two free parameters were the connection strengths between semantic and lexical representations (s-weight) and between lexical and phonological representations (p-weight), an architecture known as SP. SP has been used to explain performance on other tasks as well, such as word repetition (Dell et al., 2007), and to predict the location of neurological damage seen in clinical imaging (Dell et al., 2013), although here we will focus primarily on its relevancy to picture naming errors.

SP pertains specifically to computations that occur between semantic and phonological levels. It is assumed that the output of the model is a sequence of abstract phonemes that must then be converted into motor plans for controlling the vocal tract. We turn next to some fundamental constructs that have come out of research on how motor effectors are, in fact, controlled.

Motor control theory

At the broadest level, motor control requires sensory input to motor systems for initial planning and feedback control. It requires input for planning to define the targets of motor acts (e.g., a cup of a particular size and orientation and in a particular location relative to the body) and to provide information regarding the current state of the effectors (e.g., the position and velocity of the hand relative to the cup). Without sensory information, action is impossible, as natural (Cole & Sedgwick, 1992; Sanes, Mauritz, Evarts, Dalakas, & Chu, 1984) and experimental (Bossom, 1974) examples of sensory deafferentation have demonstrated. Sensory information has also been shown to provide critical *feedback* information during movement (Wolpert, 1997; Wolpert, Ghahramani, & Jordan, 1995), which provides a mechanism for error detection and correction (Kawato, 1999; Shadmehr, Smith, & Krakauer, 2010). When precise movements are performed rapidly, however, as in speech production, feedback mechanisms may be unreliable, due to feedback delay or a noisy environment. In this case, a state feedback control system can be supplemented with forward and inverse models (Jacobs, 1993), enabling the use of previously learned associations between motor commands and sensory consequences to guide the effectors

toward sensory goals. This arrangement implies that motor and sensory systems are tightly connected, even prior to online production or perception.

In the case of speech, the most critical sensory targets are auditory (Guenther, Hampson, & Johnson, 1998; Perkell, 2012), although somatosensory information also plays an important role (Tremblay, Shiller, & Ostry, 2003). Altered auditory feedback has been shown to dramatically affect speech production (Houde & Jordan, 1998; Larson, Burnett, Bauer, Kiran, & Hain, 2001; Yates, 1963) and changes in a talker's speech environment can lead to "gestural drift," that is, changes in his or her articulatory patterns (accent) (Sancier & Fowler, 1997). Additionally, neuroimaging experiments investigating covert speech production consistently report increased activation in auditory related cortices in the temporal lobe (Callan et al., 2006; Hickok & Buchsbaum, 2003; Okada & Hickok, 2006).

Some particularly relevant evidence for the role of the auditory system in speech production comes from neurpsychological investigations of language. Striking patterns of impaired and intact language processing abilities resulting from neurological injury have led theorists to propose separate auditory and motor speech representations in the brain (Caramazza, 1988; Jaquemot et al., 2007; Pulvermuller, 1996; Wernicke, 1874). Patients with conduction aphasia (Goodglass, 1992), for example, have fluent speech production, suggesting preserved motor representations. These patients also have good auditory comprehension and can recognize their own errors, suggesting spared auditory representations. Despite these abilities, they make many phonemic errors in production and have trouble with nonword repetition. This pattern is typically explained as resulting

from damage to the interface between the separate auditory and motor systems (Anderson et al., 1999; Geschwind, 1965; Hickok et al., 2000; Hickok, 2012). This point regarding conduction aphasia has important theoretical implications, as discussed below.

Conceptual integration

The hierarchical state feedback control (HSFC; Hickok, 2012) model provides a theoretical framework for the integration of psycholinguistic notions with concepts from biological motor control theory. This conceptual framework is organized around three central principles. The first principle is that speech representations have complimentary encodings in sensory and motor cortices that are activated in parallel during speech production, all the way up to the level of (at least) syllables. The second principle is that a particular pattern of excitatory and inhibitory connections between the sensory and motor cortices, mediated by a sensorimotor translation area, implements a type of forward/inverse model that can robustly guide motor representations toward sensory targets, despite the potential for errors in motor program selection during early stages of motor planning/activation. The third principle is that sensorimotor networks supporting speech production are hierarchically organized, with somatosensory cortex processing smaller units on the order of phonemes (or more accurately, phonetic-level targets such as bilabial closure, which can be coded as somatosensory states) and auditory cortex processing larger units on the order of syllables (i.e., acoustic targets). A schematic of the HSFC framework is presented in Figure 2.2; it is clear that the top portion (darker purple colors) embodies the two steps of SP, but breaks down the phonological component into two subcomponents, an auditory-phonological network and a motor-phonological network.

This conceptual overlap has inspired our creation of a new computational model that is directly related to the first principle, and partially related to the other two principles. We reason that the architectural assumptions of the HSFC model can be evaluated, in part, by integrating them with an established and successful computational model of naming, SP; if the architectural changes lead to improved modeling performance, this will provide support for the new framework.

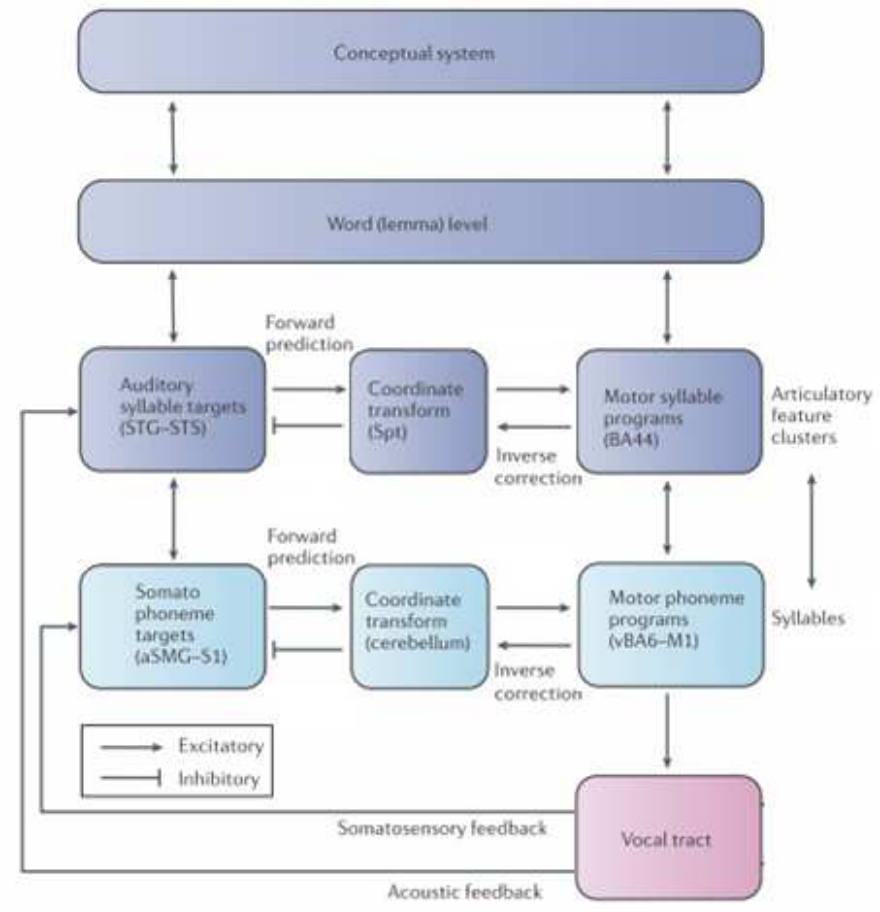

Figure 2.2. A schematic diagram of the hierarchical state feedback control (HSFC) framework (Hickok, 2012).

The SLAM model

SLAM is a computational model of lexical retrieval that divides phonological representations into auditory and motor components (Figure 2.3). The dual representation of phonemes directly follows from the first HSFC principle. The choice to label the sensory units as auditory representations is motivated by the third principle, specifically, that this level of coding is larger than the phonetic feature. Neither SP nor SLAM include inhibitory connections, and thus the second HSFC principle is not directly implemented; however, the pattern of connections in the SLAM model does implement a type of forward/inverse model that can reinforce potentially noisy motor commands. Our goal here was to modify the computational assumptions of SP as little as possible to assess the effects of the architectural assumption of separate motor and sensory phonological representations.

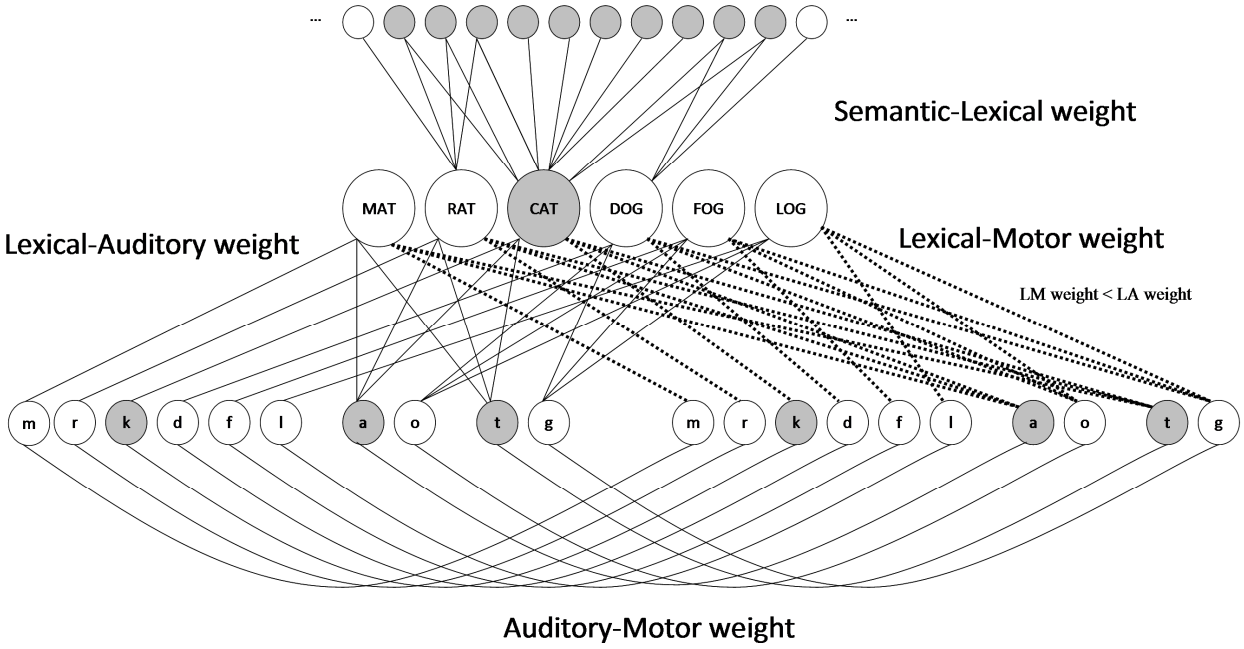

Figure 2.3. The SLAM model architecture.

During picture naming simulations, activation primarily flows from semantic to lexical to auditory to motor units, hence the model's acronym, SLAM. There is also a weaker, direct connection between lexical and motor units. The existence of this lexical-motor connection

acknowledges that speech production may occur via direct information flow from lexical to motor units, an assumption dating back to Wernicke (1874), needed to explain preserved fluency and spurts of error free speech in conduction aphasia. However, the connection is always weaker than the lexical-auditory route (again, Wernicke's original idea), motivated by several points. First, the auditory-lexical route is presumed to develop earlier and be used more frequently than the lexical-motor route. Longitudinal studies have shown that children begin to comprehend single words several months before they produce them, and they acquire newly comprehended words at nearly twice the rate of newly produced words (Benedict, 1979). Second, motor control theory dictates that motor plans are driven by their sensory targets. During development, the learner must make reference to auditory targets, in order to learn the mapping between speech sounds and motor gestures that reproduce those sounds (Hickok, Houde, & Rong, 2011; Hickok, 2012). Third, in the context of aphasia, comprehension deficits tend to recover more than production deficits (Lomas & Kertesz, 1978), suggesting a stronger association between lexical and auditory-phonological representations.

There is an important consequence of the assumption that the lexical-auditory mapping is always stronger than the lexical-motor mapping. It means that the SLAM model is not merely the SP model with an extra part; in fact, there is effectively zero overlap in the parameter space covered by SP and SLAM. The reason for this is as follows. Given the SLAM architecture shown in Figure 2.3, it is clear that one *could* implement SP simply by setting the connection weights in the lexical-auditory and auditory-motor mappings to zero, and letting the lexical-motor weights vary freely. This would make SP a proper subset of SLAM,

allowing SLAM to cover identical parameter space (and therefore fits to data) as SP. However, this architectural possibility was explicitly excluded by implementing our assumption that lexical-auditory weights are always stronger than lexical-motor weights:—if lexical-auditory weights are zero, then lexical-motor weights must also be zero and cannot vary freely—thus effectively excluding the parameter subspace used by SP. This further allows us to test SLAM's assumption that the lexical-auditory route is the primary one used in naming. We can examine model performance with the opposite constraint, namely, when lexical-auditory weights are always less than lexical-motor weights—a variant we might call "SLMA" to reflect the lexical-motor dominance—which includes SP parameter space as a subset. SLAM and SLMA have the same number of free parameters, both of which are more than SP, but with different assumptions regarding the connection strength patterns. If SLAM does better than SLMA, even though SLMA implements SP as a proper subset of its parameter space, it will demonstrate that the primacy of the lexical-auditory route is not only theoretically motivated, but also necessary for the observed improvements.

To summarize, we hypothesized that SLAM would characterize deficits in the general aphasia population at least as well as SP, and would primarily benefit the modeling of conduction aphasia. Recall that conduction aphasia is best explained as a dysfunction at the interface between auditory and motor speech representations that affects the phonological level in particular (Hickok et al., 2011; Hickok, 2012). Thus a naming model that incorporates a mapping between auditory- and motor-phonological representations should provide a better fit for speech errors resulting from dysfunction in this mapping. To

test this hypothesis, we compared the SP and SLAM model fits to a large set of aphasic picture naming data.

Computational Implementation

Patient data

All data were collected from the Moss Aphasia Psycholinguistic Project Database (Mirman et al., 2010; www.mappd.org). The database contains de-identified data from a large, representative group of aphasic patients, including responses on the Philadelphia Naming Test (PNT; Roach, Schwartz, Martin, Grewal, & Brecher, 1996), a set of 175 line drawings of common nouns. All patients in the database had post-acute aphasia subsequent to a left-hemisphere stroke, without any other diagnosed neurological comorbidities, and they were able to name at least one PNT item correctly. We analyzed the first PNT administration for all patients in the database with demographic information available, including aphasia type and months post-onset (N=255). The cohort consisted of 103 Anomic, 60 Broca's, 46 Conduction, 35 Wernicke's, and 11 others with transcortical sensory, transcortical motor, postcerebral artery, or global etiologies. The median months post-stroke was 28 [1, 381], and the median PNT percent correct was 76.4 [1, 99].

Computational models

As mentioned above, SP was first presented by Foygel and Dell (2000). The model's approach to simulating picture naming instantiates an interactive, two-step, spreading activation theory of lexical retrieval and consists of a 3-layer network, with individual units representing Semantic, Lexical, and Phonological symbols (Figure 2.1). The number of units

and the pattern of connections are intended to approximate the statistical probabilities of speech error types in English, by implementing the structure of a very small lexical neighborhood consisting of only 6 words, 1 of which is the target. There are 6 Lexical units with each connected to 10 Semantic units representing semantic features. Semantically related words share 3 Semantic units, meaning that on a typical trial, with only 1 word that is semantically related to the target, the network has a total of 57 Semantic units. Each Lexical unit is also connected to 3 Phonological units corresponding to an onset, vowel, and coda. There are 10 Phonological units total: 6 onsets, 2 vowels, and 2 codas. Words that are phonologically related to the target differ only by their onset unit, and the network always consists of 2 such words. Finally, the remaining 2 words in the network are unrelated to the target, with no shared Semantic or Phonological units. On 20% of the trials, one phonologically related word is also semantically related, creating a neighbor that has a "mixed" relation to the target.

Simulations of picture naming begin with a boost of activation delivered to the Semantic units. Two parameters, S and P, specify the bidirectional weights of Lexical-Semantic and Lexical-Phonological connections, respectively. Activation spreads simultaneously between all layers, in both directions, for eight time steps according to a linear activation rule with noise and decay. Then, a second boost of activation is delivered to the most active Lexical unit, and activation continues to spread for a further eight time steps. Finally, the most active Phonological onset, vowel, and coda units are selected as output to be compared with the target. Production errors occur due to the influence of noise as activation levels decay, which can be mitigated by strong connections. Responses are classified as Correct,

Semantic, Formal, Mixed, Unrelated, or Neologism. For a given parameter setting, a multinomial distribution over these six response types is estimated by generating many naming attempts with the model. These distributions may then be compared with those that result from naming responses produced by aphasic patients.

SLAM retains many of the details of SP, consistent with our aim to primarily assess the effects of the architectural modification. The Semantic and Lexical units remain unchanged, but there is an additional copy of the Phonological units, with one group designated as Auditory and the other as Motor (Figure 2.3). Four parameters specify the bidirectional weights of Semantic-Lexical (SL), Lexical-Auditory (LA), Lexical-Motor (LM), and Auditory-Motor (AM) connections. The LA and LM connections are identical to the P connections in the SP model, with each Lexical unit connecting to 3 Auditory and 3 Motor units, while AM connections are simply one-to-one. Simulations of picture naming are carried out in the same two-step fashion as with SP, with boosts delivered to the Semantic and then Lexical units, and phonological selection occurring within Motor units.

Fitting data

In order to fit data, the model is evaluated with different sets of parameters that yield sufficiently different output distributions, creating a finite-element map from parameter-space to data-space, and vice versa. This process involves, first, selecting a set of parameter values (e.g., S and P weights), then generating many naming attempts with the model using that parameter set, in order to estimate the frequency of each of the 6 types of responses that occur with that particular model setup. Once those frequencies have been determined,

that weight configuration becomes associated with the output distribution in a paired list called a map. Each point in the map represents a prediction about the type of error patterns that are possible when observing aphasic picture naming. One way to evaluate a model then, is to measure how close its predictions come to observed aphasic error patterns. The distance between an observed distribution and the model's nearest simulated distribution is referred to as the model's fit for that data point. The root mean square deviation (RMSD) is an arbitrary, but commonly used measure of fit, which can be interpreted as the average deviation for each response type. For example, an RMSD of .02 indicates that the observed proportions deviate from the predicted proportions by .02 on average (e.g., predicted = [.50, .50]; observed = [.48, .52]). Thus, a lower RMSD value indicates a better model fit. Immediately, the question arises of how many points one should generate, and how to select the parameters to avoid generating redundant predictions.

In their Appendix, Foygel and Dell (2000) provide guiding principles for generating a variable-resolution map of parameter-space, along with an example algorithm. They note that the particular choice of mapping algorithm likely has little impact on fit results, as long as it yields a comprehensive search; however, given the inherently high computational cost of mapping, a particular algorithm may affect the map's maximum resolution in practice. A second algorithm for parameter-space mapping is given by (Dell, Lawler, Harris, & Gordon, 2004), and these maps are considered to be the standard for SP, as they are available online and used in subsequent publications. This SP map has 3,782 points with 10,000 samples at each point, and required several days of serial computation to generate. Clearly, the computational cost associated with the mapping procedure represents a considerable

bottleneck for developing and testing models. Adding new points to the map improves the chances of a prediction lying closer to an observation, with diminishing returns as the model's set of novel predictions winnows. As Dell has suggested, because the goal is to find the best fit, adding more points to improve model performance is probably a worthy pursuit (G. Dell, personal communication, July 12, 2013). Moreover, because SLAM has two additional parameters, there was a need to modify the mapping procedure to generate maps more efficiently.

We greatly improved efficiency by redesigning the mapping algorithm to take advantage of its inherent parallelism. There are two main iterative steps in the mapping algorithm: point selection and point evaluation. The coordinates of a point in parameter-space are defined by a possible parameter setting for the model (point selection) and a corresponding point in data-space is defined by the proportions of response types generated with that parameter setting (point evaluation). The point evaluation step is extremely amenable to parallelization, because the simulations involve computations across independent units, independent samples, and independent parameter sets. Point selection, however, required a new approach to foster parallelism: Delaunay mesh refinement.

The Delaunay triangulation is a graph connecting a set of points such that the circumcircle of any simplex does not include any other points in the set. This graph has many favorable geometric properties, including the fact that edges provide adjacency relationships among the points. The new point selection algorithm takes advantage of these adjacency relationships. Beginning with the points lying at the parameter search range boundaries

and their centroid, if the separation between any two adjacent points in parameter-space exceeds a threshold distance (RMSD) in data-space, their parameter-space midpoint is selected for evaluation and added to the map. These new points are then added to the Delaunay mesh, and the process reiterates until all edges are under threshold. Thus, on each iteration, the point selection algorithm yields multiple points to be evaluated in parallel across the entire parameter search range. Parallel processing was executed on a GPU to further improve efficiency[1].

Before statistically comparing SP and SLAM's performances, we studied the effects of map resolution on model fits. First, we generated a very high resolution map for each model using a low RMSD threshold of 0.01 to encourage continued exploration of the parameter space. Each map included 10,000 samples at each point, and parameters varied independently in the range [0.0001, 0.04]. The maximum parameter values were selected to be near the lowest values that yield the highest frequency of correct responses, so that reduced values lead to more errors. Due to the use of a low mapping threshold, the algorithm was halted before completion, after generating an arbitrarily large number of points. Early termination is not a great concern because the algorithm efficiently selects points over the full search range. This fact also makes it a trivial matter to reduce the map resolution while still covering the full space.

[1] At the time of writing this manuscript, the authors were unaware of any freely available parallel algorithm to incrementally construct the Delaunay triangulation in arbitrary dimensions. We therefore implemented point evaluation and edge bisection using CUDA C and the Thrust library, executing these steps on a GPU, while the Delaunay triangulation was constructed on the CPU using the CGAL library. Performance tests comparing the parallel point evaluation step to a serial C++ implementation, running on an Nvidia Tesla K20Xm GPU and an Intel 1200 MHz 64-bit CPU, respectively, demonstrated a speedup by a factor of 26.0.

