2017

Multivariate pattern analysis of input and output representations of speech

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

MULTIVARIATE PATTERN ANALYSIS OF INPUT AND OUTPUT REPRESENTATIONS OF SPEECH

by

CHRISTOPHER J. MARKIEWICZ

B.S./B.C.S., University of Tulsa, 2009

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2017

Approved by

First Reader

Jason W. Bohland, PhD
Assistant Professor of Health Sciences

Second Reader

Frank Guenther, PhD
Professor of Speech, Language, and Hearing Sciences

Third Reader

Tyler Perrachione, PhD
Assistant Professor of Speech, Language, and Hearing Sciences

MULTIVARIATE PATTERN ANALYSIS OF INPUT AND OUTPUT REPRESENTATIONS OF SPEECH

CHRISTOPHER J. MARKIEWICZ

Boston University Graduate School of Arts and Sciences, 2017

Major Professor: Jason W. Bohland, Assistant Professor of Health Sciences

ABSTRACT

Repeating a word or nonword requires a speaker to map auditory representations of incoming sounds onto learned speech items, maintain those items in short-term memory, interface that representation with the motor output system, and articulate the target sounds. This dissertation seeks to clarify the nature and neuroanatomical localization of speech sound representations in perception and production through multivariate analysis of neuroimaging data.

The major portion of this dissertation describes two experiments using functional magnetic resonance imaging (fMRI) to measure responses to the perception and overt production of syllables and multivariate pattern analysis to localize brain areas containing associated phonological/phonetic information. The first experiment used a delayed repetition task to permit response estimation for auditory syllable presentation (input) and overt production (output) in individual trials. In input responses, clusters sensitive to vowel identity were found in left inferior frontal sulcus (IFs), while clusters responsive to syllable identity were found in left ventral premotor cortex and left mid superior temporal sulcus (STs). Output-linked responses revealed clusters of vowel information bilaterally in mid/posterior STs.

The second experiment was designed to dissociate the phonological content of

the auditory stimulus and vocal target. Subjects were visually presented with two (non)word syllables simultaneously, then aurally presented with one of the syllables. A visual cue informed subjects either to repeat the heard syllable (repeat trials) or produce the unheard, visually presented syllable (change trials). Results suggest both IFs and STs represent heard syllables; on change trials, representations in frontal areas, but not STs, are updated to reflect the vocal target.

Vowel identity covaries with formant frequencies, inviting the question of whether lower-level, auditory representations can support vowel classification in fMRI. The final portion of this work describes a simulation study, in which artificial fMRI datasets were constructed to mimic the overall design of Experiment 1 with voxels assumed to contain either discrete (categorical) or analog (frequency-based) vowel representations. The accuracy of classification models was characterized by type of representation and the density and strength of responsive voxels. It was shown that classification is more sensitive to sparse, discrete representations than dense analog representations.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Structures	
AG	Angular gyrus
CG	Cingulate gyrus
FMC/FOC	Frontal medial/orbital cortex
FO/CO/PO	Frontal/central/parietal operculum
FP	Frontal pole
H	Heschl's gyrus
IFg/IFs	Inferior frontal gyrus/sulcus
IFt/IFo	Inferior frontal gyrus pars triangularis/opercularis
IPC/IPL	Inferior parietal cortex/lobule
INS	Insula
ITg	Inferior temporal gyrus
LG	Lingual gyrus
MC	Motor cortex
MFg	Middle frontal gyrus
MTO/ITO	Middle/inferior temporal occipital gyrus
OC	Occipital cortex
PH	Posterior hippocampal gyrus
PMC	Premotor cortex
PP	Planum polare
PreSMA	Anterior portion of the supplementary motor area
PT	Planum temporale
SC	Sensory cortex
SFg	Superior frontal gyrus
SMA	Supplementary motor area
SMg	Supramarginal gyrus
SPL	Superior parietal lobule
STg/STs	Superior temporal gyrus/suclus
TF	Temporal fusiform
TOF	Temporal occipital fusiform
TP	Temporal pole
Prefixes	
v-/mid-/d-	Ventral/middle/dorsal
a-/p-	Anterior/posterior
m-	Medial

Chapter 1

Introduction

In order to perform a seemingly simple task like repeating an auditory word or syllable, our brains must rely on a series of neural representations and functional pathways. A speaker must register an auditory representation of the incoming sound sequence, map that onto learned speech items, maintain an accurate representation over any delay between the input and required output, interface that working memory representation with the motor output system, and fluently produce the target output sequence. Accordingly, to carry out this complex series of computations, the brain relies on a large, distributed set of cortical areas.