The mapping procedure generated an initial 31,593 points for the SLAM model, with parameters freely varying; then, in accordance with the SLAM architecture, all points with LM ≥ LA were removed, yielding a SLAM map with 17,786 points. The full SP map had 57,011 points. Next, we created 50 lower-resolution maps for each model by selecting subsets from the larger maps, with logarithmically spaced numbers of points from 5 to 17,000. For each map, we calculated the mean fit for the aphasic patients as a whole and for each of the diagnosis groups, excluding the heterogeneous diagnosis group. Figure 2.4 plots the fit curves. As expected for both models, adding points improves fits with diminishing returns. The relative fit patterns appear to stabilize around 2,321 points, marked by a vertical line in the figure. We therefore chose to compare SP and SLAM at this map resolution; our findings should apply to any higher resolution map comparisons, with trends favoring SLAM as resolution increases.

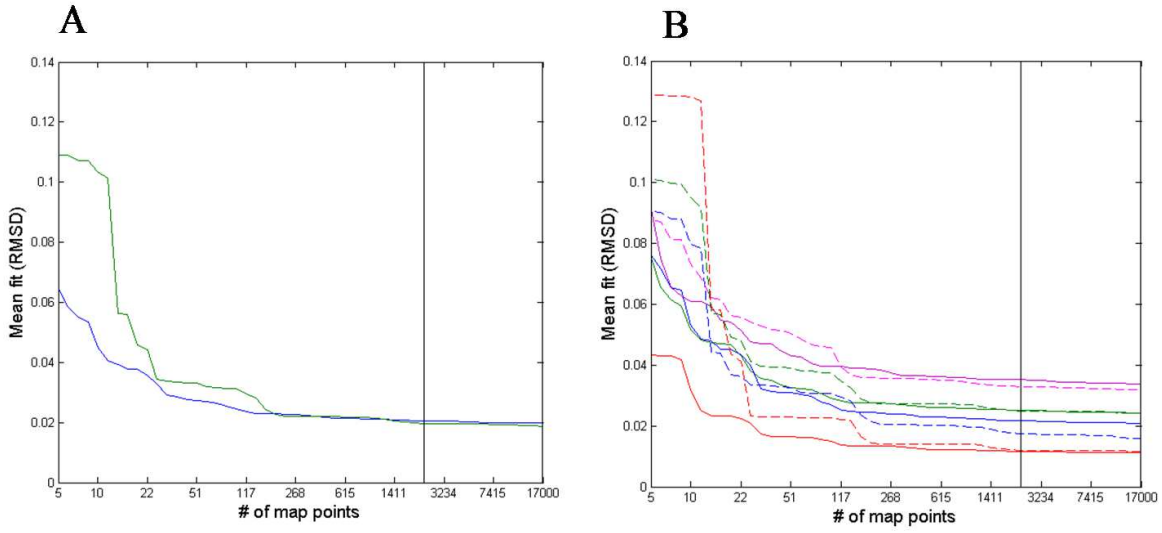

Figure 2.4. Mean fit curves for A) all patients (SP = blue, SLAM = green) and B) diagnosis groups (SP = solid, SLAM = dashed; Anomic = red, Broca's = green, Conduction = blue, Wernicke's = magenta). The black vertical line indicates the maps that were used for statistical comparisons.

To compare the new parallel generated maps with the standard serially generated maps, we also identified a parallel SP map resolution that yielded similar performance in terms of mean and maximum fit to the values reported in Schwartz et al. (2006). For this set of 94 patients, a parallel SP map with 189 points resulted in a mean and maximum RMSD of 0.0238 and 0.0785, compared with the reported values of 0.024 and 0.084, respectively. As expected, the parallel algorithm selects points much more efficiently than the serial algorithm, requiring many fewer predictions to achieve similar performance. We used this lower map resolution as a baseline, to compare the effects of adding points to the standard SP map with the effects of augmenting SP's structure. Because our fitting routine yields better fits than the standard SP maps that have been available to researchers online (Dell et al., 2004), we have provided our fitting routine, with adjustable map resolutions, along with our new model, at the following web address:
http://cogsci.uci.edu/~alns/webfit.html

Results

First, we examined our hypothesis that SLAM would fit data at least as well as SP for the general aphasia population. All analyses were performed using the MATLAB software package. As mentioned above, we chose to use RMSD as our measure of fit (lower value means better fit). Table 1 provides descriptive statistics of model fits for the entire sample of patients, as well as five subtypes of aphasia. Figure 2.5 shows a scatterplot comparing SP and SLAM fits. The solid diagonal line represents the hypothesis that the models are equivalent, and the dotted lines indicate one standard deviation of fit difference in the sample. It is clear that both models do quite well overall, with the majority of patients

clustering below .02 RMSD. While the models tend to produce similar fits in general, it is also clear that a subgroup of patients falls well outside the 1 SD boundaries. The inset in Figure 5 shows a bar graph comparing the number of patients that are better fit (> 1 SD) by SP or SLAM, demonstrating that SLAM provides better fits for a subgroup of patients without sacrificing fits in the general population.

Table 2.1. Descriptive statistics for SLAM and SP model fits.

		SP			SLAM				
		S	**P**	**RMSD**	**RMSD**	**SL**	**LM**	**LA**	**AM**
All N=255	Mean	.0241	.0232	.0194	.0187	.0242	.0152	.0300	.0254
	St. Dev.	.0118	.0075	.0189	.0189	.0121	.0059	.0085	.0135
	Median	.0256	.0224	.0127	.0111	.0250	.0151	.0300	.0300
	IQR	[.0170 , .0341]	[.0179 , .0281]	[.0071 , .0264]	[.0067 , .0246]	[.0176 , .0347]	[.0113 , .0200]	[.0225 , .0388]	[.0151 , .0388]
	Range	[.0001 , .0400]	[.0062 , .0400]	[.0010 , .1273]	[.0019 , .1977]	[.0001 , .0400]	[.0026 , .0300]	[.0010 , .0400]	[.0001 , .0400]
Anomic N=103	Mean	.0299	.0274	.0110	.0115	.0299	.0181	.0308	.0296
	St. Dev.	.0081	.0070	.0095	.0095	.0085	.0054	.0080	.0113
	Median	.0296	.0266	.0082	.0085	.0300	.0176	.0325	.0350
	IQR	[.0241 , .0376]	[.0219 , .0318]	[.0049 , .0137]	[.0056 , .0141]	[.0250 , .0400]	[.0151 , .0200]	[.0250 , .0400]	[.0204 , .0400]
	Range	[.0054 , .0400]	[.0106 , .0400]	[.0010 , .0654]	[.0019 , .0685]	[.0063 , .0400]	[.0038 , .0300]	[.0101 , .0400]	[.0001 , .0400]
Broca's N=60	Mean	.0215	.0218	.0238	.0240	.0217	.0143	.0267	.0266
	St. Dev.	.0125	.0071	.0243	.0250	.0128	.0053	.0088	.0144
	Median	.0205	.0202	.0145	.0149	.0200	.0126	.0250	.0325
	IQR	[.0139 , .0334]	[.0174 , .0262]	[.0076 , .0312]	[.0075 , .0300]	[.0144 , .0313]	[.0101 , .0176]	[.0200 , .0350]	[.0188 , .0400]
	Range	[.0001 , .0400]	[.0075 , .0400]	[.0012 , .1273]	[.0019 , .1292]	[.0001 , .0400]	[.0026 , .0275]	[.0101 , .0400]	[.0001 , .0400]
Conduction N=46	Mean	.0245	.0182	.0203	.0157	.0250	.0120	.0323	.0163
	St. Dev.	.0110	.0053	.0153	.0137	.0110	.0048	.0089	.0134
	Median	.0259	.0177	.0175	.0110	.0275	.0126	.0375	.0144
	IQR	[.0020 , .0331]	[.0145 , .0219]	[.0078 , .0282]	[.0063 , .0217]	[.0188 , .0338]	[.0088 , .0138]	[.0250 , .0400]	[.0038 , .0238]
	Range	[.0001 , .0400]	[.0062 , .0300]	[.0019 , .0720]	[.0028 , .0727]	[.0001 , .0400]	[.0038 , .0275]	[.0101 , .0400]	[.0001 , .0400]
Wernicke's N=35	Mean	.0126	.0195	.0332	.0318	.0123	.0115	.0305	.0233
	St. Dev.	.0095	.0059	.0209	.0225	.0096	.0051	.0080	.0130
	Median	.0133	.0193	.0294	.0275	.0126	.0101	.0325	.0250
	IQR	[.0039 , .0187]	[.0152 , .0248]	[.0155 , .0448]	[.0139 , .0472]	[.0032 , .0185]	[.0076 , .0151]	[.0232 , .0388]	[.0123 , .0350]
	Range	[.0002 , .0400]	[.0070 , .0327]	[.0042 , .0979]	[.0038 , .0989]	[.0001, .0400]	[.0038 , .0225]	[.0163 , .0400]	[.0001 , .0400]
Other N=11	Mean	.0180	.0255	.0275	.0283	.0178	.0173	.0293	.0246
	St. Dev.	.0148	.0073	.0222	.0221	.0148	.0057	.0080	.0123
	Median	.0173	.0257	.0143	.0125	.0176	.0176	.0275	.0275
	IQR	[.0019 , .0265]	[.0220 , .0301]	[.0087 , .0488]	[.0106 , .0488]	[.0023 , .0250]	[.0151 , .0200]	[.0250 , .0372]	[.0200 , .0319]
	Range	[.0003 , .0400]	[.0133 , .0400]	[.0049 , .0617]	[.0057 , .0628]	[.0001 , .0400]	[.0076 , .0288]	[.0151 , .0400]	[.0001 , .0400]

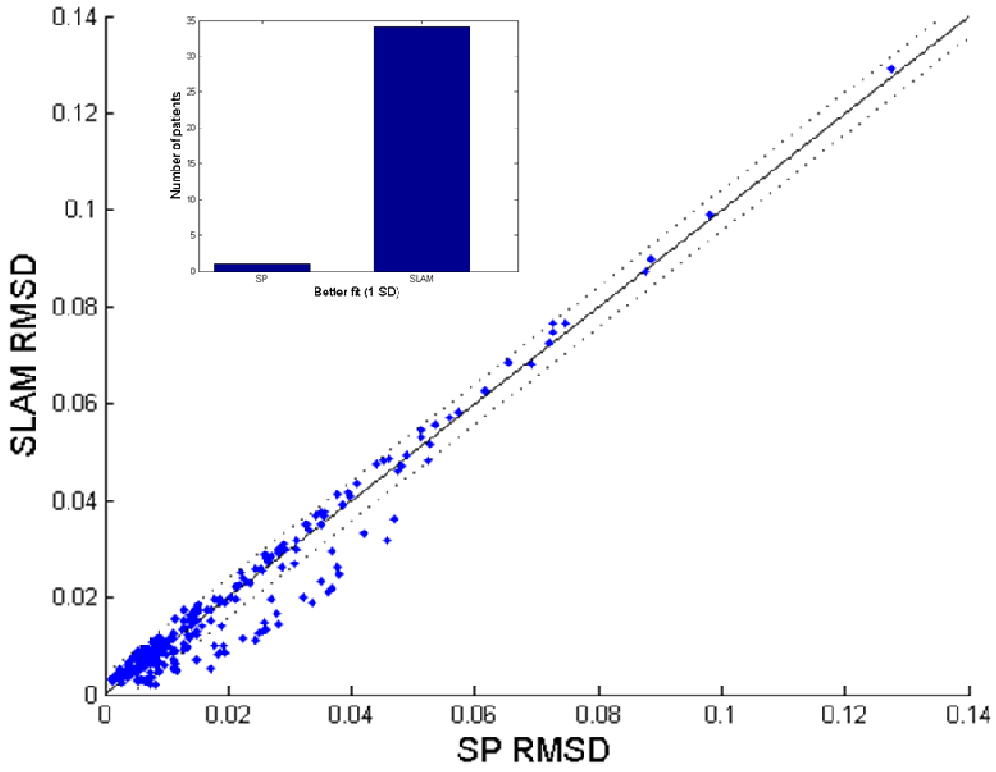

Figure 2.5. Scatterplot comparing model fit between SP and SLAM. The solid diagonal line represents equivalent fits; the dotted lines represent 1 SD of fit difference in the sample. The majority of patients are well fit by both models, and a subgroup of patients are notably better fit by SLAM (inset).

Next, we examined our hypothesis that SLAM would improve model fits specifically for Conduction aphasia. Figure 6 displays the RMSD differences between models for individual patients, grouped by aphasia type. Positive difference values indicate improved fits for SLAM over SP. It is clear that the SLAM model provides the largest and most consistent fit improvements for the Conduction group, and a majority of fits for Wernicke's patients also benefit from the new model. The fact that Wernicke's aphasia was also better fit by SLAM is consistent with the HSFC theory. Wernicke's aphasia is associated with very similar neuroanatomical damage to Conduction aphasia and acute Wernicke's aphasia often recovers to be more like a Conduction profile, suggesting a partially shared locus of

impairment. For a statistical comparison of the fit improvements between the five aphasia subtypes, we performed a one-way ANOVA on the RMSD change, which indicated at least one significant difference between group means ($p<.001$). A follow-up multiple comparison test indicated that the Conduction group benefited more from SLAM than any other group, as the 95% confidence interval for the mean fit improvement did not overlap with any other group, including Wernicke's.

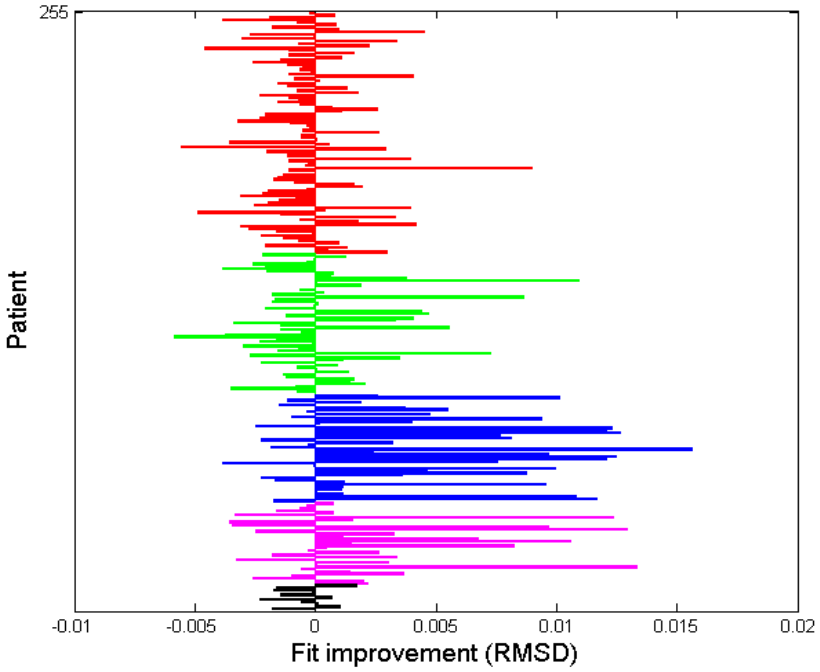

Figure 2.6. Individual fit changes between the SP and SLAM models. Positive values indicate better SLAM fits. Anomic = red, Broca's = green, Conduction = blue, Wernicke's = magenta, Other = black.

To further validate these results, we tested whether fit improvements due to increasing the SP map resolution specifically favored any of the diagnosis groups. Unlike our theoretically motivated structural changes, this method of improving model fits is not expected to favor any particular group. We compared model fits for an SP map with 189 points, which on

average is equivalent to the standard SP map in the literature, to the higher-resolution SP map with 2,321 points. For the group of 255 patients, increasing the number of SP map points significantly improved the average fit from 0.0230 RMSD to 0.0206 RMSD ($p<0.001$). The improvement in fit was significant for all diagnosis groups (all $p<0.001$); however, a one-way ANOVA with follow-up multiple comparison tests showed that there was no group that had significantly greater improvement than every other group (no disjoint confidence intervals), unlike the result produced by our structural changes, which specifically favored the Conduction group. Instead, the Wernicke's group improved most, while the Anomic group improved least, consistent with the observation that these groups are already the worst and best fit by SP, respectively. The implication is that the improvements in fit caused by our theoretically motivated manipulation of the SP model's architecture are qualitatively different than improvements gained by other methods.

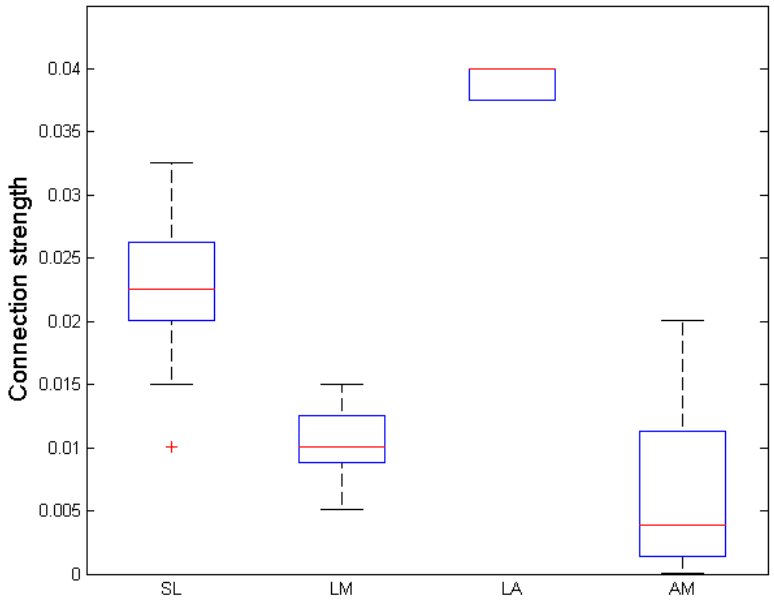

Figure 2.7. Boxplots show the SLAM weights for the group of 20 patients with the greatest fit improvements. As expected, a model profile with high lexical-auditory and low auditory-motor weights leads to the greatest improvements over the SP model.

We also hypothesized that the Conduction naming pattern should be fit by a particular SLAM configuration: strong LA and weak AM weights. For the patients who exhibited the greatest improvements in fit, this was indeed the case. Figure 2.7 uses boxplots to display the SLAM weight configurations for the 20 patients (13 Conduction, 5 Wernicke's, 1 Anomic, 1 Broca's) who exhibited the greatest fit improvements (> 2 SD). Figure 2.8 shows data from an example patient with Conduction aphasia, along with the corresponding SP and SLAM model fits. The best fitting weights in the SP model were .022 and .017, for S and P, respectively. The SLAM model yielded .023 and .013 for SL and LM, respectively, while LA weights were maximized at .04 and AM weights were minimized at .0001. For this patient, SLAM reduced SP fit error by .0135 RMSD. This example also illustrates how SLAM's largest fit improvements over SP are accompanied by a consistent increase in the predicted frequency of Formal errors, along with a consistent decrease in Semantic (and Unrelated) errors. This trade-off in Formal errors for Semantic errors is most likely to occur at the first, lexical-selection step. The dual nature of Formal errors, that they can occur during either lexical or phonological selection, is one of the hallmarks of the SP model. Foygel and Dell (2000) showed that Formal errors during lexical selection increase when phonological feedback to lexical units outweighs the semantic feedforward activation. In Conduction aphasia, large LA weights provide strong phonological feedback to Lexical units, while small AM and LM weights provide weak phonological feedforward to the Motor units. With LM greater than AM, more activation flows from the incorrect, phonologically-related lexical items, thereby increasing Formal errors at the expense of

Semantic errors. The implication, that strong auditory-phonological feedback influences lexical selection in Conduction aphasia, represents a novel prediction of our model that is supported by the data.

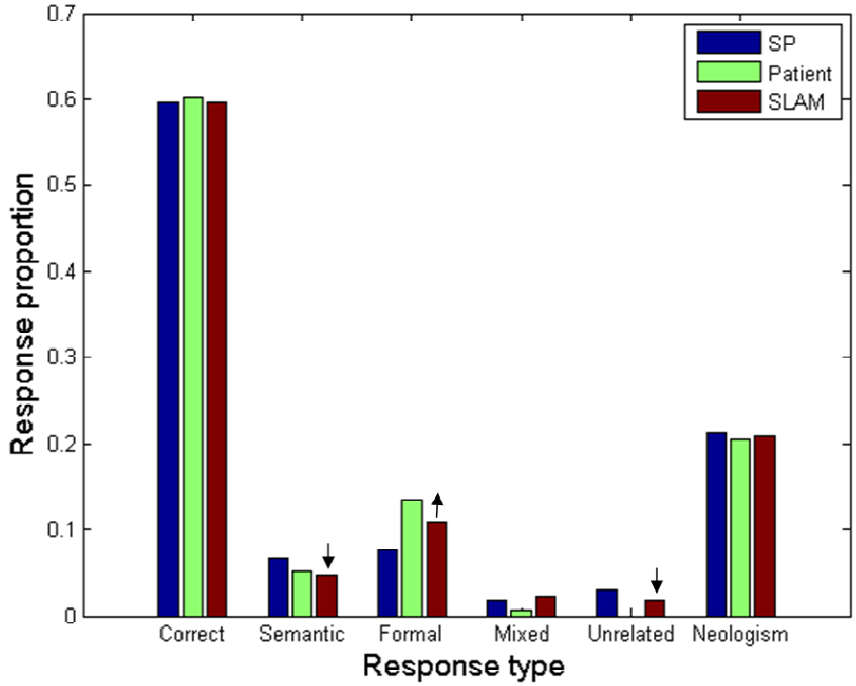

Figure 2.8. Naming response distribution from an example patient with Conduction aphasia, along with the corresponding SP and SLAM model fits. Arrows indicate how SLAM improves the fit to data, increasing Formal at the expense of Semantic and Unrelated errors. The SLAM model reduced fit error for this patient by 0.0135 RMSD.

Finally, we tested the criticality of our assumption that LA weights must be greater than LM weights. We repeated our original analysis, this time comparing SP to SLMA, an alternative version of SLAM that has lexical-motor dominance instead of lexical-auditory dominance. SLMA was fit with a 4-parameter map with 2,321 points, the same size as SLAM, culled from the 13,807 discarded SLAM points, ensuring that LM weights were always greater than or equal to LA weights. Figure 2.9 is a scatterplot comparing SP and SLMA model fits; the diagonal lines are the same as in Figure 2.5. When this alternative model architecture

was used, there were no noticeable improvements over SP; the maximum change in fit was only 0.0038 RMSD. Thus, it was not the mere presence of additional parameters in SLAM that caused the observed fit improvements; their theoretically motivated arrangement was necessary as well.

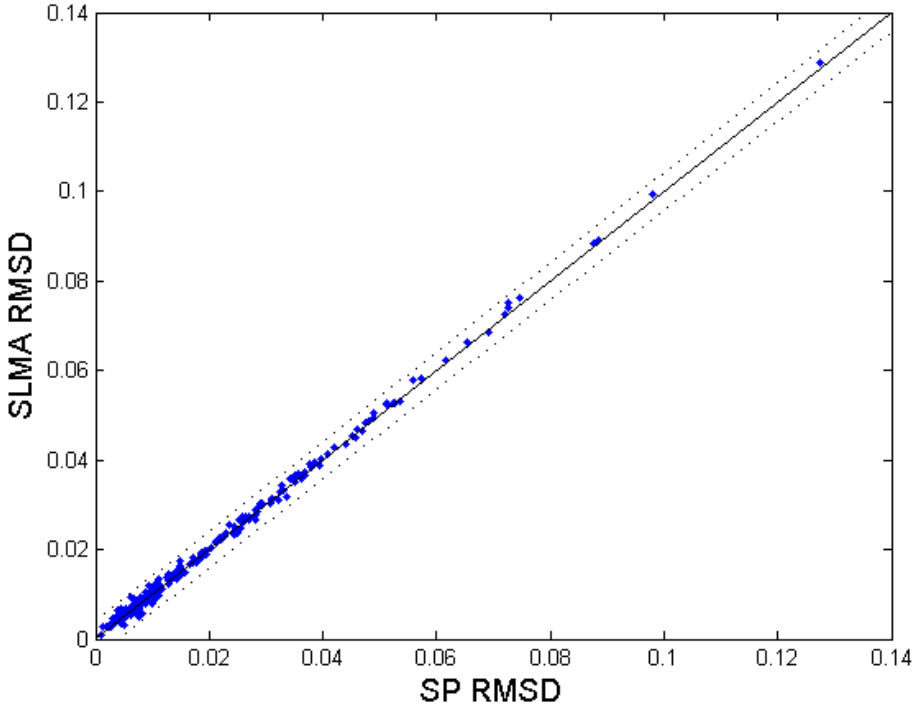

Figure 2.9. Scatterplot comparing model fit between SP and SLMA, an alternative architecture with the same number of parameters as SLAM, but with lexical-motor dominance instead. The lines are the same as in Figure 5. Unlike SLAM, SLMA provides no obvious fit improvements.

We also explored the necessity of the LM weights, testing the importance of our two routes. We fixed the LM weights at .0001 (effectively zero) by using 323 points from the full SLAM map to fit the data, thus yielding a 3-parameter model, and we compared these fits with fits from an SP map that had the same number of points. This 3-parameter model that lacked direct LM connections did much worse than the 2-parameter SP model, yielding an average fit of .10 RMSD. This catastrophic failure was due to the fact that not enough activation

reached the motor units via the lexical-auditory-motor route. Recall that activation is multiplied by a fraction at each level yielding lower activation after two-steps through the lexical-auditory-motor route compared to the one-step lexical-motor route. Without the combined input to motor units from the two routes, the model could only produce a maximum estimate of 65% correct responses. While HSFC theory does predict that direct lexical-motor connections are required for normal levels of correctness, the weaker input to motor units from the auditory-motor route raises the concern that our initial choice of SLAM parameter constraints gave more prominence to the lexical-motor route than the HSFC theory warrants. We therefore explored the SLAM parameter space further, and we discovered alternative parameter constraints that yields qualitatively similar results: in the "healthy model", SL and LA weights have the usual maximum value of .04, while LM weights have a maximum of .02, and AM weights have a maximum of .5; in aphasia, the parameters are free to vary below those values. This parameter arrangement ensures that the primary source of phonological feedback to the lexical layer is usually from auditory units, enables the auditory-motor route to provide strong activation to motor units during naming, and removes the previous constraint that in damaged states, the LM weights must always be lower than the LA weights. As with the original choice of SLAM parameter constraints, we observe similar fits as SP in the general population, with noticeable improvements for the conduction naming pattern, accompanied by high LA and low AM weights. With this alternative arrangement, a 3-parameter model with LM weights fixed at .0001 still does not perform as well as the 2-parameter SP model, although the failure is no longer catastrophic, due to compensation by strong AM weights. To summarize, these investigations confirm our main finding that a second source of phonological feedback,

predicted by HSFC theory to come from the auditory system, is the critical component for improving model fits.

Discussion

We put forward a new computational model of naming, SLAM, inspired by a recent conceptual model, HSFC, aimed at integrating psycholinguistic and motor control models of speech production. SLAM implemented the HSFC claims that sublexical linguistic units have dual representations within auditory and motor cortices, and that the conversion of auditory targets to motor commands is a crucial computation for lexical-retrieval, even prior to overt production.

We showed that augmenting the well-established SP model to incorporate auditory-to-motor conversion into the lexical-retrieval process allowed the model to explain general aphasic naming errors at least as well, while improving the model's ability to account for Conduction naming patterns in particular. The improvements in model fits were predicted to result from parameter settings with strong LA and weak AM weights. Examining the naming responses of 255 aphasic patients, the largest analysis of PNT responses to date, we confirmed our predictions, and additionally demonstrated that, unlike our theoretically motivated structural changes, improvements due to added map resolution were not specific to any aphasia type. We also discovered that the predicted weight configuration, which yielded the greatest fit improvements, did so by increasing Formal errors at the expense of Semantic errors. It's worth noting in this context that Schwartz et al. (2006) identified three anomalous subgroups whose naming patterns significantly deviate from

SP's predictions, one of which exhibits too many Formal errors. Two of the patients in this subgroup had Conduction aphasia, while the other had Wernicke's aphasia. SLAM provides a plausible explanation for this subgroup. The increase of Formal errors at the expense of Semantic errors in Conduction aphasia suggests that a significant proportion of their phonologically-related errors are generated at the lexical-selection stage, rather than the phonological-selection stage, a novel prediction of our model. We also found that two separate phonological routes were required to produce the effect. Although the auditory-motor integration loop described by HSFC theory currently is not modeled in detail within SLAM, parallel inputs and feedback to separate auditory and motor systems are a pre-requisite for state feedback control. The results of our modeling experiments thereby support the assumptions of the HSFC framework.

Although we pit SP and SLAM against one another, they share many of their essential features. Thus, much of SLAM's success can be attributed to the original SP model's assumptions. The notions of computational simultaneity, hierarchical representation, interactivity among hierarchical layers, localized damage, and continuity between random and well-formed outputs are what enabled good predictions. The fact that we were able to successfully extend the model reinforces the utility of these ideas. Similarly, much of the criticism of SP applies equally to SLAM. For instance, the very small lexicon can only approximate the structure of a real lexicon and semantic representations are arbitrarily defined. While the model is interactive, it does not include lateral or inhibitory connections, which are essential features of real neurological systems. Also, the model does not deal directly with temporal information, which constitutes a large body of the psycholinguistic

evidence regarding speech processing. Nevertheless, for examining the architectural assumptions of the HSFC, SP provides a useful test bed in that it was the best computational model available.

One further advantage of SLAM over SP (and over similar models that assume a unified phonological network) is that SLAM provides a built in mechanism for repetition. Repetition is often used in addition to naming as a test of lexical-retrieval models, because repetition involves the same demands on the motor production system as naming, but lacks the semantic search component. In order to simulate repetition, however, some form of auditory representation is necessary, even if it is implicit. In Foygel and Dell (2000), the single-route SP model was used to simulate repetition, without explicitly modeling auditory input, by assuming that perfect auditory recognition delivers a boost directly to lexical units, essentially just the second step of naming. Later, to account for patients with poor naming but spared repetition abilities, a direct input-to-output phonology route was added to the model (Hanley, Dell, Kay & Baron, 2004). This dual-route model grafts the "non-lexical" route on to SP, leaving the architecture and simulations of naming unchanged; the two routes are used only during repetition. While several studies have generated empirical support for the idea that the two routes are indeed used in repetition (Nozari, Kittredge, Dell, & Schwartz, 2010), our study suggests that both routes are used in naming as well, potentially providing a more cohesive account of the computations underlying these tasks. Given that SLAM already requires the auditory component for naming, we intend to develop it to simulate repetition as well, allowing for more direct comparisons to this alternative dual-route model in the future.

While SLAM does not employ learning or time-varying representations, another lexical retrieval model that does implement these features has also adopted a similar separation of auditory and motor speech representations. Ueno, Saito, Rogers, and Ralph (2011) present Lichtheim 2, a "neurocomputational" model, which simulates naming, repetition, and comprehension for healthy and aphasic speech processing, using a network architecture in which each layer of units corresponds to a brain region. Lichtheim 2 does not categorize speech error types according to SP's more detailed taxonomy, however, making it hard to compare directly with SLAM. Furthermore, since our goal with SLAM was to investigate the effects of the separate phonological representations, and Lichtheim 2 shares this architectural assumption, we did not compare the models directly. In Lichtheim 2, input and output phonology is represented by a pattern of phonemic features presented one cluster at a time, while semantic representations are temporally static and statistically independent of their corresponding phonological representations. The model is simultaneously trained on all 3 tasks and hidden representations are allowed to form in a largely unconstrained manner. The trained network can then be "lesioned" in specific regions to simulate aphasic performance. We see much in common between our approaches in terms of the theoretical motivations, proposing psycholinguistic representations grounded in neuroanatomical evidence. Furthermore, the use of a single network to perform multiple tasks is very much in line with our plans to develop the SLAM model. One major difference between SLAM and Lichtheim 2 is that SLAM maintains an explicit hierarchical separation between lexical units and phonological units, allowing for selection errors at either stage. This hierarchical separation was essential for making our

successful predictions regarding Conduction naming patterns. It remains to be seen how our proposed architecture will cope with multiple tasks simultaneously.

Another model of lexical production, WEAVER++/ARC (Roelofs, 2014), has been proposed as an alternative to Lichtheim 2. While this model uses spreading activation through small, fixed networks, like SP, it also employs condition-action rules to mediate task-relevant selection of the network's representations, thereby implementing a separation of declarative and procedural knowledge. Like Lichtheim 2, this model does not apply the detailed error taxonomy examined by SLAM and so we did not compare them directly. Importantly though, WEAVER++/ARC and Lichtheim 2 largely agree on most cognitive and computational issues, especially the primary one investigated by SLAM: the participation of separate auditory and motor phonological networks in speech production. Additionally, like SLAM and Lichtheim 2, WEAVER++/ARC simulates the Conduction aphasia pattern by reducing weights between input and output phonemes. The primary disagreement between WEAVER++/ARC and Lichtheim 2 is an anatomical one; should the lexical-motor connections for speech production be associated with the (dorsal) arcuate fasciculus or the (ventral) uncinate fasciculus? At present, the SLAM model is compatible with either position.[2] WEAVER++/ARC does differ from SLAM with respect to one important theoretical point, however. In WEAVER++/ARC, there is a separation of input and output lexical units, and in naming, activation primarily flows from lexical output units to motor units. Auditory units then provide stabilizing activation to motor units through an auditory

[2] One might wonder whether lexical-motor and auditory-motor connection weights were generally correlated in our sample. They were not ($r = .10$, $p = .09$). This seems to indicate that these mappings are functionally and anatomically distinct; however, WEAVER++/ARC also allows these routes to be independently lesioned, so this is not necessarily a strong point of disagreement.

feedback loop (i.e., motor-to-auditory-to-motor) rather than being activated by a single lexical layer in parallel with motor units to serve as sensory targets. This runs contrary to our finding that strong lexical-auditory feedback influenced lexical selection for Conduction aphasia. Again, it remains to be seen whether our assumption of a single lexical layer can account for multiple tasks as Lichtheim 2 and WEAVER++/ARC do, which we intend to explore in future work.

The SLAM model falls into a broad class of models that can be described as "dual-route" models, that is, models that posit separate but interacting processing streams controlling behavior. Much of this work relates directly to Hickok and Poeppel's (2000, 2004, 2007) neuroanatomical dual stream framework for speech processing in that the mapping between auditory and motor speech systems corresponds to the dorsal stream, while the mapping between auditory and lexical-semantic levels corresponds to the ventral stream. While Hickok and Poeppel discussed this cortical network from the perspective of the auditory speech system, which diverges into the two streams, picture naming traverses both streams, going from conceptual to lexical to auditory (ventral stream) and from auditory to motor (dorsal stream). One difference between the SLAM model and the Hickok and Poeppel framework is that explicit connectivity is assumed between lexical and motor-phonological networks. Hickok and Poeppel assumed (but didn't discuss) connectivity between conceptual and motor systems, but did not specifically entertain the possibility of lexical to motor speech networks. The present model, along with the HSFC, thus refine the Hickok and Poeppel dual stream framework.

References

Abel, S., Huber, W., & Dell, G. S. (2009). Connectionist diagnosis of lexical disorders in aphasia. *Aphasiology*, *23*(11), 1353-1378.

Anderson, J.M., Gilmore, R., Roper, S., Crosson, B., Bauer, R.M., Nadeau, S., Beversdorf, D.Q., Cibula, J., Rogish, M., 3rd, Kortencamp, S., et al. (1999). Conduction aphasia and the arcuate fasciculus: A reexamination of the Wernicke-Geschwind model. Brain Lang. 70, 1–12.

Benedict, H. (1979). Early lexical development: Comprehension and production. *Journal of child language*, *6*(02), 183-200.

Bossom, J. (1974). Movement without proprioception. *Brain Res, 71*, 285-296.

Callan, D. E., Tsytsarev, V., Hanakawa, T., Callan, A. M., Katsuhara, M., Fukuyama, H., & Turner, R. (2006). Song and speech: brain regions involved with perception and covert production. *Neuroimage*, *31*(3), 1327-1342.

Caramazza, A. (1991). Some aspects of language processing revealed through the analysis of acquired aphasia: The lexical system. In *Issues in reading, writing and speaking* (pp. 15-44). Springer Netherlands.

Cole, J. D., & Sedgwick, E. M. (1992). The perceptions of force and of movement in a man without large myelinated sensory afferents below the neck. *J Physiol, 449*, 503-515.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, *93*(3), 283.Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, *43*(2), 182-216.

Dell, G. S., Lawler, E. N., Harris, H. D., & Gordon, J. K. (2004). Models of errors of omission in aphasic naming. *Cognitive Neuropsychology*, *21*(2-4), 125-145.

Dell, G. S., Martin, N., & Schwartz, M. F. (2007). A case-series test of the interactive two-step model of lexical access: Predicting word repetition from picture naming. *Journal of Memory and Language*, *56*(4), 490-520.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological review*, *104*(4), 801.

Dell, G. S., Schwartz, M. F., Nozari, N., Faseyitan, O., & Coslett, H. B. (2013). Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition*, *128*(3), 380-396.

Geschwind, N. (1965). Disconnexion syndromes in animals and man. I. Brain 88, 237–294, 585–644.

Goodglass, H. (1992). Diagnosis of conduction aphasia. In S. E. Kohn (Ed.), Conduction aphasia (pp. 39-49). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review, 105*, 611-633.

Hanley, J. R., Dell, G. S., Kay, J., & Baron, R. (2004). Evidence for the involvement of a nonlexical route in the repetition of familiar words: A comparison of single and dual route models of auditory repetition. *Cognitive Neuropsychology*, *21*(2-4), 147-158.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*, 135-145.

Hickok, G. (2014a). The architecture of speech production and the role of the phoneme in speech processing. *Lang Cogn Process, 29*, 2-20.

Hickok, G. (2014b). Toward an Integrated Psycholinguistic, Neurolinguistic, Sensorimotor Framework for Speech Production. *Lang Cogn Process, 29*, 52-59.

Hickok, G., & Buchsbaum, B. (2003). Temporal lobe speech perception systems are part of the verbal working memory circuit: Evidence from two recent fMRI studies. *Behavioral and Brain Sciences*, *26*(06), 740-741.

Hickok, G., Erhard, P., Kassubek, J., Helms-Tillery, A.K., Naeve-Velguth, S., Strupp, J.P., Strick, P.L., and Ugurbil, K. (2000). A functional magnetic resonance imaging study of the role of left posterior superior temporal gyrus in speech production: Implications for the explanation of conduction aphasia. Neurosci. Lett. 287, 156–160.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron*, *69*(3), 407-422.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in cognitive sciences*, *4*(4), 131-138.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1), 67-99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393-402.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*, 1213-1216.

Jacobs, O. L. R. (1993). Introduction to control theory. (Oxford: Oxford University Press).

Jacquemot, C., Dupoux, E., & Bachoud-Lévi, A. C. (2007). Breaking the mirror: Asymmetrical disconnection between the phonological input and output codes.*Cognitive Neuropsychology*, *24*(1), 3-22.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Curr Opin Neurobiol, 9*, 718-727.

Kittredge, A. K., Dell, G. S., Verkuilen, J., & Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology*, *25*(4), 463-492.

Larson, C. R., Burnett, T. A., Bauer, J. J., Kiran, S., & Hain, T. C. (2001). Comparison of voice F0 responses to pitch-shift onset and offset conditions. *Journal of the Acoustical Society of America, 110*, 2845-2848.

Lomas, J., & Kertesz, A. (1978). Patterns of spontaneous recovery in aphasic groups: A study of adult stroke patients. *Brain and Language*, *5*(3), 388-401.

Mirman, D., Strauss, T. J., Brecher, A., Walker, G. M., Sobel, P., Dell, G. S., & Schwartz, M. F. (2010). A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive neuropsychology*, *27*(6), 495-504.

Nozari, N., Kittredge, A. K., Dell, G. S., & Schwartz, M. F. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of memory and language*, *63*(4), 541-559.

Okada, K., & Hickok, G. (2006). Left posterior auditory-related cortices participate both in speech perception and speech production: Neural overlap revealed by fMRI. *Brain and Language*, *98*(1), 112-117.

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics, 25*, 382-407.

Pulvermüller, F. (1996). Hebb's concept of cell assemblies an the psychophysiology of word processing. *Psychophysiology*, *33*(4), 317-333.

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia naming test: Scoring and rationale. *Clinical aphasiology*, *24*, 121-134.

Roelofs, A. (2014). A dorsal-pathway account of aphasic language production: The WEAVER++/ARC model. *Cortex*, *59*, 33-48.

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics, 25*, 421-436.

Sanes, J. N., Mauritz, K. H., Evarts, E. V., Dalakas, M. C., & Chu, A. (1984). Motor deficits in patients with large-fiber sensory neuropathy. *Proc Natl Acad Sci U S A, 81*, 979-982.

Schwartz, M. F., & Brecher, A. (2000). A model-driven analysis of severity, response characteristics, and partial recovery in aphasics' picture naming.*Brain and language*, *73*(1), 62-91.

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and language*, *54*(2), 228-264.

Shadmehr, R., Smith, M. A., & Krakauer, J. W. (2010). Error correction, sensory prediction, and adaptation in motor control. *Annual Review of Neuroscience, 33*, 89-108.

Tremblay, S., Shiller, D. M., & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature, 423*, 866-869.

Ueno, T., Saito, S., Rogers, T. T., & Ralph, M. A. L. (2011). Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron*, *72*(2), 385-396.

Wernicke, C. (1874). The symptom complex of aphasia: A psychological study on an anatomical basis. Reprinted in Boston Studies in the Philosophy of Science (1969), R.S. Cohen and M.W. Wartofsky, eds. (Dordrecht: D. Reidel Publishing Company), pp. 34–97.

Wolpert, D. M. (1997). Computational approaches to motor control. *Trends Cogn Sci, 1*, 209-216.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science, 269*, 1880-1882.

Yates, A. J. (1963). Delayed auditory feedback. *Psychological Bulletin, 60*, 213-251.

CHAPTER 3: Distinguishing SLAM from Post-Lexical Processing (LPL)

The Semantic-Lexical-Auditory-Motor (SLAM) model of speech production (Walker & Hickok, 2015) represents an attempt to evaluate the effects of a theoretically motivated architectural modification of the Semantic-Phonological (SP) model of lexical retrieval (Foygel & Dell, 2000). The modification involved splitting the phonological layer into two parts: an auditory and a motor component. This was motivated by neuroscience data and motor control theory, which both highlight the importance of sensorimotor interaction in controlling movement, including speech (Hickok, 2012). Part of the neuroscience data that motivated the architecture came from conduction aphasia, which can be conceptualized as a sensorimotor deficit in the linguistic domain, and thus we specifically predicted the SLAM model would provide better fits compared to SP for naming error patterns in this syndrome. Our prediction was confirmed without sacrificing fits for other types of aphasia. Furthermore, we used the clinical description of the conduction syndrome to predict the model configuration that would lead to fit improvements: strong auditory-lexical connections and weak auditory-motor connections. This prediction was confirmed. Moreover, we discovered that this model configuration improved fits for the conduction aphasia group specifically by accounting for sound-related errors via the interaction of the lexical and phonological levels as opposed to dysfunction at the phonological level alone. We took SLAM's success as support for the idea that an integration of psycholinguistic, motor control, and neuroscience was (is) feasible (Hickok, 2014a, 2014b).

Goldrick (2015) is unconvinced, however, that SLAM represents any real theoretical progress. He argues instead that SLAM does better than SP because it is an approximation of a lexical + post-lexical phonological theory (henceforth LPL) proposed by Goldrick and Rapp (2007), which he claims provides a better account of the sound-related errors in conduction aphasia by placing their source at the post-lexical level. In partial support of his claim, he presents a regression analysis showing that SLAM's fit improvements over SP for conduction aphasia are correlated with the number of sound-related errors that the patients produced. In reply, we make three points. First, Goldrick is comparing his conceptual model against a computational implementation of a *part of* our own conceptual framework. His arguments hold no water against our broader theoretical perspective. Second, because Goldrick has not used LPL to make quantitative predictions about the same data as SLAM, there is no objective metric to evaluate the claim that SLAM's improvement is due to its approximating LPL. Third, when we implement the LPL theory in a computational model, we find that it fails to provide the same fit improvements as SLAM.

On the Relation Between SLAM, HSFC, and LPL

Goldrick begins his commentary by correctly noting that we presented an implementation of *aspects of* Hickok's (2012) Hierarchical State Feedback Control (HSFC) theory. He then fails to notice that (a) the unimplemented aspects provide exactly the post-lexical component he calls for and (b) the goal of SLAM was to assess precisely the one bit that we changed, not a full-blown implementation of the entire system as we understand it. In Goldrick's (2015) Figure 1, the SLAM architecture is redrawn to show its similarity to the

LPL theory and to highlight the difference with respect to the existence of a post-lexical stage of phonological processing, present in LPL and absent in SLAM. But this is misleading with respect to the broader theoretical context in which SLAM is situated. Figure 3.1 here compares the architectures of the LPL conceptual theory with the HSFC conceptual theory. HSFC proposes the existence of sensorimotor loops that correspond to different hierarchical levels of phonological processing. One could readily map the auditory-phonological loop onto LPL's lexical phonological level and the somatosensory-phonological loop onto the post-lexical level. With this alignment there are no presumed architectural advantages of LPL in terms of selection levels. That is, if we implemented this more complete architecture we would, according to Goldrick's arguments, provide a better fit to more of the data[3]. This remains to be seen, of course, but it is an interesting and potentially fruitful direction for further development. And now that we better understand the computational effects of sensorimotor loops in word production models, which was the aim of developing SLAM, we are in a good position to take the next step.