This dissertation seeks to clarify the nature and neuroanatomical localization of speech sound representations in perception and production through multivariate analysis of neuroimaging data. While a large number of studies have sought to associate cortical areas with specific component processes during speech, the studies presented here ask to what extent their activation patterns reflect the specific speech sounds heard, planned and produced. Multi-voxel pattern analysis is a technique well-suited to probing for neural correlates of categorically defined cognitive states, and speech is composed of categorically distinct sounds, or phonemes; these studies induce subjects to perceive, prepare and produce speech sounds, permitting the use of MVPA to probe for phonological content. The remainder of this chapter presents the theoretical, experimental, and methodological basis of this dissertation.

1.1 Organization of speech processing

The classical breakdown of speech into perceptual (i.e., for processing auditory inputs) and production (i.e., for preparing and executing motor outputs) related components (Lichtheim, 1885) has been, in part, supported by differential effects of lesions to the posterior or anterior portions of the left hemisphere, with perception localized to posterior superior temporal cortex and production to inferior frontal cortex. There remains considerable controversy, however, in how the brain encodes speech sounds; while there are various accounts for deriving speech content from acoustic signals through auditory and general cognitive processes (*e.g.*, Hickok and Poeppel, 2004; Diehl et al., 2004; Massaro and Chen, 2008; Stasenko et al., 2013, 2015), some suggest, for example, that the premotor or motor cortices (traditionally associated with speech output) are automatically engaged and serve an important role in perception (Liberman et al., 1967; Fadiga and Craighero, 2003). In addition, a number of theoretical models suggest the activation of auditory cortical areas for directing speech production (*i.e.*, target readout; Guenther et al., 2006; Hickok et al., 2011). Thus, a clean segregation of circuitry between these two components is unlikely and not well supported by the existing evidence.

1.1.1 The dual streams model

The theoretical framework that guides this dissertation is based on a combination of the dual streams speech processing model (Hickok and Poeppel, 2004, 2007; Rauschecker and Scott, 2009) and the GODIVA neurocomputational model (Bohland et al., 2010), which treats output-related processes of syllable sequence representation and production. A unified theory requires the specification of how speech is represented at a high level of detail across both receptive and expressive speech processes. The

links between such input and output-related components are perhaps most apparent in auditory repetition, variations on which form the core of the studies presented in this work. Repetition tasks have been suggested to rely upon the so-called *dorsal stream* (Hickok and Poeppel, 2007; Saur et al., 2008), a pathway which maps between auditory and motor representations of speech sounds. The dorsal stream is considered to be left hemisphere dominant, projecting from the posterior superior temporal gyrus (STg) and interconnecting the planum temporale (PT), including the region at the posterior portion of the Sylvian fissure at the parietotemporal junction (Spt), inferior parietal cortex, and premotor and inferior frontal areas. Damage to the dorsal pathway – in particular to area Spt (Buchsbaum et al., 2011) and/or to left posterior temporoparietal cortex (Baldo et al., 2012) – can result in conduction aphasia, a language disorder characterized in part by impaired repetition and problems with phonological short-term memory.

The ventral stream, in contrast, is proposed to map from auditory to conceptual-semantic representations. This pathway is proposed to project from posterior STs / STg, and interconnect the temporo-parieto-occipital junction (Hickok and Poeppel, 2000, 2004), mid-MTg (Indefrey and Levelt, 2004), anterior STs and Broca's area (Friederici et al., 2000a,b), constructing progressively larger (lexical) and more meaningful (semantic/grammatical) representations. For single syllable repetition, any ventral stream access is likely to be limited to incidental or opportunistic representation, in the absence of any explicit semantic or grammatical demands.

1.1.2 Speech perception and working memory

In most theoretical models of repetition, auditory inputs are mapped to motor outputs first via a phonological layer (e.g., Hartley and Houghton, 1996; Hanley et al.,