[3] Goldrick's primary argument hinges on the claim that SLAM is a limited implementation of HSFC and therefore accidentally captures the details of the LPL theory rather than the intended theory. The criticism implies that Walker and Hickok (in press) overlooked the presence of post-lexical errors in the data, and because SLAM is an approximation of LPL, it is accidentally accounting for these errors in order to improve the fit. We note that, while SP and SLAM are both limited implementations of larger theories, the models both attempt to account for the theoretical notion of post-lexical errors through a practical implementation of lenient scoring. For patients with obvious articulatory-motor impairment, including verbal apraxia (e.g., Romani & Galluzzi, 2005; Romani, Galluzzi, Bureca, & Olson, 2011; Galluzzi, Bureca, Guariglia, & Romani, 2015), scoring is such that responses with a single addition, deletion, or substitution of a phoneme or consonant cluster are scored as correct. This scoring procedure is based on the assumption that the error occurred after the correct selection at the phonological layer of the model (Schwartz et al., 2006). This means that many of the sound-related errors that Goldrick assumes would be better explained by LPL and, by extension, SLAM were actually excluded from the analysis. Nevertheless, according to the LPL theory, any patient has the potential to exhibit a post-lexical processing error. Thus, while the analysis of SLAM clearly did not overlook the potential for post-lexical errors, it is possible that our efforts did not remove these effects from the data entirely (See Goldrick, Folk, and Rapp, 2010 for further discussion). We therefore tried to evaluate Goldrick's claim that SLAM's improvements are in fact due to its approximation of the LPL theory.

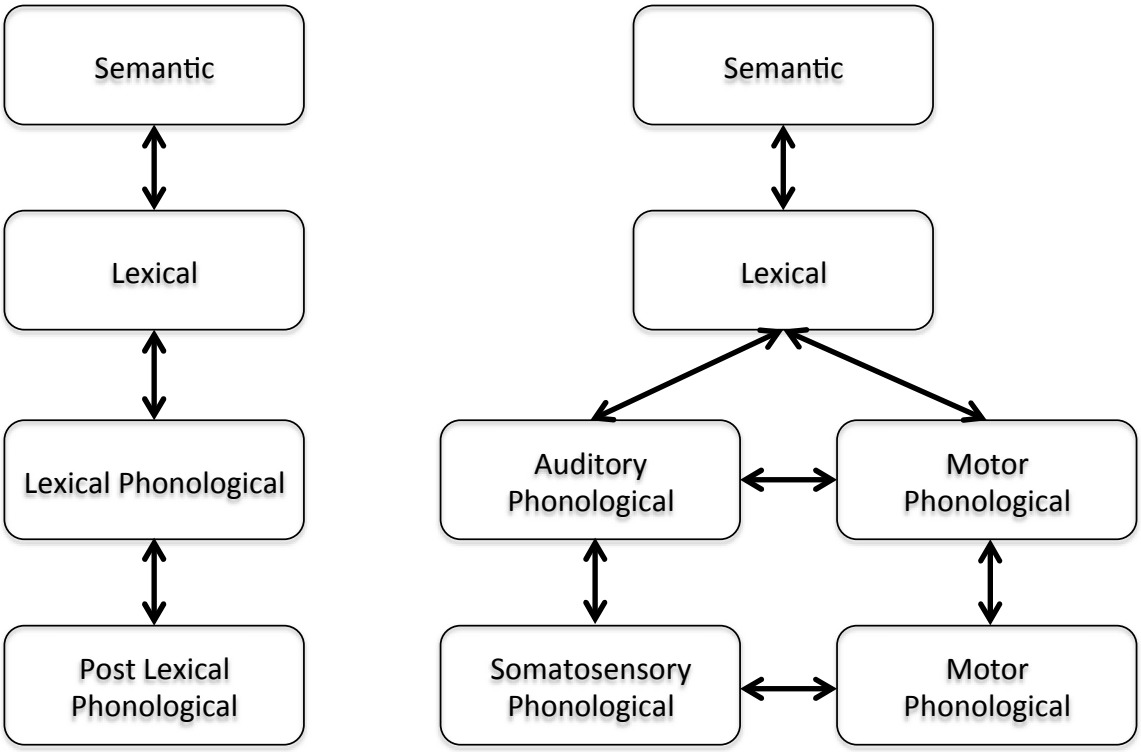

Figure 3.1. The lexical + post-lexical (LPL) model (left) compared with the Hierarchical State Feedback Control (HSFC) model (right).

How do we know if LPL accounts for the same data as SLAM?

The LPL theory is a conceptual one, not a computational implementation. We can readily compare the amount of variance accounted for by the two computational models we evaluated, SP and SLAM, and apply quantitative metrics to determine which one does a better job. Goldrick does not present a quantitative metric to evaluate the performance of the LPL theory relative to these models, making it impossible to confirm or refute his claim based on his arguments. For example, Goldrick asserts that SLAM has "great difficulty" in accounting for his prototypical case of a patient with only sound-related errors. What we know is that SLAM predictions for this case deviate (using the RMSD metric) from the observed values on average by 1.5% (Figure 3.2), which is better than the typical fit error values across all patients. Is this a good fit, or is SLAM having great difficulty? The only

way to tell is to compare it with the fit of a competitor model. If the competitor's fit differs by 10%, then SLAM provides a good fit. If the competitor's fit is off by 0.01%, then indeed SLAM is having great difficulty. Since Goldrick provides no quantitative comparison, his assertions are vacuous.

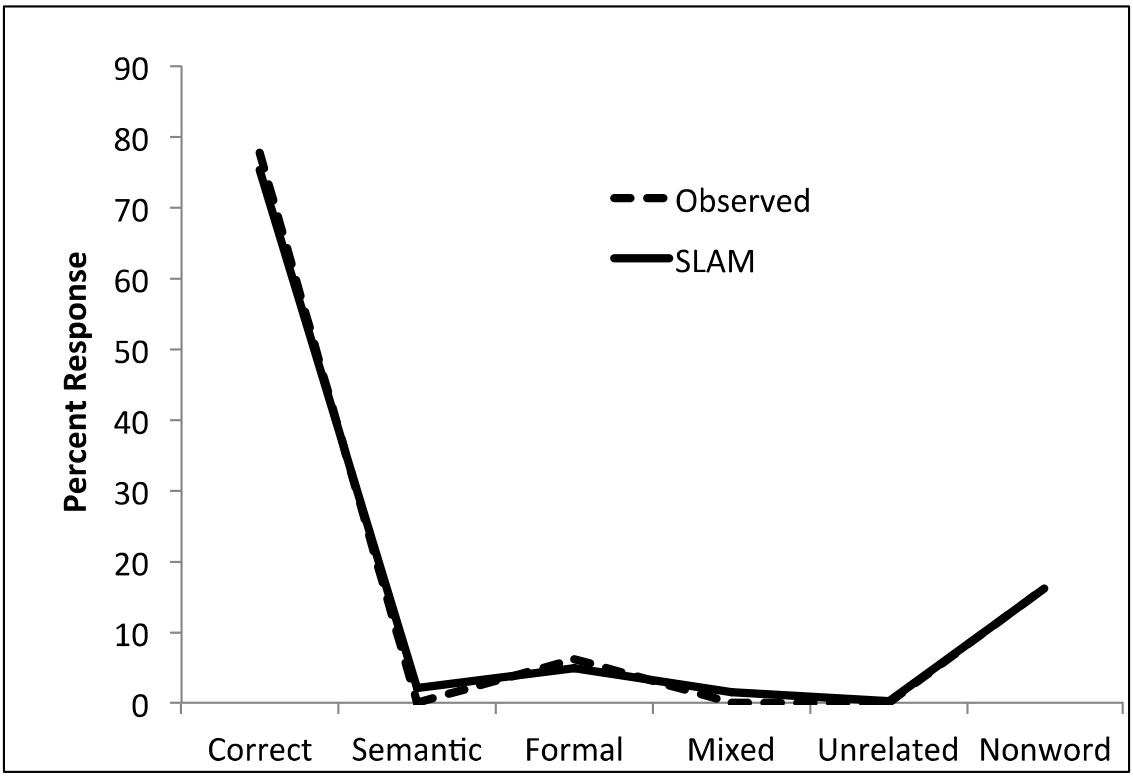

Figure 3.2. Observed response pattern in a patient with no semantic errors and SLAM's best fit predictions.

The same lack of quantitative comparison undermines Goldrick's critique of our methodological approach. The critique is based on a set of simulations involving the SP model reported previously (Goldrick, 2011). The simulations involved generating artificial datasets with SP as well as two alternative models. Data from the three models were then fit with parameters from the SP model. Goldrick claims that "the degree of fit was equivalent for all three artificial case series." From this observation, he suggests that model

fitting cannot always discriminate between theoretical accounts. However, his statements are misleading and his simulation strategy is inadequate. In order to discriminate models quantitatively, it is not enough to look at fits from a single model and judge them good or poor, equivalent or different. Rather, one needs to evaluate the fits of each model and assess whether one provides a better account of the data than the other.

To demonstrate this point and to evaluate whether our methods can indeed discriminate SP and SLAM, we ran simulations similar to those carried out in Goldrick (2011). Specifically, we generated three artificial data sets using the SP model and three using SLAM. Two of the artificial datasets from each model were generated by simulating 175 naming attempts from each of 1,000 simulated patients, following Goldrick (2011); the third artificial dataset was generated by simulating 175 naming attempts from each of 255 simulated patients (see below). For the first two artificial datasets, a given patient was simulated using a set of model parameters selected randomly from a continuous distribution of parameters. Goldrick (2011) used a single continuous distribution of parameters in his simulations, namely, all weights were independently and normally distributed with a mean of 0.025, a standard deviation of 0.01, and truncated on the interval [0.0001, 0.05]. Since this parameter space distribution is an arbitrary choice, we used two different arbitrary distributions applied to each model (SP and SLAM) to provide a more thorough evaluation of model discriminability. The first was a normal distribution with mean 0.02 and standard deviation 0.01 truncated on the interval [0.0001, 0.04]. These values were selected because we used a maximum weight of 0.04 in our original simulations for both SP and SLAM. Also, in accordance with our original simulations of

SLAM, the LM-weight was re-sampled until it was less than the LA-weight (Walker & Hickok, in press). The normal distribution assumes that most aphasic patients will have weights around 0.02, and few will have extreme weight values. In the second simulated dataset, we used a uniform distribution of weight values, that is, a distribution in which any value is equally likely to be selected over the interval [0.0001, 0.04]; again we constrained the LM-weight to be less than the LA-weight for the SLAM simulations. Our third simulation used the empirical distribution of weight configurations in our sample rather than randomly sampling from a continuous distribution of parameter values. For each of the 255 patients in our sample, we generated a new simulated patient with 175 naming attempts using the best fit parameter values from each model. These procedures generated six datasets: three generated by SP and three generated by SLAM. We then fit each dataset with each of the models, using the same maps with 2,321 points and 10,000 naming attempts that we used in our previous studies.

For each data set, we used a paired, two-tailed t-test to assess whether the models produced significantly different fits to the data on average. We note that null hypothesis testing is not the only way to quantitatively compare models, but it provides a familiar frame of reference. Our null hypothesis was that the models have equal fit to the data on average and thus cannot be discriminated with our method. Table 3.1 shows the average fit of each model to each data set, along with the preferred model if we had enough evidence to reject our null hypothesis and successfully discriminate between them.

Table 3.1. Results from paired, two-tailed t-tests comparing mean RMSD for the SP and SLAM models. The data sets that were generated with parametric distributions each have

1,000 simulated patients. The data sets that were generated with empirical distributions each have 255 simulated patients.

Data set	SP fit (mean RMSD)	SLAM fit (mean RMSD)	Preferred model	p-value
SP_normal	0.0118	0.0124	SP	0.0021
SP_uniform	0.0116	0.0123	SP	0.0006
SP_empirical	0.0099	0.0106	SP	0.00006
SLAM_normal	0.0123	0.0117	SLAM	0.0155
SLAM_uniform	0.0120	0.0116	SLAM	0.0512
SLAM_empirical	0.0105	0.0096	SLAM	0.00004

In each comparison, the true model that generated the data provided a better fit compared to the alternative model. A statistical test of the difference in fits between the two models shows that the difference is statistically significant in each case (or all but one case if the p = 0.0512 is counted as non-significant). This makes it a non-trivial finding that there are enough individuals from a broad sample of post-acute aphasic patients concentrated in regions of parameter space to detect a difference between the models; if SLAM did not truly improve fits over SP for the aphasia population, then our model comparisons would have indicated this. We note that these effect sizes would be much larger if the analysis was applied to the empirical distribution of only the conduction patients. We also note that our observed effect sizes[4] are much smaller than those reported by Goldrick (2011). If we had used Goldrick's (2011) model evaluation method (comparing datasets, not models), we would simply observe that the average fit of the SP model to the SP_normal data is very similar to the average fit of the SP model to the SLAM_normal data, 0.0118 and 0.0123, respectively, and conclude without further analysis that it is impossible to discriminate

[4] Goldrick (2011) reports that the SP model fits the SP-generated dataset with 0.01 mean RMSD and other datasets with 0.017 mean RMSD. Even though this represents a 70% increase in error, and is judged by Goldrick (2011) to be "equivalent", these effect sizes regard the difference in fit between datasets, not the difference in fit between models. These comparisons should not be confused with our effect sizes in fit between models, which is a more clearly interpretable quantity.

between the models. This illustrates the importance of comparing different models of the same data.

How does an implemented version of LPL fare?

Goldrick contends that the only reason SLAM improves fits over SP is because SLAM includes an intervening layer between lexical selection and the model's output, accidentally capturing the critical components of the LPL theory. He explains that the SLAM architecture can be converted into an implementation of the LPL theory simply by removing the lexical-motor weight, and more importantly, adding a further selection step at the auditory layer. Indeed, the added selection step is the crucial difference between the theories under consideration. SLAM does not include an extra selection step, so minimizing the LM weight is the best it can do (without becoming a different model) to approximate LPL. But we already demonstrated that restricting this weight for all patients (i.e., approximating LPL) leads to worse predictions (Walker & Hickok, 2015). Thus, it is clear that SLAM, as we previously implemented it, does not improve over SP by approximating the structure of LPL. We therefore considered the possibility that a different model that better represents the LPL theory might account for the same data as SLAM.

We created a new LPL model by modifying SLAM in accordance with Goldrick's suggestions: removing the lexical-motor weight[5], adding a selection step at the auditory layer, and removing the lexical input to phonemes after they are selected. This last

[5] In order to reuse our earlier code to map a 4-parameter space, the LM-weight was allowed to vary between 0 and 1e-8, then all points greater than or equal to 5e-9 were removed, as they likely represent duplicate predictions. If an LM-weight was not exactly zero in the final map, it remained several orders of magnitude smaller than the activation levels in the network.

modification implements Goldrick's assumption that post-lexical processing is a distinct stage. Quoting Goldrick (2015): "Note that this is a distinct stage of production processing in that it follows the explicit selection of an abstract phonological representation. In general, such selection mechanisms serve to reduce interactions across processing levels, increasing the degree to which distinct subprocesses can exhibit distinct patterns of impairment." Thus, by removing lexical input to phonemes after they are selected, we are implementing this theoretical position. In the LPL implementation, the *phonological* units correspond to SLAM's *auditory* units, and the *phonetic* units correspond to the *motor* units. The phonetic units can be thought of as localist representations of feature bundles, and none of the phonemes in the artificial lexicon share phonetic features. We refer to the connections as S-weights, P-weights, and PL-weights. For the LPL model to be viable, the boost of activation to each phonological unit should be large enough to successfully propagate over a further number of timesteps to produce mostly correct responses in the healthy model. We verified that delivering a boost of 150 to each phonological unit and running the model an additional 8 timesteps with the weights set at the maximum (0.04) was able to approximate the normal pattern (~97% correct). We then used the same procedures that we used previously (Walker & Hickok, in press) to fit the same patient data with our implementation of LPL, using a parameter map that included 2,321 points. As can be seen in Figure 3.3 (left), a scatterplot comparing the models' fits (RMSD) shows that LPL offers no improvements over SP: that is, fit differences are within 1SD of the fit difference between SLAM and SP. Figure 3.3 (middle) compares model fits for LPL versus SLAM. Note the cloud of points that fall outside the 1SD range; these indicate SLAM's fit advantage over LPL. Figure 3.3 (right) shows that the conduction patients are once again fit better by SLAM

compared to LPL. It is clear that SLAM significantly outperforms LPL in the same way that it outperforms SP.

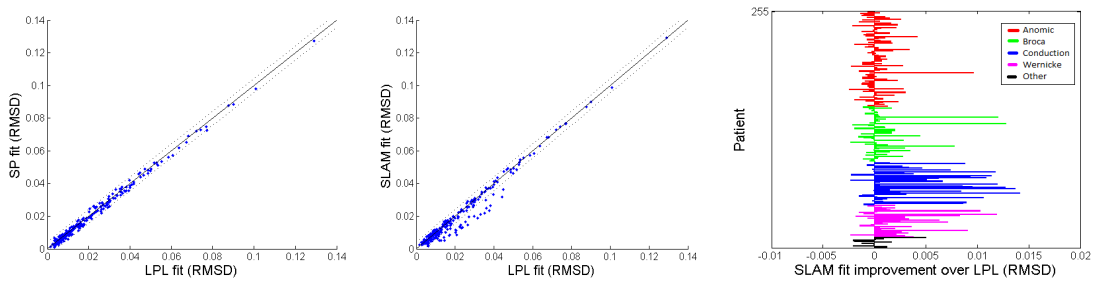

Figure 3.3. (Left) The scatterplot compares model fits (RMSD) for LPL and SP. (Middle) The scatterplot compares model fits (RMSD) for LPL and SLAM. The solid diagonal line represents equivalent fits, and the dotted lines represent 1SD of fit difference between SLAM and SP. (Right) The bar graph shows SLAM's fit improvements over LPL, grouped by aphasia type

The above simulation shows that SLAM's improvement over SP cannot be accounted for simply by the addition of another processing level. Nevertheless, Goldrick (2015) proposes a further modification of the SLAM model. He argues that SLAM's inclusion of strong feedback (following SP) renders the model empirically inadequate. Instead he promotes Rapp and Goldrick's (2000) Restricted Interaction Account (RIA). We tested this assertion computationally by adding the RIA assumptions to the LPL model implementation described above, creating LPL/RIA[6]. According to RIA, there should be no feedback from the lexical to the semantic layer, and "limited" feedback from the phonological to the lexical layer. Following Ruml, Caramazza, Shelton, and Chialant (2000), who also implemented the RIA assumptions in a computational model to fit patient data, we used a feedback attenuation value of 0.1 for the P-weights, meaning that the connection strength was 1/10

[6] We only implemented assumptions regarding feedback. Other versions of RIA have made different assumptions regarding the size of the lexicon, the implementation of damage as noise, the strength of the boosts, the number of timesteps, and an additional stage of pre-lexical, conceptual processing, which we did not address.

of the value in the reverse direction compared to the forward direction. The PL-weights remained fully interactive[7], and again lexical influences were removed during post-lexical processing. The scatterplot (Figure 3.4) shows that this LPL/RIA model makes worse predictions than the fully interactive SP model.

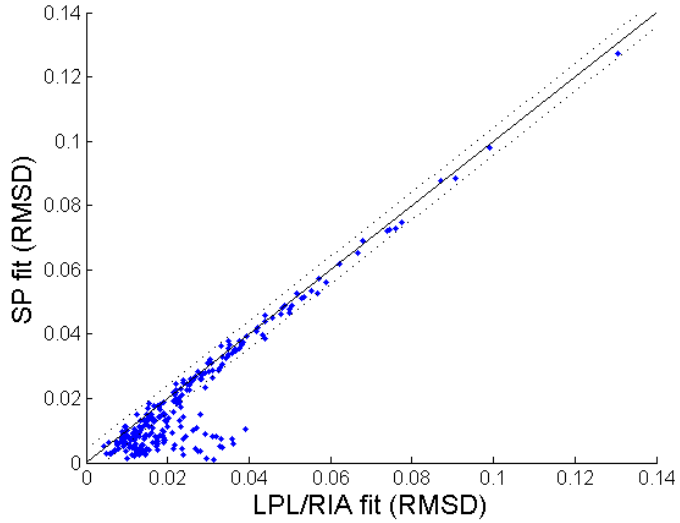

Figure 3.4. The scatterplot compares model fits (RMSD) for LPL/RIA and the (fully interactive) SP. The solid diagonal line represents equivalent fits, and the dotted lines represent 1SD of fit difference between SLAM and SP.

Finally, we hypothesized that our implementation of LPL might be able to approximate SLAM if it incorporated the same mechanism that we previously identified as driving the fit improvements: strong phonological feedback to the lexical level that influences the weak phonological feedforward link to the output layer (Walker & Hickok, 2015). The only way we were able to do this was to drop Goldrick's proposed restrictions on interaction, both with respect to the RIA assumptions and the post-lexical interaction restriction. That is, we allowed lexical representations to influence "post-lexical" processing, in order to capture the effects that SLAM suggests are occurring via interaction with the lexical level, creating

[7] The connections are 1-to-1, so the interactivity has no effect; a stronger boost could simply compensate for reduced feedback.

LPL/SLAM. To be clear, this is a test of the feedback mechanism as the explanatory factor in model improvement over SP for conduction aphasia, rather than a test of either the LPL or the SLAM models. Due to the interactivity, the lexical layer received a strong boost of feedback activation during phonological selection and, coupled with weak feedforward activation across low PL-weights, this may have the same effect as the two separate phonological sources in SLAM. Although LPL/SLAM still yields a worse fit than SP and SLAM on average, this implementation does capture many of the improvements observed with SLAM, accompanied by strong P-weights and weak PL-weights (Figure 3.5). The implication is that this mechanism, strong phonological-lexical interaction that influences weak phonological selection, can lead to improved fits for conduction aphasia naming regardless of the other theoretical or computational details of the model. The mechanism provides constraints on assumptions about interactivity of phonological and lexical processing for future models of conduction aphasia.