2004; Nozari and Dell, 2013), and substantial work has focused on localizing such an "input buffer." At early stages of the cortical hierarchy, auditory / phonetic representations of speech inputs are supported by the posterior superior temporal lobes bilaterally (Hickok and Poeppel, 2000). Auditory association areas, which receive these inputs, then can be considered candidates for more abstract or phonological representations of a speech input sequence. The planum temporale (PT) has been shown to be functionally subdivided, with lateral portions likely involved in general auditory processing (see *e.g.*, Binder et al., 1996) and medial portions important for processing self-produced feedback (Tremblay et al., 2013b). Furthermore, responses of anterior and middle PT have been shown to reflect the statistical structure and phonological complexity of speech sequences (Tremblay and Small, 2011; Tremblay et al., 2013a; Deschamps and Tremblay, 2014), leading to a suggested role in converting auditory inputs into phonological representations (Deschamps and Tremblay, 2014). The posterior PT has been further subdivided into lateral and medial (area Spt) portions, with the former shown to be sensitive to subsegmental manipulations in a nonword repetition task, and the latter to the number of syllables used (McGettigan et al., 2011). In their review of seven speech perception studies, Hickok and Poeppel (2007) showed that the mid to posterior STg and superior temporal sulcus (STs) respond preferentially to syllable (CV or CVC) stimuli over a variety of non-speech acoustic controls, and argued that these regions support phonological-level processes. Overt and covert speech *production* also induce responses in posterior STg, STs and PT (Paus, 1996; Hickok et al., 2000; Okada et al., 2003; Okada and Hickok, 2006b); activation of these areas even during silent speech is consistent with a possible role in representing auditory targets for online correction of speech (Guenther et al., 2006; Hickok et al., 2011). Similarly, posterior STs responds to both heard and

recalled words (Wise et al., 2001), possibly storing transient auditory representations retrieved from sensory or long-term memory.

The inferior parietal cortex (IPC) is often considered part of the dorsal stream in part due to its interconnections with superior temporal and inferior frontal regions via the arcuate fasciculus (Catani et al., 2005). The IPC contains at least seven cytoarchitectonically distinct areas (Caspers et al., 2006), but most previous studies have focused on the roles of the macroanatomically defined supramarginal and angular gyri. The importance of IPC for auditory repetition is backed by a voxel-based lesion symptom mapping (VLSM) study showing a strong association between left supramarginal gyrus (SMg) integrity and performance on word and nonword repetition tasks (Rogalsky et al., 2015). The left SMg has been commonly associated with phonological working memory (PWM; Paulesu et al., 1993), but its activation is not consistently found across speech studies (Buchsbaum and D'Esposito, 2008). Pugh et al. (2001), in a review of previous literature, found both angular and supramarginal gyri to respond more to reading pseudowords than words, while a PET working memory study requiring subjects to maintain lists of words or pseudowords found no significant inferior parietal response (Fiez et al., 1996), indicating a possible sensitivity to encoding demands rather than to short-term maintenance of linguistic materials. Studies by Jonides et al. (1998) and Awh et al. (1996), on the other hand, associated IPC engagement with storage, retrieval and memory load, but not encoding, in visual language tasks. Ravizza et al. (2004) characterized dorsal and ventral IPC as being sensitive to working memory load and encoding, respectively. While the roles different IPC subregions play in PWM remain unclear, a syllable repetition task with strong working memory demands might be expected to induce phoneme-specific responses to in IPC.

On the other hand, the existence of PWM as an independenty faculty with distinct anatomical localization (Baddeley, 1992) has been called into question. Instead, the phonological input and output buffers (Jacquemot et al., 2007) and the conversion mechanisms between them (*i.e.*, the dorsal stream) are proposed as sufficient substrate for PWM (Jacquemot and Scott, 2006; Acheson and MacDonald, 2009; Hickok, 2009; Perrachione et al., 2017).

1.1.3 Speech production

It is universally accepted that the frontal cortex is critical in directing the output of speech during repetition. Rolandic cortex is activated strongly and bilaterally during overt articulation, and with some degree of left lateralization when articulation is covert (Wildgruber et al., 1996; Riecker et al., 2000), possibly reflecting prepared articulatory commands and their predicted somatosensory consequences (Corfield et al., 1999; Lotze et al., 2000). Left lateralized activation is also seen under passive speech listening conditions with no distractor task (Wilson et al., 2004; Pulvermüller et al., 2006), and TMS priming of left motor cortex has been shown to assist phoneme discrimination in noise (D'Ausilio et al., 2009), suggesting a role for classically-defined motor output areas in some aspects of speech perception and/or perceptual judgments and their potential importance even in the *input* portions of repetition tasks. (However, see Hickok (2010) for a critique of these views.

The adjacent ventral premotor cortex (vPMC) has been implicated in phonetic encoding specifically at the level of syllables (Peeva et al., 2010), and may function as a neural correlate of a *mental syllabary* (Levelt and Wheeldon, 1994) or Speech Sound Map (Guenther et al., 2006), storing sensorimotor programs for well-practiced sounds.

1.1.4 Evidence for phonological encoding

A phonological *output buffer* is a commonly proposed component in speech planning and production models (e.g., Dell et al., 1997; Roelofs, 1997; Goldrick and Rapp, 2007), and the existence of separate phonological representations for encoding speech inputs and planned outputs is supported by clinical studies (Martin et al., 1999; Jacquemot et al., 2007). The GODIVA model (Bohland et al., 2010) posits the existence of parallel phonological output buffers in the left inferior frontal sulcus (IFs) and pre-supplementary motor area (pre-SMA), with the left IFs serving to encode phonemic sequences and the pre-SMA representing abstract syllable frames (see also MacNeilage, 1998). Activation of the IFs is sensitive to both the phonological complexity of produced syllables and the complexity of planned syllable sequences (Bohland and Guenther, 2006). The IFs forms the dorsal boundary of the inferior frontal gyrus (IFg). Papoutsi et al. (2009) observed a dorsal-ventral segregation of function within IFg (see also Molnar-Szakacs et al., 2005), with the dorsal portion playing a role in phonological encoding, and the ventral portion having a more motoric role in phonetic encoding. Long et al. (2016), found a timing, but not articulatory, effect when cooling IFg, supporting a role for IFg in sequencing over articulation. Activation of the pre-SMA and adjacent medial premotor areas is also frequently observed in speech perception experiments without explicit production requirements, though any causal role for their activation during perception remains unclear (see also Adank, 2012).

To what extent different brain areas are explicitly involved in the encoding or representation of speech sounds is difficult to assess through traditional fMRI paradigms. This is, in large part, due to the possibility that differences in activation could be observed for a number of reasons other than a requirement to *represent* one or more

speech sounds. Recent studies have begun to fill in existing knowledge gaps with protocols and analysis techniques that allow neural representations to be probed more explicitly. Repetition suppression (RS) paradigms have been used as an indicator that a neuronal population treats two stimuli (e.g., two instantiations of the same syllable or phoneme) as the same or different, with the assumption that trial-to-trial neuronal adaptation gives rise to suppression or enhancement of the BOLD signal. Phonological RS has implicated left posterior STg (Graves et al., 2008) in an auditory pseudoword repetition task and bilateral STs and IFg (Vaden et al., 2010) in a task requiring listening to words of with varying degrees of phonological similarity. Vaden and colleagues also reported *increased* activation in bilateral SMg in response to word lists with repeated phonological content, compared to word lists with purely novel phonological content (Vaden et al., 2010). Consistent with phonological representations in dorsal IFg / IFs, Myers et al. (2009) found an increased response to between-phoneme differences in voice onset time, with little sensitivity to changes within a phonetic category, of perceived syllables in this region. This was in contrast with left superior temporal regions, which showed both within- and across-category changes in activation. Additionally, electrocorticography studies have shown reliable phoneme identification in left STg (Chang et al., 2010), although this may be supported by a phonetic, feature-based representation (Mesgarani et al., 2014).

Although there is general agreement on the structures involved in speech perception and production, there remain questions as to the precise role of these regions in the representation of speech sounds. Studies have used a variety of methods to localize phonological sensitivity, but there is a scarcity of systematic studies of the representational units of speech perception and production.

1.2 fMRI methods

Functional magnetic resonance imaging (fMRI) depends on the blood-oxygenation-level-dependent (BOLD) signal (Ogawa et al., 1990) as a proxy for neuronal activity. This signal reflects changes in blood volume, blood flow and oxygen use (Mandeville et al., 1999) associated with the vascular response to metabolic activity, and has been shown to correlate with local field potentials (Logothetis et al., 2001). A hemodynamic response function (HRF) is the model impulse response of the BOLD signal to a metabolic event, characterized by a peak response lagging the event onset by approximately 5s and an undershoot following the event offset (Friston et al., 1998). This slow response temporally smooths neural activations, and is a common basis for estimating neural responses from observed BOLD signals (Josephs et al., 1997; Hinrichs et al., 2000).

1.2.1 Sparse Acquisition

Studying speech with fMRI presents two physical challenges to experimental design: the noise of the MR scanner can interfere with subjects' perception of speech stimuli and the auditory feedback of their own voices, and overt speech during MR acquisition has also been shown to induce speech-related head motion (Gracco et al., 2005) and susceptibility artifacts resulting from changes in vocal tract configuration (Birn et al., 1998). Further, there is evidence that speech perception makes different demands of auditory (Du et al., 2014) and motor (Meister et al., 2007) systems under noisy conditions. Sparse and clustered volume acquisition paradigms introduce delays in volume acquisition (Eden et al., 1999; Edmister et al., 1999; Hall et al., 1999), which can be exploited to allow subjects to hear and produce speech in relative quiet and reduce the impact of motion and susceptibility artifacts.