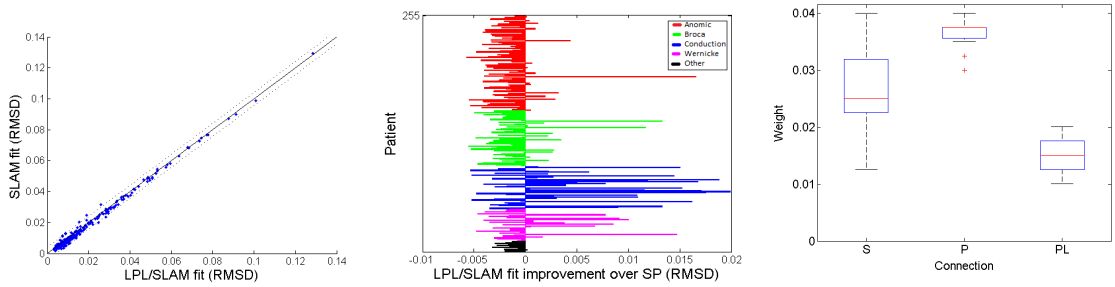

Figure 3.5. (Left) The scatterplot compares model fits (RMSD) for LPL/SLAM and SLAM. (Middle) The bar graph shows LPL/SLAM's fit improvements over SP, grouped by aphasia type. (Right) The boxplots show the weight configurations for the 19 patients with LPL/SLAM fit improvements over SP greater than .01 RMSD.

Summary

The question raised by Goldrick (2015) is whether SLAM represents an improvement over his LPL model. He argues that it offers no improvement. We have argued here that it does. First, we pointed out that SLAM implements a portion of a conceptual model (HSFC) that encompasses LPL. Second, we showed that SLAM's explanatory advantage is not a result of approximating the architectural or computational assumptions of LPL. Finally, we showed that abandoning some core theoretical assumptions of LPL—making it more like SLAM in terms of interactivity—allowed LPL to capture some of the same effects as SLAM. SLAM therefore provides new modeling constraints regarding interactions among processing levels, while also elaborating on the structure of the phonological level. We view this as evidence that an integration of psycholinguistic, neuroscience, and motor control approaches to speech production is feasible and may lead to substantial new insights (Hickok, 2014a, 2014b).

References

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, *43*(2), 182-216.

Galluzzi, C., Bureca, I., Guariglia, C., & Romani, C. (2015). Phonological simplifications, apraxia of speech and the interaction between phonological and phonetic processing. *Neuropsychologia*, *71*, 64-83.

Goldrick, M. (submitted). Integrating SLAM with existing evidence: Comment on Walker and Hickok (in press). *Psychonomic Bulletin & Review*.

Goldrick, M. (2011). Theory selection and evaluation in case series research. *Cognitive neuropsychology*, *28*(7), 451-465.

Goldrick, M., Folk, J. R., & Rapp, B. (2010). Mrs. Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language*, *62*(2), 113-134.

Goldrick, M., & Rapp, B. (2007). Lexical and post-lexical phonological representations in spoken production. *Cognition*, *102*(2), 219-260.

Hickok, G. (2000). Speech perception, conduction aphasia, and the functional neuroanatomy of language. In Y. Grodzinsky, L. Shapiro & D. Swinney (Eds.), *Language and the brain* (pp. 87-104). San Diego: Academic Press.

Hickok, G. (2001). Functional anatomy of speech perception and speech production: Psycholinguistic implications. *Journal of psycholinguistic research, 30*, 225-234.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*(2), 135-145.

Hickok, G. (2014a). The architecture of speech production and the role of the phoneme in speech processing. *Lang Cogn Process, 29*(1), 2-20.

Hickok, G. (2014b). Toward an Integrated Psycholinguistic, Neurolinguistic, Sensorimotor Framework for Speech Production. *Lang Cogn Process, 29*(1), 52-59.

Hickok, G., Erhard, P., Kassubek, J., Helms-Tillery, A. K., Naeve-Velguth, S., Strupp, J. P., et al. (2000). A functional magnetic resonance imaging study of the role of left posterior superior temporal gyrus in speech production: implications for the explanation of conduction aphasia. *Neuroscience Letters, 287*, 156-160.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron, 69*(3), 407-422.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4*, 131-138.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition, 92*, 67-99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393-402.

Mirman, D., Strauss, T. J., Brecher, A., Walker, G. M., Sobel, P., Dell, G. S., & Schwartz, M. F. (2010). A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive neuropsychology*, *27*(6), 495-504.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological review*, *107*(3), 460.

Romani, C., & Galluzzi, C. (2005). Effects of syllabic complexity in predicting accuracy of repetition and direction of errors in patients with articulatory and phonological difficulties. *Cognitive Neuropsychology*, *22*(7), 817-850.

Romani, C., Galluzzi, C., Bureca, I., & Olson, A. (2011). Effects of syllable structure in aphasic errors: Implications for a new model of speech production.*Cognitive psychology*, *62*(2), 151-192.

Ruml, W., Caramazza, A., Shelton, J. R., & Chialant, D. (2000). Testing assumptions in computational theories of aphasia. *Journal of Memory and Language*, *43*(2), 217-248.

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and language*, *54*(2), 228-264.

Walker, G. M., & Hickok, G. (2015). Bridging computational approaches to speech production: The semantic–lexical–auditory–motor model (SLAM). *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-015-0903

Chapter 4: Replication Studies

In this chapter, we examine data from a new cohort of aphasic patients. One of the goals of our modeling work is to find ways of reliably measuring latent properties of the speech production system, and to that end, we want our models to apply generally. We therefore collected new data using the same picture naming test, but with a different set of experimenters and a different set of patients, from a different research lab in a different geographical region, and recruited for treatment studies rather than basic science research. This approach is rather rare in the aphasia literature, given the substantial resources required for the collection and handling of the data. In collaboration with Dr. Julius Fridriksson and colleagues at the University of South Carolina, we were also able to analyze the patients' neuroimaging data, attempting to replicate previous localizations of lexical network components. While the modeling results are promising, the neuroanatomical analyses did not replicate, which reveals some important points in itself. We begin with a description and comparison of the data collected from the previously studied cohort (MR; Moss Rehabilitation Research Institute) and the new patient cohort (SC; University of South Carolina).

Patient Cohort Comparison

Moss Rehabilitation Research Institute Patient Database

The picture naming data studied in Chapters 2 and 3 comes from the Moss Aphasia Psycholinguistic Project Database (Mirman et al., 2010; www.mappd.org), which contains picture naming and other data from post-acute aphasic individuals in the Philadelphia region. In this chapter, data from the first Philadelphia Naming Test (PNT) administration

for all patients with a Western Aphasia Battery (WAB; Kertesz, 1982) diagnosis were included (N=275); 20 patients were added to the online database since our query for Chapter 2. Patients' responses to each of the 175 pictures are classified into one of 8 category types: correct, semantic, formal, mixed, unrelated, neologism, abstruse neologism, or omission. Modeling analyses collapse the neologism and abstruse neologism categories. The omission category is not explicitly modeled by the connectionist network; omissions are assumed to occur independently from production errors (see Dell et al., 2004 for discussion of omission errors). Semantically related responses are judged by the scorer as having a taxonomic or associative relation to the target. Phonologically related responses share the initial or final phoneme with the target, or a single phoneme in the same word position aligned from left to right, or two phonemes in any word position; that is, only a weak phonological relationship is necessary. Lenient scoring to correct for articulatory motor impairment was not applied to the data, to be more commensurate with other naming measures, like the WAB or Boston Naming Test (BNT).

The University of South Carolina Patient Database

These patients were recruited as part of a larger stroke study at the University of South Carolina, investigating single-event ischemic stroke (e.g., Basilakos, Rorden, Bonilha, Moser, & Fridriksson, 2015). Participants were excluded if they had a history of other neurological diseases, alcohol or drug addiction, or developmental language abnormality. All participants (N=95) provided informed consent in accordance with the Institutional Review Board of the University of South Carolina and underwent an extensive speech, language, and hearing assessment, as well as magnetic resonance imaging of their brain. As

part of the assessment, patients were video recorded performing the Philadelphia Naming Test. The author (G. Walker) trained two Speech-Language Pathology graduate students at the University of South Carolina to transcribe and score the naming attempts with the same protocol used by the Philadelphia research group.

Results

Comparisons of demographic, clinical, and naming measures for the SC and MR cohorts are presented below. Demographic variables include gender, race, age, and months post stroke. Clinical variables include aphasia type diagnosed from the WAB, as well as the WAB aphasia quotient (AQ) which is a measure of general aphasia severity, and the presence or absence of apraxia severe enough to interfere with naming. The naming measures that we compare are based on PNT responses, including the number of Correct, Semantic (Semantic + Mixed), Phonological (Formal + Nonword), Unrelated (Unrelated + Abstruse Neologism), and Omission errors. Tests of statistically significant differences were carried out using online statistical calculators: variables with two categories were compared with a Fisher's exact test (http://graphpad.com/quickcalcs/contingency1/), variables with more than two categories were compared with a Chi-Square test of independence (http://vassarstats.net/newcs.html), and means were compared with a t-test (http://graphpad.com/quickcalcs/ttest1/?Format=SD). Categorical variables with frequencies below 5 in either cohort were excluded from statistical analyses.

With respect to demographic variables, differences between the two patient samples were not significant for gender (Fisher's exact test, two-tailed p-value = .63), age (unpaired t-test,

two-tailed p-value = .27), or months post stroke (unpaired t-test, two-tailed p-value = .27). There was a significant difference for the racial makeup of the samples (Fisher's exact test, two-tailed p-value = .032).

Table 4.1. Gender comparison between MR and SC patient cohorts.

Gender	MR	SC
Female	118 (43%)	38 (40%)
Male	157 (57%)	57 (60%)

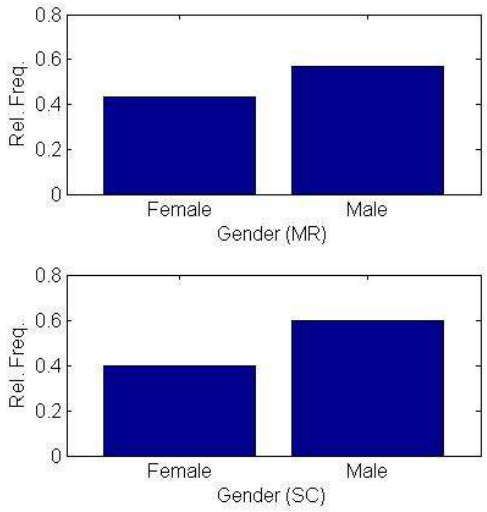

Figure 4.1. Gender comparison between MR and SC patient cohorts.

Table 4.2. Race comparison between MR and SC patient cohorts.

Race	MR	SC
Caucasian	145 (53%)	82 (86%)

African American	112 (41%)	13 (14%)
Hispanic	3 (1%)	0
Asian	1 (0.4%)	0
Unknown	14 (5%)	0

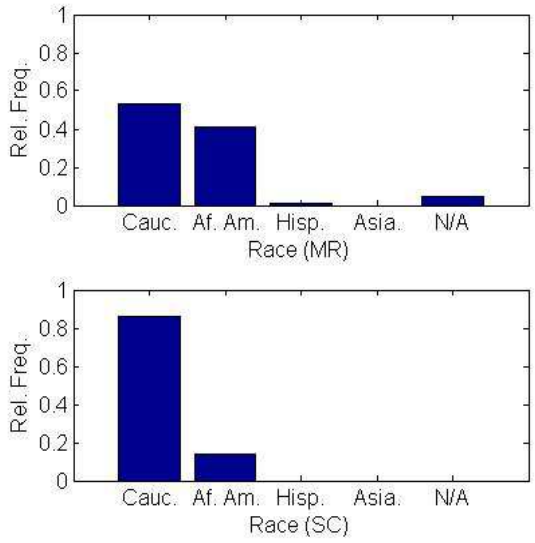

Figure 4.2. Race comparison between MR and SC patient cohorts.

Table 4.3. Age comparison between MR and SC patient cohorts.

Age	MR	SC
N	261	93
mu	58.8	60.5
sigma	13.1	11.7
min	22	36
median	59	62
max	86	83

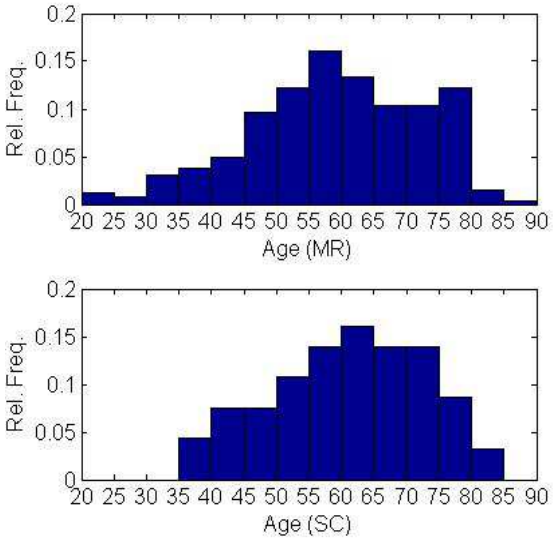

Figure 4.3. Age comparison between MR and SC patient cohorts.

Table 4.4. Months post onset comparison between MR and SC patient cohorts.

Mos. Post	MR	SC
N	261	92
mu	29.9	36.1
sigma	46.3	45.9
min	1	6
median	11	21
max	381	276

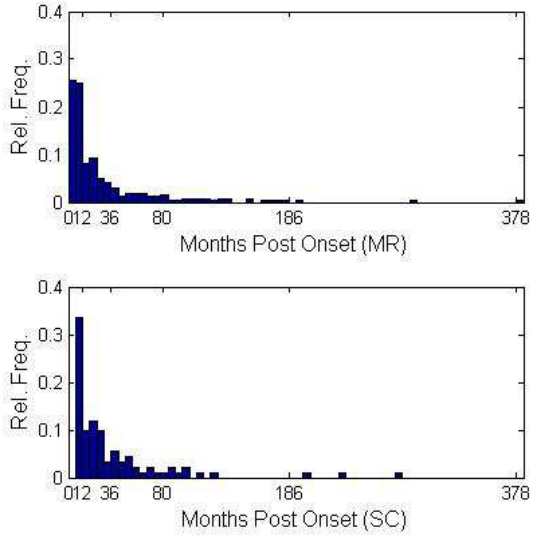

Figure 4.4. Months post onset comparison between MR and SC patient cohorts.

With respect to clinical measures, there were significant differences between the patient samples for aphasia type (Chi-Square test, p-value = .018), WAB AQ (unpaired t-test, two-tailed p-value < .0001), and the presence of apraxia of speech (Fisher's exact test, two-tailed p-value = .0008). Specifically, the SC sample had a higher proportion of Broca's and Global patients at the expense of Conduction and Wernicke's patients, along with a lower mean WAB AQ and a higher incidence of apraxia.

Table 4.5. Aphasia type comparison between MR and SC patient cohorts.

Aphasia type	MR	SC
Anomic	118 (43%)	35 (37%)
Broca	62 (23%)	32 (34%)
Conduction	48 (17%)	9 (9%)
Wernicke	38 (14%)	6 (6%)

Global	1 (0.4%)	9 (9%)
TCS	5 (2%)	0
TCM	3 (1%)	1 (1%)
Unknown	0	3 (3%)

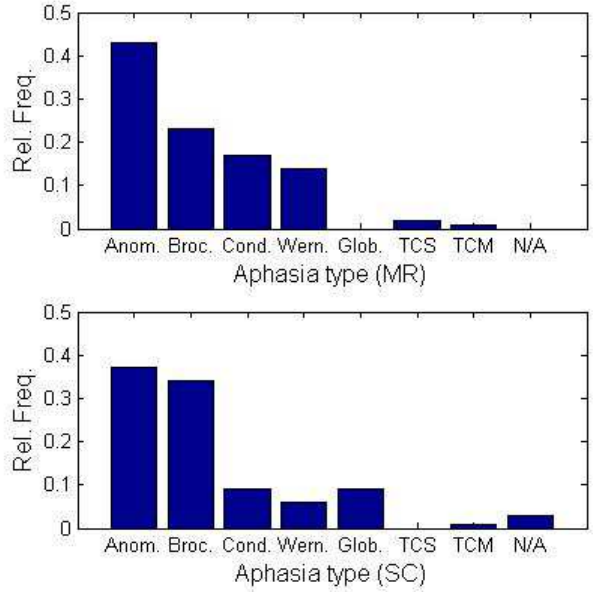

Figure 4.5. Aphasia type comparison between MR and SC patient cohorts.

Table 4.6. WAB Aphasia Quotient comparison between MR and SC patient cohorts.

WAB AQ	MR	SC
N	183	93
mu	72.7	61.5
sigma	18.0	26.4
min	25.2	8
median	76.5	67.2
max	97.9	96.5

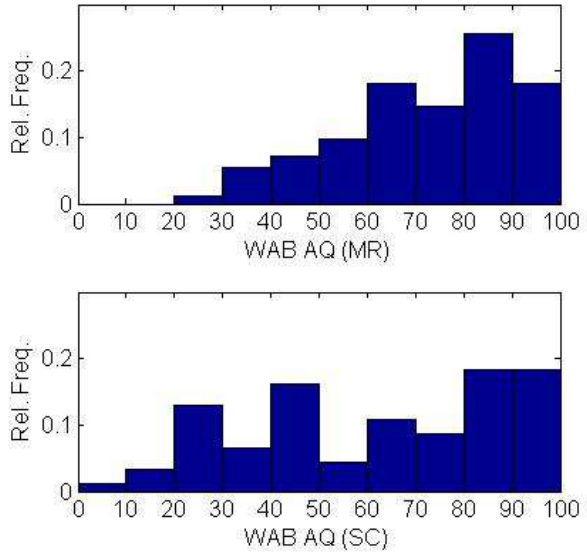

Figure 4.6. WAB Aphasia Quotient comparison between MR and SC patient cohorts.

Table 4.7. Apraxia comparison between MR and SC patient cohorts.

Apraxia	MR	SC
Present	54 (20%)	36 (38%)
Absent	221 (80%)	59 (62%)

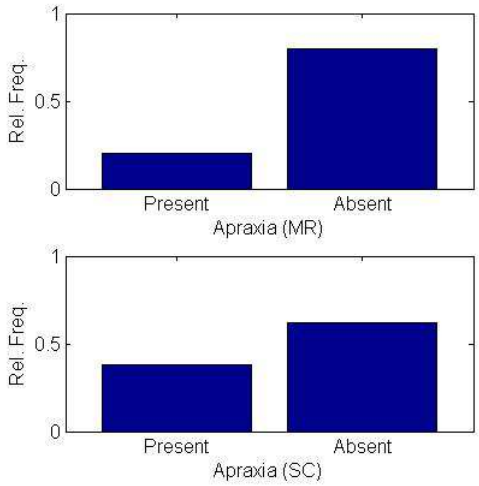

Figure 4.7. Apraxia comparison between MR and SC patient cohorts.

With respect to naming measures, the SC sample had fewer Correct responses (unpaired t-test, two-tailed p-value < .0001), more Unrelated errors (unpaired t-test, two-tailed p-value = .0033), and more Omission errors (unpaired t-test, two-tailed p-value < .0001). The MR cohort excludes patients who do not make any correct responses or who do not produce enough naming attempts, and there are 7 such patients (7.4%) in the SC cohort, but the significant differences remain even after excluding these patients. The mean differences were not significant for Semantic errors (unpaired t-test, two-tailed p-value = .70) or Phonological errors (unpaired t-test, two-tailed p-value = .14).

Table 4.8. PNT Correct comparison between MR and SC patient cohorts.

PNT C	MR	SC
mu	105.0	73.8
sigma	49.2	58.1
min	2	0
median	120	65
max	172	168

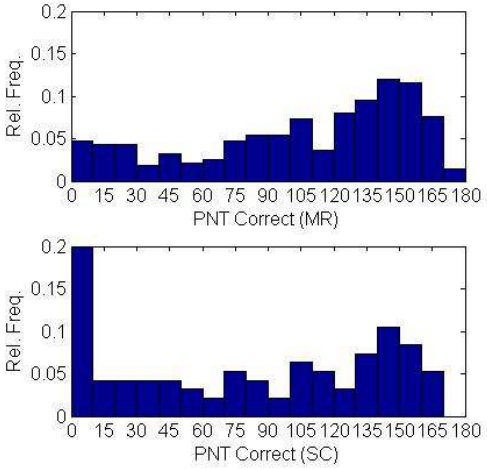

Figure 4.8. PNT Correct comparison between MR and SC patient cohorts.

Table 4.9. PNT Semantic comparison between MR and SC patient cohorts.

PNT S+M	MR	SC
mu	11.7	11.0
sigma	7.7	8.0
min	1	0
median	9	9
max	44	33

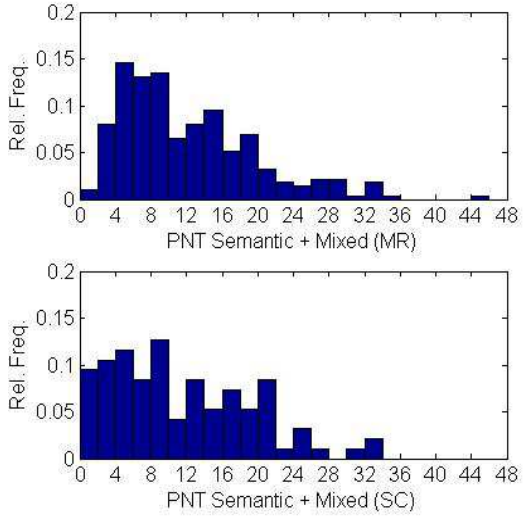

Figure 4.9. PNT Semantic comparison between MR and SC patient cohorts.

Table 4.10. PNT Phonological comparison between MR and SC patient cohorts.

PNT F+N	MR	SC
mu	24.0	28.4
sigma	23.3	30.2
min	0	0

median	16	17
max	104	119

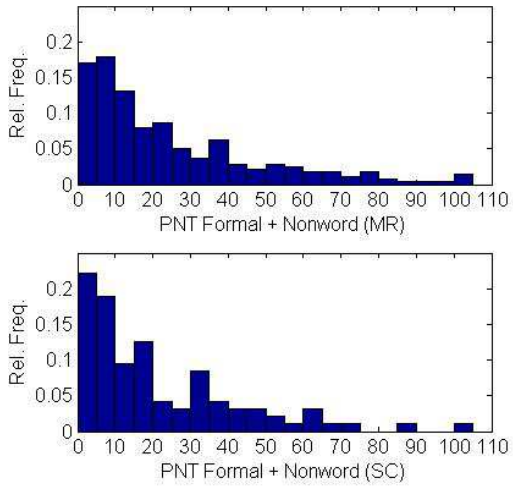

Figure 4.10. PNT Phonological comparison between MR and SC patient cohorts.