1.2.2 Univariate Contrasts

To a first approximation, most fMRI analyses begin by describing conditions of interest as on-off time series, convolving these series with an HRF to form a *design matrix*. The strength of response to each condition is estimated *at each voxel* by regressing the voxel's time series against the design matrix. These parameter estimate, or *beta*, maps may be used to construct contrasts, for instance by subtracting betas associated with control trials from betas associated with task trials, to identify voxels that receive increased blood flow when engaged in the task than at rest. A standard contrast analysis uses T or F tests to determine an uncorrected significance level. A number of methods exist to account for multiple comparisons (Nichols and Hayasaka, 2003), including cluster-wise thresholding, in which a cluster-defining threshold is applied, and clusters (contiguous super-threshold regions) are further thresholded by a cluster-size threshold to achieve the desired false positive rate.

1.2.3 Multivariate Pattern Analysis

An alternative approach to examining fMRI data is to treat a beta map as a feature vector in a multivariate statistical test (commonly referred to as machine learning). This class of analyses is known as multi-voxel or multivariate pattern analysis (Haxby et al., 2001; Norman et al., 2006, MVPA). In contrast to univariate approaches, in which a region is deemed significant if sufficiently many adjacent voxels (a cluster) respond more strongly to one condition than another, a multivariate analysis takes into account the covariance structure of multiple the responses of multiple voxels. The earliest such studies identified a small number of regions of interest (ROIs), and applied *ad hoc* classification techniques (see, *e.g.*, Haxby et al., 2001) to characterize the differentiability of trial conditions, such as visual stimulus category, based only

on BOLD responses within each ROI.

An important development in multivariate techniques is the "searchlight" method (Kriegeskorte et al., 2006; Chen et al., 2011), which defines one ROI per voxel: a geometrically-defined neighborhood of constant extent centered around that voxel. A figure of merit is calculated within each neighborhood, and mapped to the central voxel. From a purely statistical perspective, searchlight analysis provides a solution to the "double-dipping" problem (Kriegeskorte et al., 2009), in which ROI definition and any subsequent statistical tests must be performed on independent datasets. From an fMRI analytic perspective, searchlight provides whole-brain maps, which admits spatial comparisons that are difficult to make in the absence of tiled ROIs of uniform size. Searchlight analysis does, however, reintroduce the multiple-comparisons problem present in univariate analysis, though most family-wise error correction methods no longer apply. In particular, the distributions of figures of merit (such as cross-validation accuracy (Stone, 1974; Kohavi, 1995)) are difficult to characterize, rendering T and binomial tests against a theoretical chance accuracy unreliable. Additionally, as MVPA is typically applied to un-smoothed data, and the smoothing inherent to searchlight analysis – neighboring voxels have overlapping searchlights – has not been well-characterized, parametric cluster thresholding methods cannot be applied. Thus, there is little alternative to computationally-intensive non-parametric significance tests (Stelzer et al., 2013).

The quantity that multi-voxel analyses measure is the degree to which a set of voxels correlates with stimulus category (or some other aspect of the study design), often termed "information"[1], and a significant result is said to be indicative of "informative voxels". Anderson and Oates (2010) argue against model inspection – attribution of

[1]Note that this is unrelated to any formal information theoretic term.

weights to voxels or time points – as a strategy to impute information to specific voxels, as the idiosyncracies of the model itself may have more influence on a voxel's selection than that voxel's timecourse. They further demonstrate that different voxel patterns may be discriminable only by certain classes of models. Searchlight analyses present an additional set of interpretive challenges, as the spatial extent of the searchlight must be considered in associating a region with a result found in that region, as discussed in Etzel et al. (2013). For example, a small, highly informative region may produce a large cluster of above-chance classification accuracies due to the number of searchlights containing it, while a large, moderately informative region may produce a small cluster of above-chance accuracy if most voxels are required to accurately classify the stimulus, or fail to be detected if the searchlight is too small.

1.3 Organization of dissertation

The studies described here are somewhat exploratory works – rather than attempting to address any one specific hypothesis or theory, we employed simple, repetition tasks, motivated by current theories of speech perception and production that propose complex, reciprocal interactions between the two processes (*e.g.*, Guenther et al., 2006; Jacquemot and Scott, 2006; Hickok et al., 2011; Majerus, 2013). Additionally, we were motivated to determine to what extent the neural machinery used in perceiving and maintaining a representation of a heard syllable overlapped the substrates necessary for holding and enacting a plan for motor output. These questions derived theoretically both from case studies that showed dissociations in phonological input and output processes (*e.g.* Jacquemot et al., 2007) and from the controversies surrounding the use of frontal and motor circuitry in passive speech perception (Pulvermüller et al., 2006; Iacoboni, 2008; Hickok, 2010). Using stimuli designed to vary

systematically across phonemic categories, we generated datasets with rich metadata, that could be queried on multiple dimensions.

Chapter 2 describes a functional magnetic resonance imaging (fMRI) experiment that uses a delayed syllable repetition task in order to engage input and output representations - neural processes necessary to achieve accurate perception and production - of the same speech token. Estimating responses to multiple temporally-spaced events in individual trials, we apply searchlight multi-voxel pattern analysis (MVPA) (Haxby et al., 2001; Norman et al., 2006; Kriegeskorte et al., 2006) techniques to detect correlates of phonemes and syllables at the input and output stages of the task.

Chapter 3 describes a second fMRI experiment, which extends the experimental protocol developed in Chapter 2 to dissociate the phonological content of the auditory stimulus and vocal target, and thus to disambiguate responses in areas traditionally associated with motor output to perception from vocal preparation, and responses in auditory cortex to production from auditory memory.

MVPA considers how information content across multiple voxels correlates with discrete stimulus classes (Haxby et al., 2001; Norman et al., 2006; Kriegeskorte et al., 2006). One barrier to assessing and interpreting neural representations of speech sounds is the correlation of acoustic features and phonological sound categories. To address this, Chapter 4 proposes a preliminary model of cognitive signals, which is used to explore the properties of our multivariate analyses.

Chapter 2

Mapping the cortical representation of speech sounds in a syllable repetition task

This chapter has been published in modified form at NeuroImage (Markiewicz and Bohland, 2016).

2.1 Introduction

Speech repetition relies on a series of distributed cortical representations and functional pathways. A speaker must map auditory representations of incoming sounds onto learned speech items, maintain an accurate representation of those items in short-term memory, interface that representation with the motor output system, and fluently articulate the target sequence. A "dorsal stream" consisting of posterior temporal, inferior parietal and premotor regions is thought to mediate auditory-motor representations and transformations, but the nature and activation of these representations for different portions of speech repetition tasks remains unclear. Here we mapped the correlates of phonetic and/or phonological information related to the specific phonemes and syllables that were heard, remembered, and produced using a series of cortical searchlight multi-voxel pattern analyses trained on estimates of BOLD responses from individual trials.

Multi-voxel pattern analysis (MVPA) considers how information content across multiple voxels correlates with discrete stimulus classes (Haxby et al., 2001; Norman et al., 2006; Kriegeskorte et al., 2006). MVPA variants have been employed by

a number of researchers to determine brain areas whose response patterns predict some aspect of speech stimuli (Formisano et al., 2008; Kilian-Hütten et al., 2011; Lee et al., 2012; Merrill et al., 2012; Abrams et al., 2013; Du et al., 2014; Arsenault and Buchsbaum, 2015; Correia et al., 2015; Evans and Davis, 2015; Zhang et al., 2016). Formisano et al. (2008) found responses able to discriminate between different vowels distributed bilaterally across the mid-posterior STg / STs, with responses able to discriminate different speakers of those vowels more focal and right lateralized. Using a /ba/-/da/ discrimination task, Lee et al. (2012) found that the patterns of activation in voxel clusters within the left IFg, pre-SMA, and STg were predictive of the syllable that subjects perceived, though whether these areas discriminated phonemes or syllables was unclear. In a simple listening experiment, Zhang et al. (2016) found predictive patterns for consonants and vowels, across different syllables, along the mid-STg bilaterally. Du et al. (2014) presented subjects with syllables that spanned four consonant classes under different background noise conditions. Using a multivariate method, they found that the ability for responses in different brain regions to predict the phoneme class in noise differed, with inferior frontal cortex most resilient to noise, followed by ventral premotor and inferior parietal cortex, with no discernibility under any noise conditions in STg. In recent work using representational similarity analysis (Kriegeskorte et al., 2008), to compare the differential response patterns of local cortical areas to syllable and syllable-like acoustic inputs to those predicted by simple theoretical models, Evans and Davis (2015) have proposed a hierarchical organization of speech representations. In particular, traditional "output" areas (somatomotor cortex) appeared to have the most abstract representation of the syllable, with early auditory areas retaining the most acoustic detail. The vast majority of studies employing MVPA and related techniques have focused on speech perception

rather than production; here we extend the approach to syllable repetition, which allows analysis of both input and output related representations.