Table 4.11. PNT Unrelated comparison between MR and SC patient cohorts.

PNT U+AN	MR	SC
mu	9.9	16.8
sigma	16.9	25.9
min	0	0
median	2	6
max	84	147

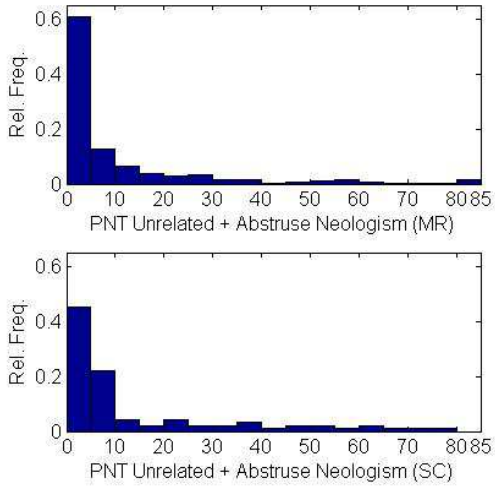

Figure 4.11. PNT Unrelated comparison between MR and SC patient cohorts.

Table 4.12. PNT Omission comparison between MR and SC patient cohorts.

PNT O	MR	SC
mu	24.5	44.9
sigma	28.7	50.6
min	0	0
median	13	20
max	148	175

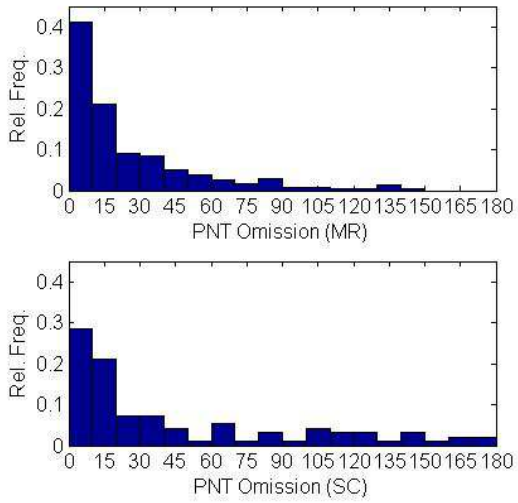

Figure 4.12. PNT Omission comparison between MR and SC patient cohorts.

Replication of SLAM Modeling Results

The patient samples exhibit some important differences with respect to their language impairments. Given that these differences emerge even in relatively large samples, it is important to determine whether our models are able to generalize to the larger population. We therefore used the same fitting procedures that we used previously (Walker & Hickok, 2015) to fit the SP and SLAM models to the new patient cohort. Six patients (2 Broca's, 4 Global) were excluded because they did not make any correct responses, and one patient (Broca's) was excluded for making only a single response that happened to be correct, leaving 88 total patients from the SC cohort for model fitting.

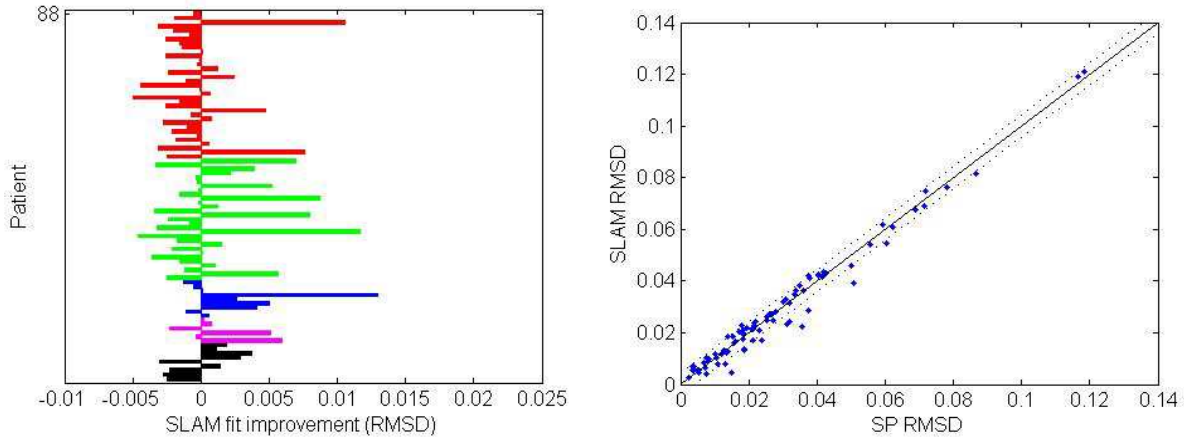

Figure 4.13. (Left) Scatterplot comparing SP model fit (abscissa) and SLAM model fit (ordinate). The solid diagonal line indicates identity, and the dotted lines indicate 1SD of fit change. 13 patients fall below the lines indicating that SLAM fit better; 3 patients fall above the lines indicating SP fit better. (Right) Bar graph showing the SLAM fit improvement, sorted by aphasia type. Red = anomic, green = Broca's, blue = Conduction, magenta = Wernicke's, and black = Other.

The modeling results in the new patient cohort reveal some important similarities with the previous analyses. SP and SLAM both fit well overall; however, with average RMSD of .0279

and .0275 respectively, there is a statistically significant increase in average model error from the MR cohort (p=.0008). Using one standard deviation of change in model fit between SLAM and SP from the previous MR analysis to define a region of practical equivalence (±0.0042 RMSD), we found that SP fit 3 patients better than SLAM, while SLAM fit 13 patients better than SP (Figure 4.13). The greatest improvement in model fit of SLAM over SP was observed in a patient with Conduction aphasia, who was assigned strong lexical-auditory connections and weak connections to the motor units (SL=.0126, LA=.0375, LM=.0076, AM=.0126). This pattern of results is consistent with our previous findings in the MR cohort; it supports the claims of the Hierarchical State Feedback Control theory, which predicts this configuration for Conduction aphasia due to a disconnection of the coordinated representations in sensory and motor cortices.

Replication of Neuroanatomical Localization Results

A popular method in cognitive neuropsychology involves localization of cognitive functions in the brain by identifying relationships between damage to specific brain regions and cognitive deficits. The approach, formally known as voxel-based lesion symptom mapping (VLSM; Bates et al., 2003), divides a 3-dimensional brain template into many small regions called voxels, and each patient's brain damage is represented in this space, typically as a binary variable indicating the presence or absence of damage in each voxel. Then, t-tests are used to compare the average symptom measures between patients with and without damage to each voxel. Because this is a mass univariate approach, corrections are applied to control for multiple comparisons, thereby revealing the locations where damage has a significant impact on symptom measures. The method has been successfully applied to the

localization of naming symptoms in the MR cohort, including semantic errors in anterior temporal lobe (Schwartz et al., 2009; Walker et al., 2011), phonological errors in the dorsal stream (Schwartz, Faseyitan, Kim, & Coslett, 2012), and SP model parameters in similar regions as the corresponding error types (Dell, Schwartz, Nozari, Faseyitan, & Coslett, 2013). We therefore attempted to localize the same cognitive processes in the SC cohort using VLSM as well.

Of the 88 patients who were included in the previous modeling analysis, 83 had lesion masks available for anatomical investigation. Details about the creation of lesion masks, such as MRI data acquisition, lesion segmentation, or registration to standard space, can be found in Basilakos et al. (2015). It is worth noting that all of these details are different from the MR cohort, and while the different lesion segmentation methods should converge on similar results, the variance in methods across research labs remains a possible source of disagreement. We used the Nii_Stat software packages in Matlab for statistical analysis of neuroimaging data, and we used MRIcron for visualization. For the VLSM analysis, we included voxels that were damaged in at least 5 patients, and we used a False Discovery Rate (FDR; Genovese, Lazar, & Nichols, 2002) threshold of q=.05 to identify voxels with a significant relationship to symptom measures. There is no conventional level of significance for FDR, and while .05 is frequently used in the literature, it is a fairly relaxed criterion. Methodology for analyzing neuroimaging data is an active area of research that is progressing quickly; we chose to use the FDR threshold mainly for comparison with the earlier work. We examined 8 different naming measures: counts of correct responses,

semantic only, semantic + mixed, nonword + abstruse neologism, formal + nonword, omissions, and the S-weight and P-weight parameters from the SP model.

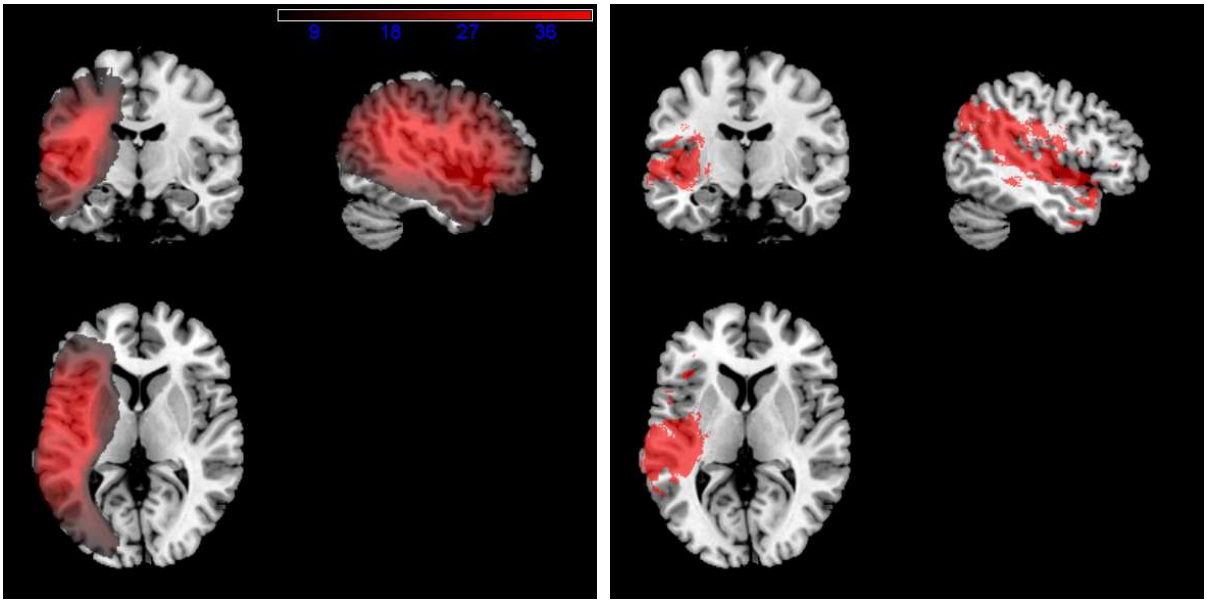

Figure 4.14. (Left) The lesion overlap distribution for 83 patients from the SC cohort. The sagittal, coronal, and axial slices of the ch2better template are shown at coordinates x=44, y=116, z=77. The color indicates the number of patients damaged in the voxel, from 5 (black) to 40 (red). The peak voxel had 57 patients. (Right) Voxels where damage is significantly associated with decreased naming accuracy (Z<-3.25). The template coordinates are the same as the left of the figure.

There were 429,957 voxels that were lesioned in at least 5 patients, with a maximum overlap of 57 patients in 2 voxels (x=42, y=104, z=96 and 98) in the left precentral gyrus. Figure 4.14 (left) shows the lesion overlap distribution. The only significant relationship between lesion status and symptom measure was for overall accuracy (Figure 4.14, right), yielding 67,703 voxels with Z-values below the critical value of -3.25, encompassing most of the left perisylvian region. None of the voxels survived a post hoc analysis in which accuracy was regressed against lesion volume. We thus failed to replicate the published localization results that were found in the MR cohort.

There are several possible reasons for the failure to replicate. As mentioned above, the lesion segmentation methods differ; the methodology is still an active area of research, so it is not likely to be standardized across labs any time soon. Nevertheless, the differences in segmentation procedures should converge on similar results, so this is unlikely to be the major source of contention. The fact that the samples differ with respect to their language measures, in particular, more severe (Broca's) aphasia with a higher incidence of apraxia in the SC cohort, may have impeded the anatomical localization of damaged functions. Localization requires heterogeneous lesions; if there is not enough variance in the type of damage to the lexical network in the sample, the data will not be able to reveal the consequences of circumscribed injury. Since the only significant association with lesion status was overall accuracy, and this was confounded by lesion volume, it may be the case that the sample of SC patients were more homogenous than the MR cohort with respect to the location of their damage within the lexical network. The lesion data from the MR cohort was not directly available for comparison, so this remains a speculation. Additionally, we can only speculate about what is causing the differences between the samples, though it stands to reason that the offer of treatment during recruitment may convince more severe patients to participate.

Although we did not replicate the VLSM analyses, we learned some things that will be relevant as we move forward with our research on aphasia. The heterogeneity of the aphasic syndromes can hardly be understated; even in relatively large patient samples, clear differences can emerge with respect to speech and brain functions. Nevertheless, the two-step models of lexical retrieval were still able to fit this new patient cohort's data quite

well, and the SLAM model seemed to offer similar improvements over the SP model, supporting the generalizability of our modeling efforts.

References

Basilakos, A., Rorden, C., Bonilha, L., Moser, D., & Fridriksson, J. (2015). Patterns of Poststroke Brain Damage That Predict Speech Production Errors in Apraxia of Speech and Aphasia Dissociate. *Stroke*, *46*(6), 1561-1566.

Bates, E., et al. (2003). Voxel-based lesion–symptom mapping. *Nature neuroscience, 6*(5), 448-450.

Dell, G. S., Schwartz, M. F., Nozari, N., Faseyitan, O., & Coslett, H. B. (2013). Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition*, *128*(3), 380-396.

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, *15*(4), 870-878.

Kertesz, A. (1982). *Western aphasia battery test manual*. Psychological Corp.

Schwartz, M. F., Faseyitan, O., Kim, J., & Coslett, H. B. (2012). The dorsal stream contribution to phonological retrieval in object naming. *Brain*, *135*(12), 3799-3814.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Faseyitan, O., Brecher, A., Dell, G. S., & Coslett, H. B. (2009). Anterior temporal involvement in semantic word retrieval: voxel-based lesion-symptom mapping evidence from aphasia. *Brain, 132*(12), 3411-4427.

Walker, G. M., & Hickok, G. (2015). Bridging computational approaches to speech production: The semantic–lexical–auditory–motor model (SLAM). *Psychonomic Bulletin & Review*, 1-14.

Walker, G. M., Schwartz, M. F., Kimberg, D. Y., Faseyitan, O., Brecher, A., Dell, G. S., & Coslett, H. B. (2011). Support for anterior temporal involvement in semantic error production in aphasia: new evidence from VLSM. *Brain and Language*, *117*(3), 110-122.

CHAPTER 5: A Bayesian Approach to Connectionist Model Fitting

In this chapter, we adopt a Bayesian approach to evaluating a connectionist model of lexical retrieval (Foygel & Dell, 2000), by first reformulating the cognitive model as a statistical model, then applying cross-validation methods which are considered the gold-standard for statistical model assessment. The probabilistic approach enables (i) the quantification of uncertainty in parameter estimates due to limited observations, (ii) formal model comparisons, (iii) clear semantic interpretations of statistical constructs, and (iv) proven methods for estimating parameters of complex, hierarchical models. We begin by briefly reviewing the critical assumptions of both the connectionist cognitive model and the multinomial statistical model, then explain the details of the probabilistic approach to model fitting and assessment, and finally demonstrate its application.

The Connectionist Cognitive Model

Connectionist models are a useful mathematical tool for studying the cognitive processes underlying word production. These models quantify the activation levels of interconnected nodes (or units), which evolve over time as activation decays and spreads along the network's connections. A mental representation can be associated with a single node or a group of nodes, and representations can compete with one another for selection via their activation levels.

The two-step, interactive model of lexical retrieval was proposed by Foygel and Dell (2000) to explain patterns of word production errors in aphasia. The model simulates a small, 3-layer, lexical network, and the pattern of connections between representational levels

(semantic, lexical, phonological) is meant to approximate the statistical structure of the English lexicon. The connection pattern is fixed, so the model does not simulate learning; however, the strength of connections between representational levels are free parameters that can be reduced to simulate a damaged network, while the maximum connection strength simulates the speech production patterns for healthy adults. This instantiates an important theoretical assumption: aphasic errors exist on a continuum between well-formed speech and random output. Additionally, the connections in the model are bidirectional, enabling both semantic and phonological influences on lexical selection. The two retrieval steps in the model involve selecting a lexical unit first, and then selecting the phoneme units. Errors can occur during either of these steps, because each unit's activation includes noise that is both intrinsic and activation-dependent. Visual processing is assumed to always be successful, so simulations of a picture naming attempt begin with a boost of activation to the target word's semantic units. Activation flows through the network for a number of time steps, and then the most active lexical unit is selected to receive a boost of activation. Activation continues to flow through the network for another number of time steps, and then the most active phonological units—grouped by onset, vowel, and coda—are selected for production. The output of the model is scored as correct, semantically related, phonologically related, mixed relation, unrelated, or not a word (nonword or neologism), producing a categorical distribution of response types. It is this categorical distribution that we use to tie the cognitive model to a statistical model.

In fitting the connectionist model to a person's picture naming data, the connection strengths that allow the model to best approximate the person's data pattern are

estimated. In previous work (e.g., Chapters 2 & 3), these connection strengths were estimated by simulating many naming attempts at many parameter settings and selecting the setting that comes closest to producing the subject's data pattern. While this simulation approach has led to significant insights, identifying only point estimates of model parameters precludes use of the wide variety of modeling tools developed in probability theory. Instead of focusing on a single, most likely parameter value, probability theory considers the likelihood of any possible parameter value. These tools therefore require a likelihood function: a mathematical expression relating any given set of model parameters to the probability of observed data.

The Multinomial Statistical Model

A likelihood function returns the probability of observing data given a set of model parameters, which, in the case of our cognitive model, are connection strengths. Once this function has been defined, we can use Bayes' theorem to update our prior beliefs about the connection strengths, using observed data to estimate the likelihood of any set of parameters, yielding a posterior distribution. These posterior distributions can be examined to answer questions about the parameters; for example, which parameter values are the most likely to have produced the data? The answer should be similar to the point estimates obtained via simulation searches, but there are further questions that can be answered regarding the model's quality using the likelihood function, by providing a range of plausible parameter values, for example.

The first step in defining the likelihood function is to consider the type of data we are observing. In this case, we collect counts of each response type, so we assume the data comes from a multinomial distribution. The multinomial distribution assumes that each naming attempt is independent and identically distributed, and although this assumption is demonstrably false (p<.0001, Smith & Batchelder, 2008, permutation test for homogeneity of items and subjects), the cognitive model also makes the same simplifying assumption. The likelihood (probability mass) function for a multinomial distribution[8],

$$f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k)$$

$$= \begin{cases} \dfrac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, & \text{when } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise,} \end{cases}$$

requires specification of the probability of each response type ($p_1, \ldots, p_k$), and the total number of responses (n). Because there are six naming response types, and their probabilities must sum to one, we need to derive five independent probabilities from the two connection strength parameters. We do this by realizing that the response probabilities depend critically on two functions that are specified by the network: the activation function and the selection function.

The activation function describes the activation of a network unit over time. The activation (A) of a unit (i) at a given time (t) depends on its previous activation combined with a decay term (D), the activations (A^*) of its connected units (j) combined with their connection strengths (W), and both intrinsic and activation-dependent noise (ε). Noise is applied after

[8] Image from https://en.wikipedia.org/wiki/Multinomial_distribution

updating the decay and input, which can lead to negative activation levels, but negative activation does not spread.

$$A_j^* = \begin{cases} A_j & \text{if } A_j \geq 0 \\ 0 & \text{if } A_j < 0 \end{cases}$$

$$A_{i,t}^- = A_{i,t-1} \times (1 - D) + \sum_{j=1}^{J} (A_{j,t-1}^* \times W_{i,j})$$

$$A_{i,t} = A_{i,t}^- + \varepsilon_{\text{in}} + (A_{i,t}^- \times \varepsilon_{\text{act}})$$

The key insight for redefining the cognitive model as a statistical model is realizing that the activation noise is normally distributed, and we can therefore describe the probability of a unit's activation at a given time with the parameters of a normal distribution. The mean (μ) is centered on the activation value based solely on the decay and input activation terms, while the standard deviation (σ) accounts for the noise terms. Because negative activation is prevented from spreading, the input activation is represented with a truncated normal mean (μ^*) and standard deviation (σ^*).

$$\mu_{i,t} = \mu_{i,t-1} \times (1 - D) + \sum_{j=1}^{J} (\mu_{j,t-1}^* \times W_{i,j})$$

$$\sigma_{i,t} = \sigma_{i,t-1} \times (1 - D)^2 + \sum_{j=1}^{J} (\sigma_{j,t-1}^* \times W_{i,j}^2) + {\sigma_{\text{in}}}^2 + (\mu_{i,t} \times \sigma_{\text{act}})^2$$

Now, at each timestep, a unit has a distribution of possible activation levels, represented by μ and σ, rather than just a single activation level, represented by A. The parameters of the

activation distribution for each node thus evolve over several time steps until the selection function is required.

The selection function takes the activation means and variances of competing units, and returns the probability of selecting a given unit over its competitors. Selection is simply based on having the greatest activation, so we can calculate the probability that a normal random variable is greater than a set of competing normal random variables with the following integral. The probability of selecting a unit (i) over competing units (k) is

$$P(S_i) = \int_{-\infty}^{\infty} \phi(\mu_i, \sigma_i, \alpha) \prod_{k=1}^{K} \Phi(\frac{\alpha - \mu_k}{\sigma_k}) \mathrm{d}\alpha$$

where $\phi(\mu, \sigma, \alpha)$ is the PDF of the normal distribution with mean μ and standard deviation σ at point α, and $\Phi(x)$ is the standard normal CDF. Because there is no closed-form solution, this integral is evaluated numerically for each of the competing units, and then the approximated probabilities are normalized to sum to one.

After defining the activation function and the selection function, it is possible to define the multinomial likelihood function by calculating the probability of each response type. Given a set of connection strength parameters, the probability of selecting each lexical unit is calculated after iteratively updating the network's activation means and standard deviations for a number of time steps. Then, for each possible lexical selection, the probability of selecting each phonological unit is calculated, similarly updating the parameters over a number of time steps after delivering the corresponding lexical boost.