In this study, we used systematically constructed CVC syllable stimuli to identify regions that correlate with phonological information at segmental and suprasegmental levels. We employed a delayed, single syllable repetition task to engage input, working memory maintenance, and output-related representations of these syllables, and a sparse fMRI paradigm to most effectively capture responses to stimulus and vocalization events. Using MVPA, we sought to localize cortical areas whose response patterns to either the input or output portions of the repetition task made significant predictions about the linguistic class labels for each stimulus. We hypothesized that MVPA analyses of the input-related responses to the auditory stimulus would highlight phonemic representations in the left posterior superior temporal sulcus and possibly Spt and/or the inferior parietal cortex. We anticipated that analyses of output-related responses would reveal a phonemic output buffer in the left inferior frontal sulcus and/or dorsal portion of the inferior frontal gyrus pars opercularis, and a syllabic representation in the left ventral premotor cortex. Because Spt has been proposed as a critical, bidirectional sensory-motor interface, we expected any predictive information in this area would also be observed during output.

2.2 Materials and methods

2.2.1 Participants

15 right-handed native American English speakers participated in this study (9 females, 6 males, mean age = 25.0, SD = 5.5, range = 19-33). No participants reported any history of speech, language, or hearing disorders. All participants gave informed consent under the protocol approved by the Institutional Review Board of Boston

University. The data from one female subject (S13) were removed from the analysis due to an abnormally high error rate during the task (see Results). For two subjects, one run (see description of sessions below) was omitted due to scanner technical problems, while two runs were omitted in a third subject. Thus, for each of 14 subjects, from 6 to 8 runs of data were analyzed.

2.2.2 Task design

Each trial began with the presentation of an auditory stimulus (syllable; see details below) and a gray fixation point, followed by an 8-9 second delay / maintenance period, after which a GO signal (the fixation point changing from gray to green) was presented, cueing participants to repeat the perceived syllable aloud. During the delay, EPI scans were triggered at 2s and 5s after the beginning of the trial. These volume acquisitions were timed to align with theoretical peak hemodynamic responses to the GO signal of the previous trial, and the stimulus presentation of the current trial, respectively (see Figure 2.1). The delay varied from 8 to 9s, in 0.1s increments, to prevent participants from anticipating the GO signal.

Sessions were broken into 8 runs, each consisting of 48 trials. In 44 trials of each run, participants were presented with syllable stimuli, while control stimuli (noise) were presented in two trials, and no stimulus (silence) was presented in two trials. As a control for motor output on silence and noise trials, participants were instructed to press a button with any finger on their right hand. For two subjects, there were four noise trials and no silence trials per run.

Prior to the session, participants were informed that the stimuli they would hear were constructed from the three chosen vowels and consonants and spoken by four different talkers. They were instructed to naturally produce the syllables they heard

without attempting to mimic detailed acoustic features of the specific stimulus.

2.2.3 MR-data acquisition

All measurements were performed using a 3T Philips Achieva MRI scanner with an 8 channel head coil at the Boston University Center for Biomedical Imaging. T1-weighted anatomical images were acquired for anatomical reference and coregistration with functional data ($0.98\times0.98\times1.2$ mm^3 voxels, 150 sagittal slices, 256×254 matrix, repetition time = 6.8 ms, echo time = 3.1 ms, P reduction (AP) SENSE factor = 1.5, S reduction (RL) SENSE factor = 2). Functional volumes consisted of 40 echo-planar transverse slices (3mm thickness), acquired in ascending order, with no gap ($3.03\times3.03\times3$ mm voxels, 76×75 matrix, acquisition time = 2500ms, echo time = 35ms, flip angle = 90°, P reduction SENSE factor = 2). Functional volumes were acquired in a sparse acquisition paradigm (Figure 2.1), triggered externally by a TTL pulse delivered from the stimulus delivery computer, mimicking a cardiac gating signal (Markiewicz, 2016). Two additional volumes were acquired at the end of each run in order to capture residual hemodynamic activity in response to previous experimental events.

Auditory stimuli were delivered through Sensimetrics S14 MRI-compatible insert earphones, and were pre-filtered to equalize the frequency response at the earphones. Subject vocalizations were recorded using an Optoacoustics FOMRI II fiber optic microphone attached to the head coil and digitized for offline analysis.