For each phoneme, the selection probabilities are then combined as a weighted average of their lexical selection probabilities. Finally, the probability of each response is calculated as the joint probability of its constituent phonemes being selected. The entire set of calculations are performed separately for lexical neighborhoods with and without mixed error opportunities, and the response probabilities are combined as a weighted average of the frequency of mixed error neighborhoods. This function thus takes a given set of connection strength parameters and returns the probability of each response type, satisfying the requirements for a multinomial likelihood function.

With the likelihood function defined, the model's parameters can then be fit to an individual's data. Next, we demonstrate the use of the likelihood function in two modeling studies of the picture naming data from the MR cohort (Chapter 4): a validation study in which all of the data is used to estimate the most likely connection parameters to compare with parameter estimates from simulation searches, and a cross-validation study in which a subset of the data is used to train the model, and then the trained model is used to predict an independent data set.

Validation Study

We implemented the statistical model and parameter estimation using JAGS (Plummer, 2003). The Bayesian approach to parameter estimation relies on Gibbs sampling, a Monte Carlo Markov Chain procedure. We use naive prior densities, assuming that all possible values are equally likely for a given connection parameter. Beginning with a beta(1,1) (i.e., uniform) prior distribution for each of the parameters, a random starting set of values is

chosen (i.e., numbers between 0 and 1), and they are linearly transformed to the S-weight and P-weight scale ranging [0.0001, 0.04]. The likelihood of the data is then calculated using the given connection strengths and the likelihood function. Then, a new parameter value is randomly selected from the prior distribution, and the new likelihood is compared with the previous one. The new parameter value is accepted or rejected based on the ratio of these likelihoods, and the process repeats for each parameter until an arbitrary number of samples have been collected. Under fairly broad conditions, the chain of samples is guaranteed to converge to the most likely distribution of parameters, eventually. Multiple chains that start at different, random points can be run to test for convergence. Typically, the initial set of samples from the chains prior to convergence are discarded as burn-in. The remaining samples can be combined to form a posterior distribution that describes the likelihood of the parameter values. We take the expected value (mean) of the posterior distribution as a point estimate of the most likely parameter value. A 95% credible interval (CI) contains 95% of the samples around the center of the posterior distribution, providing a range of credible parameter values (Lee & Wagenmakers, 2014).

An alternative approach to estimating parameters involves simulating many naming attempts at many parameter settings, and then selecting the parameter setting that generates the most similar pattern to the observed data (minimizing RMSD), resulting in a point estimate (see Chapter 2, for details). For comparison with the Bayesian estimates, the S-weight and P-weight parameters were also estimated for each patient using the simulation method, with 57,011 points (parameter value sets) and 10,000 naming

simulations at each point. It is worth noting that the minimization of RMSD versus maximization of likelihood is expected to lead to slightly different results.

To estimate the connection strength parameters, S-weight and P-weight, for each patient, 4 chains were run, with 500 samples each. Visual inspection of the chains indicated convergence after approximately 25 samples, so the first 50 samples were discarded as burn-in. Combining chains resulted in 1,800 samples of the posterior distribution for the S-weight and P-weight parameters. From these posterior distributions, we obtained point estimates and credible intervals of the parameters for each patient. The sampling chains of the S-weight and P-weight parameters for an example patient are shown below (Figure 5.1), with each of the 4 chains plotted in a different color. The plots demonstrate convergence and good mixing.

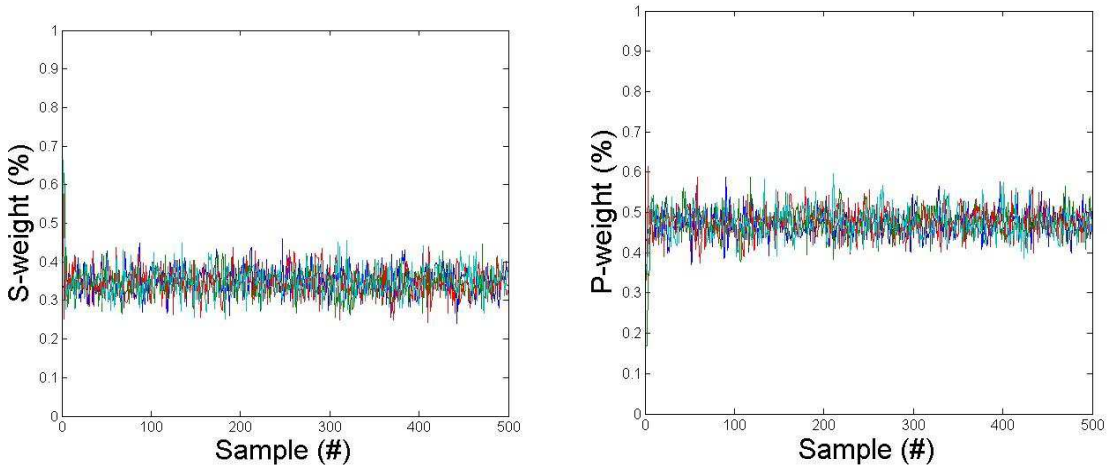

Figure 5.1. Chains of parameter samples showing convergence and mixing.

The posterior mean S-weight accounted for 94.9% of the variance (R^2, assuming y = x) in the simulation point estimates across all patients , and the posterior mean P-weight accounted for 95.7% of the variance (Figure 5.2).

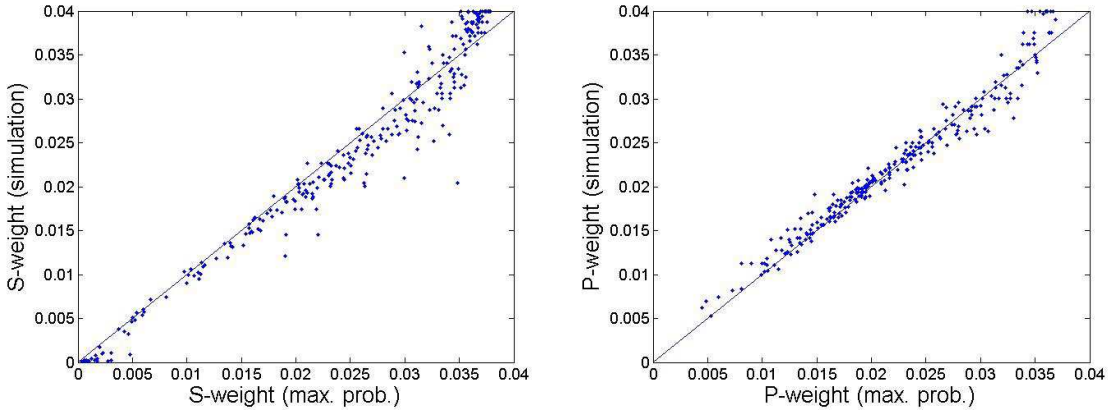

Figure 5.2. Scatterplots comparing parameter estimates from the Bayesian approach (abscissa) with parameter estimates from the simulation search approach (ordinate). The diagonal line indicates identity.

The results could be interpreted as validating our implementation of the likelihood function, demonstrating that our probabilistic point estimates are consistent with the simulation method used in previous work. The results could also be viewed as validating the earlier simulation method, given that our probabilistic estimates are derived directly from the mathematical structure of the model and are more stable. One of the additional benefits of the probabilistic approach is the ease with which it quantifies uncertainty in the parameter estimates.

We use the CI half-width as a statistic to quantify the precision of the parameter estimates obtained from the PNT. For the entire patient sample, the average CI half-width for S-weight was 11.7%, and the average CI half-width for P-weight was 10.5%. Generally speaking, this means that, for a patient with a parameter point estimate of 50%, the parameter can reasonably be expected to lie somewhere between 40-60%. For individual patients, the S-weight CI half-widths ranged [1.1% , 33.4%] with a standard deviation of 4.4%, and the P-weight CI half-widths ranged [3.9% , 19.8%] with a standard deviation of

4.0% for P-weight. Data from an example patient is shown below, along with posterior densities for the corresponding parameter estimates.

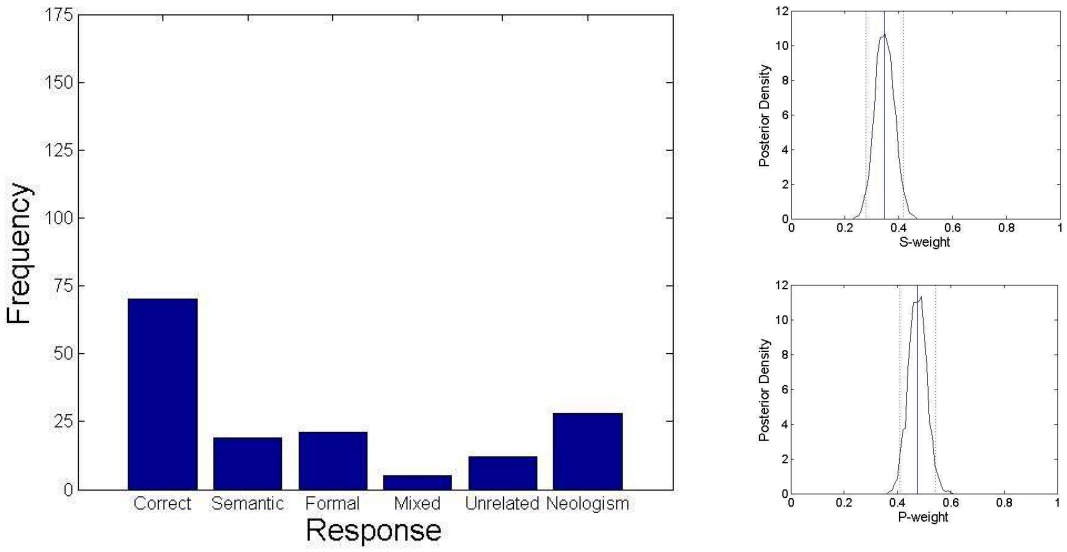

Figure 5.3. Response distribution and posterior densities for an example patient. The posterior mean is indicated by the solid vertical line, and the CI is indicated by the dotted vertical lines.

We seem to be able to make fairly precise inferences about what the most likely model parameters might be for a given set of observed data, but is the model accurately capturing the underlying patterns in the data? Next, we turn to the predictive value of the model.

Cross-validation Study

A cross-validation study uses a subset of the available data to estimate model parameters, and then predictions are generated and tested against an independent data set. Because we have 175 naming attempts for each patient, we decided to train the model using ⅗ of the data. Although we used data from 105 items, omission errors are not explicitly modeled, and therefore *n* is the sum of all non-omission responses. Effectively, omission errors do

not affect the parameter point estimates, but they do increase the credible intervals. The choice of which items to include in the training set is not a trivial matter given the heterogeneity of the items. To determine which items to include in the training set, we randomly generated 9,999 unique ways of partitioning the data by ⅗, and we identified the partition that yielded the median average absolute difference for percent accuracy. That is, for each partitioning of the data, we calculated the absolute difference in percent accuracy between the training and testing sets for each patient. These absolute differences were averaged across patients to produce a value representing the overall balance of difficulty between the training and testing sets for that partition. The partitions were then ordered by this value, and the 5,000th partition was selected, so that the partitioning of the items did not unduly affect the cross-validation results.

As before, we used Gibbs sampling with 4 chains of 500 samples to estimate the S-weight and P-weight parameters given the training data. For each sample of parameters, we also randomly generated a new multinomial distribution (n = the sum of the non-omission responses in the testing set) to serve as a prediction for the testing data, creating a posterior predictive distribution for each response type. We took the mode of the posterior predictive distribution as a point estimate. We evaluated 3 further statistics from these distributions: 1) the average RMSD between response proportions in the testing set and point estimates for each patient, 2) the variance in the testing set that is accounted for by point estimates (R^2, using the identity line and the best-fit regression line) for each response type across all patients, and 3) the posterior predictive likelihood of the observed responses. The posterior predictive likelihood is the probability that the model will

correctly predict the observed data. We estimated this probability using the proportion of samples from the posterior predictive distribution that equal the observed counts in the testing data. We used thresholds (α) of .01 and .05 for the posterior predictive checks; i.e., to pass, the model must successfully predict the testing data better than 1 out of 100 (or 1 out of 20) times. Figure 5.4 shows posterior predictive distributions for two example participants; the model fails the posterior predictive check for the first participant (Correct, Semantic, and Mixed, $p<.01$) and succeeds for the second.

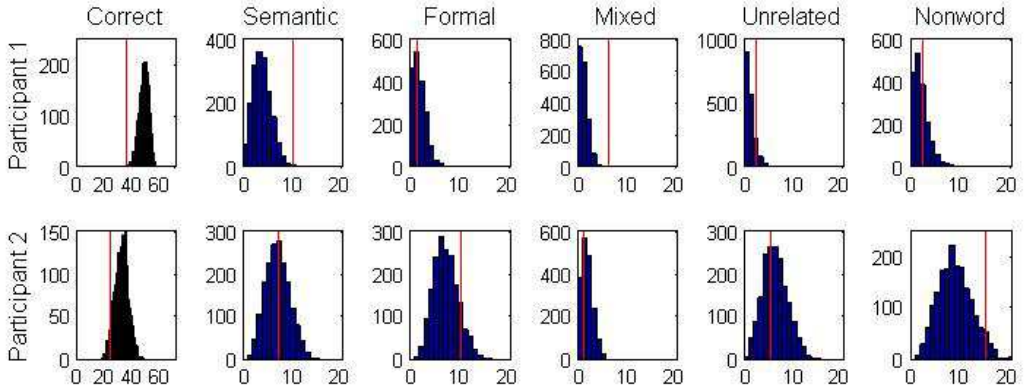

Figure 5.4. Posterior predictive distributions for each response type from two example participants. The abscissa represents the number of responses of a particular type, and the ordinate represents the number of samples (out of 1800) for which that number was predicted. The vertical red line indicates the actual number of responses in the testing set.

The RMSD statistic was calculated for each patient. It can be interpreted as the expected error between point estimates of response proportions and observed proportions of each response type, i.e., the average distance between the peaks of the posterior predictive distributions and the red lines in Figure 5.4 (expressed as proportions instead of counts, for comparison with previous work). The average RMSD across all patients was .044 (ranging .0004 to .165), which is significantly larger than the estimated error from the simulation searches (Chapters 2 & 4), as expected due to the prediction of independent

data sets. Point estimates accounted for variance in the testing data very well for Correct and Nonword responses, moderately well for Formal and Unrelated responses, and poorly for Semantic and Mixed responses (Table 5.1). Because Semantic and Mixed errors are less frequent than other responses, they have less variance to be explained.

Table 5.1. The proportion of variance in the testing data accounted for by model point estimates.

Response	Correct	Semantic	Formal	Mixed	Unrelated	Nonword
R^2 (identity)	.968	.156	.542	-.516	.697	.907
R^2 (best fit)	.968	.204	.605	.062	.828	.924

Passing the posterior predictive check simply means that the observed data is reasonably consistent with the model's predictions, where 'reasonably consistent' is defined by α; if we consider the model as a null hypothesis, there would not be enough evidence to reject it. There were 219 patients (79.6%) for whom the model passed the posterior predictive check on all response types at the α=.01 level, and there were 127 patients (46.2%) for whom the model passed at the α=.05 level on all response types. Table 5.2 shows the proportion of patients that passed the posterior predictive checks for each response type.

Table 5.2. The proportion of patients for whom the posterior predictive likelihood of the testing data is greater than or equal to α.

Response	Correct	Semantic	Formal	Mixed	Unrelated	Nonword
Passed check (α=.01)	.960	.960	.946	.942	.953	.996
Passed check (α=.05)	.796	.898	.836	.876	.913	.924

Typically when the model failed for a patient, it was on a single response type; this was the case for 71.6% of the failures in the α=.05 test. In the cases where the model fails, there may be impairments outside of the modeled lexical system, which would mean that even if the model parameters are accurate, they would still not yield accurate predictions. Alternatively, the model's simplifying assumptions or the use of limited sampling chains may have led it astray. Nevertheless, in this large sample of aphasic patients, the assumptions of the two-step lexical retrieval model were able to capture many of the underlying patterns in the data.

To summarize, we reformulated a cognitive connectionist model as a statistical multinomial model, placing the cognitive model on a solid statistical foundation. We applied Bayesian inference techniques for parameter estimation and model assessment. Our validation study confirmed that our likelihood function is instantiated properly and yields similar results as previous simulation searches. Furthermore, we were able to assess the precision of our parameter estimates, finding them to be generally adequate. Finally, our cross-validation study provided a stringent test of the model's predictive value, finding that the model made adequate predictions for a large portion of the data. Although much more work remains in our modeling efforts, the results presented here suggest some very encouraging ways forward.

References

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, *43*(2), 182-216.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 125). Technische Universit at Wien.

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*(4), 713-731.

CHAPTER 6: Modeling More Speech Production Tasks

In this chapter, we use the connectionist networks to simulate speech repetition tasks in addition to naming, under the assumption that they share some of the same speech production components. As reviewed in Chapter 1, this approach already has precedent in the literature (Dell, Martin, & Schwartz, 2007; Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Dell, Schwartz, Nozari, Faseyitan, & Coslett, 2013; Hanley, Dell, Kay, & Baron, 2004; Martin, Dell, Saffran, & Schwartz, 1994; Nozari, Kittredge, Dell, & Schwartz, 2010). Here, we re-evaluate the SP model for word repetition using our new Bayesian method for estimating the model's parameters and evaluating the model's predictions. We also simulate word and nonword repetition with SLAM, similarly investigating whether the assumption of shared network components enables predictions across tasks.

The SP Model of Repetition

Dell et al. (1997) suggested that their model of picture naming should also be able to successfully predict speech repetition, because both tasks require the retrieval and assembly of phoneme sequences. Repetition does not require a semantic processing component, and so this task can be modeled simply as the second step of naming, under the assumption that perfect auditory recognition can directly activate the lexical layer of the model. This model has been called the single lexical-route model, because it effectively assumes that there is only one connection between auditory inputs and speech outputs, via the lexical system. An alternative version, called the dual-route model, assumes a second direct connection between auditory inputs and phonological outputs (Hanley et al., 2004; Nozari et al., 2010). Like the lexical-route model, perfect auditory recognition activates the

lexical layer, but also simultaneously activates phoneme units via a non-lexical (nl) connection. In both models, s-weights and p-weights are estimated using picture naming data, while the dual-route model additionally uses nonword repetition data to estimate the nl-weight. Once these weights have been fixed, word repetition is predicted without the use of any free parameters.

Dell et al. (2007) directly compared the lexical-route and dual-route repetition models using data from 30 aphasic patients. To adhere to the perfect recognition assumption, this cohort excluded patients with excessively low scores on 5 tests of auditory input processing, although the threshold was biased toward including patients since z-scores were standardized with patient data rather than healthy controls. Comparing their model predictions of repetition response proportions to the actual data, they found average rmsd values of .046 and .055 for the dual-route and lexical-route models, respectively, although this difference was not significant. When patients had high naming accuracy, both models predicted high repetition accuracy; however, the lexical-route model underpredicted repetition scores for a subset of patients. The dual-route model specifically improved fits for 4 patients whose repetition performance was much worse than expected according to the lexical-route model, but in general, the dual-route model tended to overpredict repetition accuracy. According to the rmsd measure of fit, 19 patients were fit better by the lexical-route model, 9 were fit better by the dual-route model, and there were 2 ties. Given these somewhat ambiguous results, we decided to apply our Bayesian analysis framework to re-evaluate the single lexical-route model alongside our own dual-route model of repetition implemented by SLAM.

We began by examining data from the 103 patients studied by Dell et al. (2013), looking at 6 behavioral measures available in the MAPPD online database: picture naming, word repetition, nonword repetition, phonological discrimination, and auditory lexical decision between words and pseudowords. The details of these tests are provided elsewhere (e.g., Dell et al., 2007). Word repetition is measured with the Philadelphia Repetition Test, which includes the same 175 items from the naming test; pre-recorded audio stimuli are presented to participants over computer speakers. Word repetition responses are scored in the same manner as naming responses: Correct, Semantic, Formal, Mixed, Unrelated, Nonword, or Omission. The nonword repetition test includes 60 items derived from the naming targets by changing 2 phonemes. The responses are scored as Correct, Formal, or Other (this data provided by A. Brecher of Moss Rehabilitation Research Institute, as the error type data was not available online). The majority of errors in the Other category are nonwords, but it also includes a small proportion of omissions. Phonological discrimination tests the ability to detect a single phoneme difference between two words or pseudowords presented consecutively without delay (20 pairs), while lexical decision tests the ability to recognize words and reject pseudowords (80 of each). We used a high threshold to identify a group of 28 patients with unimpaired auditory input processing: 90% or above on phonological discrimination and both lexical decision measures. It is likely that some of the same participants from the Dell et al. (2007) study are also included here.

In calculating the likelihood of picture naming data, we already calculated the probability of each response type given a correct lexical selection. Because the SP model assumes perfect

auditory recognition, these probabilities can be used to generate a posterior predictive distribution for repetition responses directly, given a certain number of attempts. We used JAGS to re-estimate the parameters that fit the picture naming data, and for each parameter sample, we generated a prediction for word repetition using a multinomial distribution, parameterized with the response probabilities following correct lexical selection and the number of repetition attempts. We took the modes of the posterior predictive distributions as point estimates. As in Chapter 5, we examined the rmsd between point estimates and observed response proportions, the variance accounted for by point estimates, and the posterior predictive likelihood of the observed response frequencies.

The average rmsd for the SP point estimates was .037, ranging 0 to .213; this is smaller than the average rmsd reported by Dell et al. (2007), though it is unclear whether this difference is due to sampling error or model fitting error. The overall pattern of results was otherwise consistent with Dell et al. (2007): the model correctly predicted that repetition errors are mostly Formal or Nonword, while Semantic, Mixed, and Unrelated are extremely rare. Given the rarity of these responses, the variance accounted for (R^2) is difficult or impossible to interpret; not a single patient made a Semantic error, so there was no variance to explain. A negative value for variance explained indicates that the sample average is a better predictor than our model point estimates, which can be driven by the presence of outliers; the best-fit line effectively adds 2 more parameters to correct for this possibility.

Table 6.1. The proportion of variance in the word repetition data accounted for by model point estimates.

Response	Correct	Semantic	Formal	Mixed	Unrelated	Nonword
R^2 (identity)	-.689	NA	.040	-10.4	.462	-.590
R^2 (best fit)	.660	NA	.405	.051	.482	.665

Given that the same lexical network is used for naming and repetition though, it is logically possible to produce Semantic errors during repetition (a rare disorder known as deep dysphasia is characterized by this symptom), so it is noteworthy that the model correctly predicted the absence of these responses, as indicated by the posterior predictive checks.

Table 6.2. The proportion of patients for whom the posterior predictive likelihood of the word repetition data coming from the SP model is greater than or equal to α.

Response	Correct	Semantic	Formal	Mixed	Unrelated	Nonword
Passed check (α=.01)	.607	1.00	.857	1.00	1.00	.750
Passed check (α=.05)	.429	1.00	.643	1.00	.964	.500

Predictions of word repetition accuracy do explain a sizable amount of variance in the observed frequencies, though when the model fails, it tends to dramatically underpredict repetition accuracy (Figure 6.1).

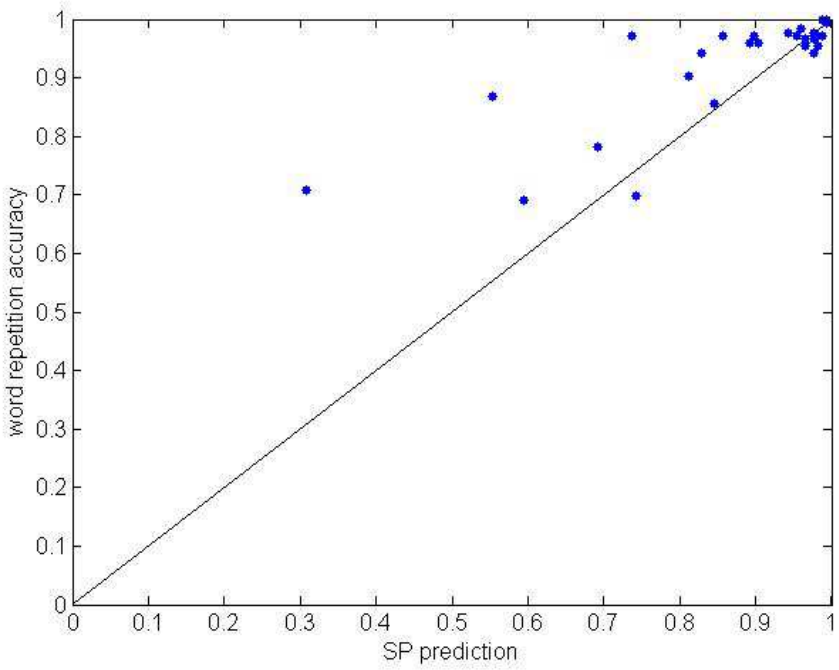

Figure 6.1. Comparison of SP model predictions of word repetition accuracy and obtained accuracy for the 28 patients. The diagonal line represents identity, i.e., perfect prediction.

The SLAM Model of Repetition

Simulating repetition tasks with the SLAM network requires some further assumptions to be made. Naturally, both word and nonword repetition are simulated with a boost of activation to the auditory units. Because the model is intended to describe a fairly high level of speech representation (i.e., syllabified phoneme sequences), we also adopt the assumption of perfect recognition and investigate the same 28 patients as above. To find our initial parameter settings, we used a custom MATLAB script implementing the likelihood function (as described in Chapter 5) for SLAM, and experimented with different values for the timesteps, activation boosts, the phonological units representing a nonword, and the number of phonological neighbors for a nonword. We settled on 4 timesteps to balance activation spreading with computational efficiency, as the Bayesian estimation

procedure has a substantial computational load. We chose to represent a nonword target with units [3, 8, 9], having phonological neighbors [3, 7, 9] and [4, 7, 9], to mimic the situation in word repetition where the word target is the same as in naming (i.e., [1, 7, 9]) and thus has the same 2 phonological neighbors. Because nonwords do not have a semantic component, and also for computational efficiency, we did not run the nonword repetition model with a lexical neighborhood that includes mixed error opportunities, although we did for word repetition. To match the patient data, model responses for nonword repetition were scored using the 3 categories Correct, Formal, or Other; word repetition was scored with the same 6 categories as picture naming. We chose to begin nonword repetition simulations with an activation boost of 5 to each auditory unit, which was the smallest integer value that led to near perfect accuracy with the healthy model setup, i.e., all weight parameters set to 0.04. It is known that word repetition is easier than nonword repetition, and all of the patients demonstrate the ability to discriminate words and nonwords, so we chose to begin word repetition simulations with a boost of 10 instead of 5. This idea is similar to the double boost of activation used in the dual-route model of Hanley et al. (2004); however, the boost comes from a single source in SLAM and can be prevented from activating the lexical layer by reducing the lexical-auditory weight. Another major difference is that SLAM assumes all weights participate in all tasks, and we estimate the parameter values using picture naming and nonword repetition data simultaneously[9]. Word repetition is also simulated by the model for the purposes of generating posterior predictive distributions, but this data is not used in the estimation of model parameters.

[9] This is not only theoretically motivated, it is also required to properly constrain all the parameters, i.e., for the Monte Carlo sampling chains to converge. If only naming data is used, the auditory-motor weight is unconstrained; if only repetition data is used, the lexical-semantic weight is unconstrained.

Again, the parameter estimation procedure is implemented in JAGS (Plummer, 2003). We used uniform prior densities, which were linearly transformed to the [0.0001, 0.04] scale, for the SL, LA, and AM weights. SLAM constrains the LM weight to be less than the LA weight, and we therefore multiplied the LM sample by the LA weight to obtain the LM weight. We used 4 chains with 500 samples, and discarded 50 samples as burn-in, yielding 1800 samples of the posterior likelihood densities for each of the 4 weight parameters along with posterior predictive distributions for each of the 3 tasks. Visual inspection of chains indicated reasonable convergence for all parameters, although the LA weight seemed like it may have benefitted from more samples and/or thinning. The behavioral data, sampling chains, posterior likelihood densities, and posterior predictive distributions are illustrated for an example conduction patient.

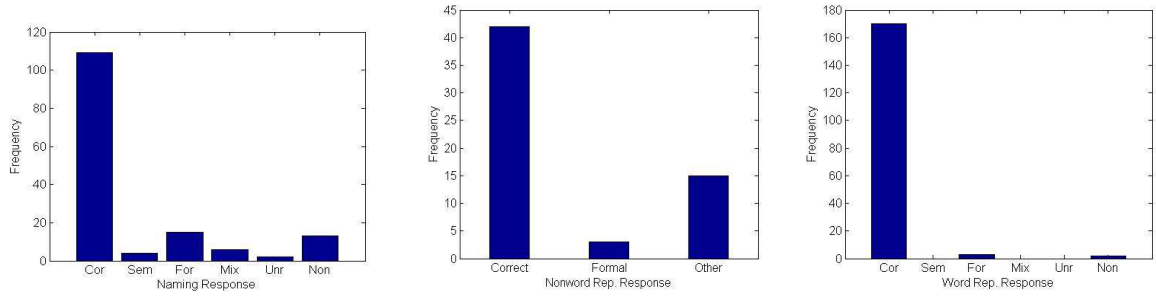

Figure 6.3. Response frequencies on picture naming (left), nonword repetition (middle), and word repetition (right), from an example conduction patient.

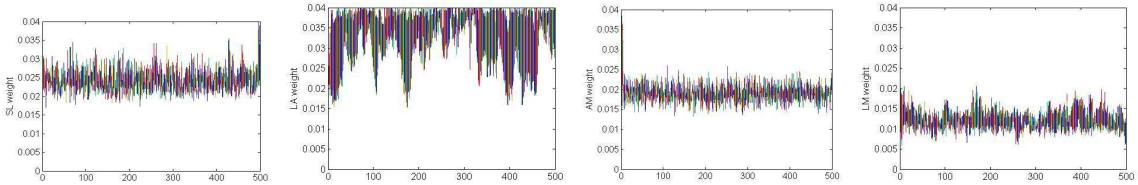

Figure 6.4. Sampling chains for the SL, LA, AM, and LM weight parameters.

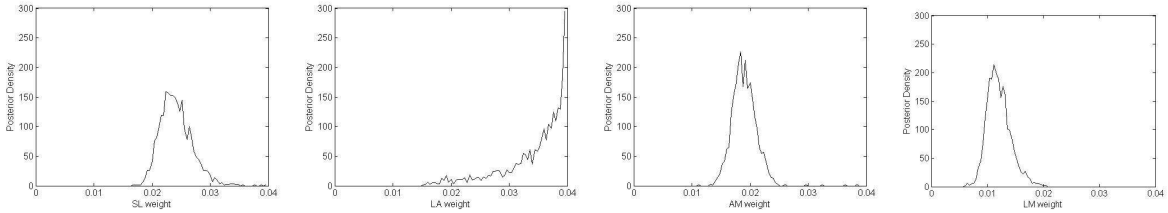

Figure 6.5. Posterior likelihood densities for the SL, LA, AM, and LM weight parameters.

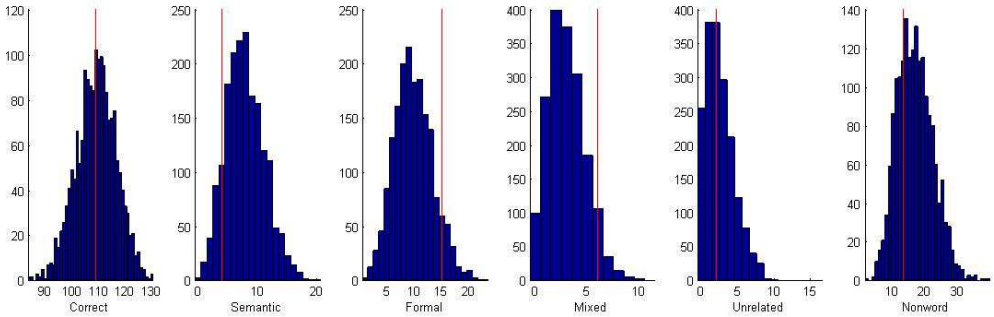

Figure 6.6. Posterior predictive distributions for the frequency of each naming response. The vertical red line indicates the observed frequency. Axes are set for maximal visibility of distribution shape.

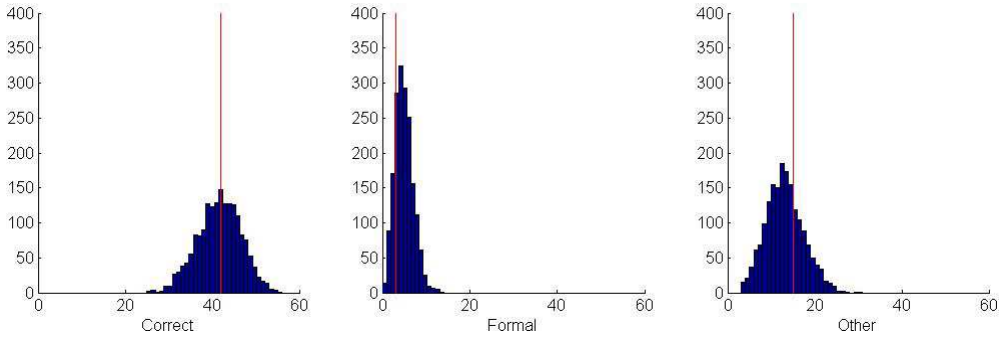

Figure 6.7. Posterior predictive distributions for the frequency of each nonword repetition response. The vertical red line indicates the observed frequency. Axes are fixed to show relative distribution shape and position.

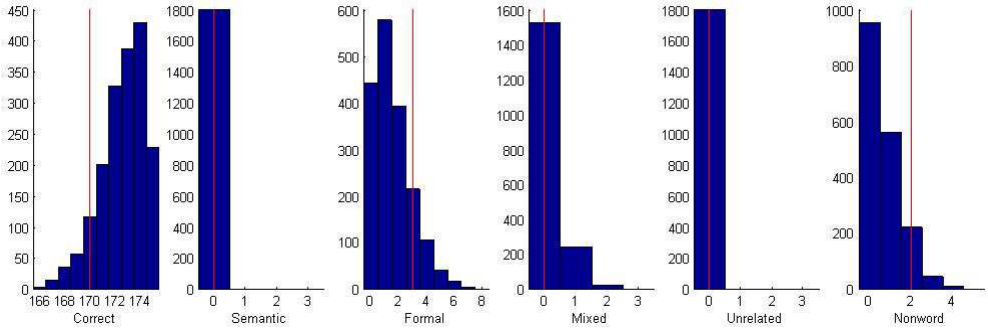

Figure 6.8. Posterior predictive distributions for the frequency of each word repetition response. The vertical red line indicates the observed frequency. Axes are set for maximal visibility of distribution shape.

The 4-parameter SLAM model was able to find good fits for the 6 naming and 3 nonword repetition responses simultaneously; there were 24 patients (86%) who passed the posterior predictive check (α=.01) for all 9 responses. The model also made correct predictions (α=.01) of word repetition accuracy for 17 patients (64%), and correctly predicted all 5 error types for 13 patients (46%).

Table 6.3. The proportion of patients for whom the posterior predictive likelihood of the picture naming data coming from the SLAM model is greater than or equal to α.

Response	Correct	Semantic	Formal	Mixed	Unrelated	Nonword
Passed check (α=.01)	1.00	1.00	.964	.929	1.00	1.00
Passed check (α=.05)	.750	.929	.821	.786	.964	.893

Table 6.4. The proportion of patients for whom the posterior predictive likelihood of the nonword repetition data coming from the SLAM model is greater than or equal to α.

Response	Correct	Formal	Other
Passed check (α=.01)	1.00	.964	1.00
Passed check (α=.05)	1.00	.893	.929

Table 6.5. The proportion of patients for whom the posterior predictive likelihood of the word repetition data coming from the SLAM model is greater than or equal to α. The proportion of patients for whom SP made correct predictions are presented again for comparison.

Model	Response	Correct	Semantic	Formal	Mixed	Unrelated	Nonword
	Passed check (α=.01)	.643	1.00	.750	1.00	.964	.643

SLAM	Passed check (α=.05)	.357	1.00	.643	1.00	.964	.357
SP	Passed check (α=.01)	.607	1.00	.857	1.00	1.00	.750
	Passed check (α=.05)	.429	1.00	.643	1.00	.964	.500

Table 6.6. The proportion of variance in the word repetition data accounted for by SLAM model point estimates. The proportion of variance accounted for by SP is presented again for comparison.

Model	Response	Correct	Semantic	Formal	Mixed	Unrelated	Nonword
SLAM	R^2 (identity)	.398	NA	-.276	-6.26	-.077	.203
	R^2 (best fit)	.546	NA	.380	.095	NA	.421
SP	R^2 (identity)	-.689	NA	.040	-10.4	.462	-.590
	R^2 (best fit)	.660	NA	.405	.051	.482	.665

When we compare the repetition predictions for SP and SLAM quantitatively, we obtain a mixed picture, similar to previous lexical-route and dual-route comparisons (Dell et al., 2007; Nozari et al., 2010). The average rmsd for the SLAM point estimates was .029, ranging [0.0, .11]) which is smaller than SP but not significantly (2-tail p=.29). The improvement is driven by 2 patients who had perfect word repetition which was correctly predicted by SLAM but not by SP. Otherwise, in terms of predicting individual response types in the overall sample, the SP model tends to do slightly better than SLAM. SLAM notably outperforms SP by improving the predictions of repetition accuracy for a handful of patients that SP fits poorly. When we compare SP and SLAM in terms of the posterior

predictive probability of the observed repetition accuracy, there were 14 patients fit better by SLAM, 13 patients fit better by SP, and 1 tie. However, SLAM improved predictions for 4 out of 5 conduction patients and 1 (of 1) transcortical motor patient, in accordance with our expectations. Furthermore, all 5 conduction patients were assigned high lexical-auditory weights (posterior mean LA = [.65, .61, .74, .71, .87]) and low auditory-motor weights (posterior mean AM = [.26, .43, .37, .52, .47]). For this sample of aphasic patients, using nonword repetition accuracy alone or using the second step of naming alone will dramatically underpredict word repetition accuracy; including the auditory-motor weight in the model provides a better description of the relationships between these tasks.

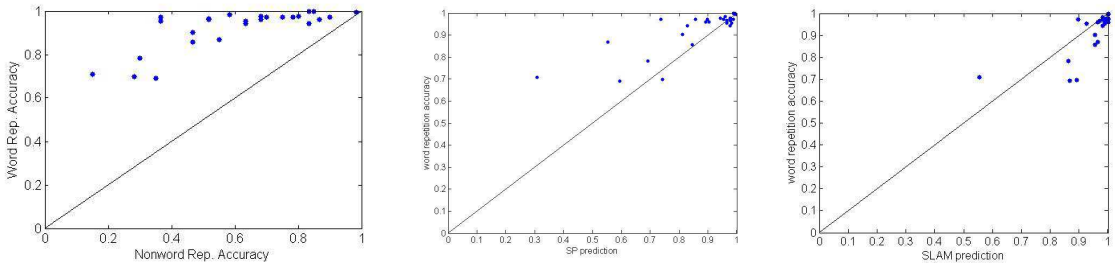

Figure 6.9. (Left) Comparison of nonword repetition accuracy and word repetition accuracy. (Middle) Comparison of SP predictions and obtained word repetition accuracy. (Right) Comparison of SLAM predictions and obtained word repetition accuracy. The diagonal line indicates identity.

Summary

In this chapter, we demonstrated that it is possible to apply our Bayesian approach to SLAM while modeling multiple speech production tasks. We also re-evaluated the SP model's simpler approach to word repetition for comparison. In terms of overall prediction for the sample of 28 patients with unimpaired hearing, the SP model does quite well, especially given its simplicity. Nevertheless, SLAM substantially improves predictions of word repetition for a subset of patients who are fit poorly by SP, notably the conduction

patients. The data supports our assumptions that auditory representations play an active role in word retrieval, and that the activation of auditory targets is preserved while connections to motor representations are damaged in conduction aphasia. The use of nonword repetition data to constrain the AM-weight lends more credence to the claim that our units really are simulating auditory representations. This, in turn, bolsters the claim that auditory representations are participating in naming, since the model is able to find good fits for both tasks with the same weights, as well as generating viable predictions for new data from a different task. All 5 of the conduction patients were assigned strong LA and weak AM weights, meaning that under the processing and architectural assumptions of the SLAM network, the best explanation for conduction aphasia is consistent with the Hierarchical State Feedback Control theory (Hickok, 2012). The model was unable to fit the rate of Formal errors in naming for 1 conduction patient, because this patient made an excessive number of these errors; however, the pattern, if not the magnitude, is still consistent with the concept of preserved auditory feedback to the lexical level (see Chapter 3).

Future work could re-examine some of our simplifying assumptions. The perfect recognition assumption could be relaxed by adding a parameter to account for the probability of mishearing, allowing many more patients to be modeled. Different probabilities for nonword targets slipping to word errors may improve predictions. The sampling procedure can likely be made more efficient to yield more reliable estimates. One of the primary benefits of the computational model is that it can be applied to many types of data from different tasks, as we have demonstrated; if we can simulate the task, it can be

used to constrain the model. Finding more data and considering new processing levels may falsify or lend further support to our theoretical assumptions.

References

Dell, G. S., Martin, N., & Schwartz, M. F. (2007). A case-series test of the interactive two-step model of lexical access: Predicting word repetition from picture naming. *Journal of Memory and Language*, *56*(4), 490-520.

Dell, G. S., Schwartz, M. F., Nozari, N., Faseyitan, O., & Coslett, H. B. (2013). Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition, 128*(3), 380-396.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological review*, *104*(4), 801.

Hanley, J., Dell, G. S., Kay, J., & Baron, R. (2004). Evidence for the involvement of a nonlexical route in the repetition of familiar words: A comparison of single and dual route models of auditory repetition. *Cognitive Neuropsychology*, *21*(2-4), 147-158.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, *13*(2), 135-145.

Martin, N., Dell, G. S., Saffran, E. M., & Schwartz, M. F. (1994). Origins of paraphasias in deep dysphasia: Testing the consequences of a decay impairment to an interactive spreading activation model of lexical retrieval.*Brain and language*, *47*(4), 609-660.

Nozari, N., Kittredge, A. K., Dell, G. S., & Schwartz, M. F. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of memory and language*, *63*(4), 541-559.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 125). Wien, Austria: Technische Universit at Wien.

CHAPTER 7: Summary and Conclusions

In this dissertation, we examined a new computational model of speech production, the Semantic-Lexical-Auditory-Motor model (SLAM). The model instantiates a small lexical network where representations are retrieved with spreading activation, and connections between representations can be reduced to simulate damage. The model's novel architecture was designed to test a critical assumption of the Hierarchical State Feedback Control Theory (HSFC; Hickok, 2012), which posits that auditory representations play a crucial role in speech planning, serving as targets even prior to overt production. Conduction aphasia has been leveraged as supporting evidence, because it has been theorized that these speech patterns can be explained by the preservation of auditory targets that are disconnected from their corresponding motor sequences. If we assume that the lexical network has these connections, as SLAM does, we do indeed find that speech production data from conduction patients is best explained in accordance with the HSFC account. The coordination of auditory and motor representations during speech planning appears to be a viable mechanism for predicting speech production behavior in aphasia.

Along the way, a number of technical developments were achieved. The SP and SLAM models were translated into 5 different computer programming languages: MATLAB, C++, CUDA, PHP, and JAGS, all for slightly different purposes. Versions of our models have been made available at a new website : www.cogsci.uci.edu/~alns/webfit.html. A computational model was converted into a fully-specified statistical model and Bayesian inference techniques were applied to evaluate it. Finally, multiple behavioral tasks were used to estimate the properties of the lexical network simultaneously.

There were also some important auxiliary findings. While the localization study did not replicate, the potentially large sampling variability that exists even in fairly large samples of people with aphasia became evident. This no doubt presents a challenge as we move forward in our attempts at understanding the underlying mechanisms that explain the wide variety of observed behaviors, but appreciating the scope of that variety is an important first step. Of course, as we acknowledged from the outset, our models are not perfect; when they fail, it may be due to oversimplification, or because the theory is wrong, though we've shown so far that the relevant points seem to find support in the data. Further examination of the cases where the model fails may therefore lead to better assumptions. The assumption of item homogeneity is particularly suspect (Smith & Batchelder, 2008), though it seems amenable to relaxation. Also, the assumption of perfect auditory recognition could be replaced with a parameter that accounts for mishearing. Data from new tasks could be simulated, possibly by extending the lexical representations to the time-varying domains of speech articulation and perception. Finally, the methods for associating model parameters with neurological data might be improved, especially via better linking hypotheses. We have shown that the Bayesian approach developed here can flexibly incorporate different types of data into the model, which could include neurological data (Turner et al., 2013). Ideally, the data and methods presented in this dissertation will provide momentum for developing new ideas and tools in the service of treating neurological injuries.

References

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, *13*(2), 135-145.

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*(4), 713-731.

Turner, B. M., Forstmann, B. U., Wagenmakers, E. J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193-206.