2.2.4 Materials

Stimuli were consonant-vowel-consonant (CVC) syllables designed to permit independent analysis of three segments, including a comparison of consonants in two distinct

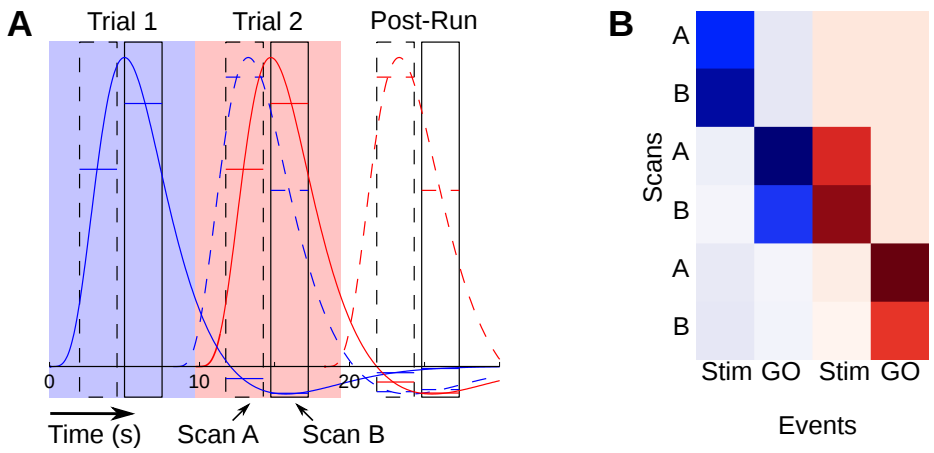


Figure 2.1: **Specification of sparse event-related design.** (A) Schema of unevenly-spaced sparse acquisition paradigm. The solid blue curve indicates the expected hemodynamic response function (HRF) associated with the presentation of the stimulus at $t = 0s$. Scans, represented by black-outlined boxes, are acquired (with TA = 2500ms) beginning at $t = 2s$ (A) and $t = 5s$ (B). The subject is cued to produce the syllable at $t = 8.5s \pm 0.5s$, and the dashed blue curve indicates the expected response associated with production, timed to the cue. Red curves are associated with stimulus and production cues for a second trial. Horizontal lines indicate the mean values of theoretical HRFs across the duration of a scan, which are used as regressors in a general linear model. (B) Design matrix to estimate BOLD response amplitudes to each event from the acquired scans. The blue columns depict the contributions of each scan to trial 1 events, and the red columns depict contributions to trial 2 events. See Figure 2.3 for further details on event estimation.

syllable positions. Stimuli were constructed from the consonants /m/, /t/ and /l/ and the vowels /ɪ/, /ɛ/ and /ʌ/ (Table 2.1). With the constraint that two different consonants must be used in each syllable, this produced 18 unique syllables. To select these phonemes, we parsed the CELEX English Frequency, Syllables corpus, constructed predominantly from written sources (Baayen et al., 1993), and calculated the distribution of frequencies of all 18 CVC syllables composed of any sets of three consonants and three vowels. We chose this phoneme set to generate syllables with a wide range of frequencies of occurrence, from very infrequent (<1 per million syllables) to moderately frequent (584 per million). All syllables were phonotactically legal in American English.

Two male and two female native English speakers recorded the stimuli, and five recordings of each syllable per speaker were used to allow for additional acoustic variations in the auditory tokens that subjects heard. On any given trial (see below), subjects heard a randomly selected recording of the chosen syllable, and they heard each recording from 1 to 7 times over the course of the experiment. Speech-shaped control stimuli were generated by amplitude modulating pink noise by the Hilbert envelopes of the original stimuli.

Mean formant frequencies were extracted from the approximate mid-points of vowels in both the presented stimuli and recordings of subject vocalizations using custom PRAAT (Boersma and Weenink) scripts (You et al., 2015), and random subsets were hand-checked for accuracy. Trials with vocalization formants that could not be extracted were excluded from acoustic analysis.

	/ɪ/		/ɛ/		/ʌ/	
/m/	/mɪl/ (33)	/mɪt/ (248)	/mɛl/ (17)	/mɛt/ (298)	/mʌl/ (28/48)	/mʌt/ (133/0)
/l/	/lɪm/ (32)	/lɪt/ (175)	/lɛm/ (2)	/lɛt/ (407)	/lʌm/ (95/9)	/lʌt/ (88/0)
/t/	/tɪl/ (540)	/tɪm/ (120)	/tɛl/ (584)	/tɛm/ (308)	/tʌl/ (1/0)	/tʌm/ (185/17)

Table 2.1: The consonants /m/, /t/, /l/, and vowels /ɪ/, /ɛ/, /ʌ/ were selected to generate CVC stimuli that span a wide range of syllable frequencies (in parentheses, per million syllables) in General American English. For the vowel /ʌ/, the first number indicates frequency of the reduced schwa vowel sound.

2.2.5 Behavioral assessment

All subjects' vocalizations were verified against the presented stimuli. Due to at times inconsistent recording quality (see Results), errors were marked only if the successfully recorded portions of vocalization were inconsistent with the stimulus. For instance, if only the sound sequence /ɪt/ was audible, this vocalization would be considered a match to either a target of /mɪt/ or /lɪt/. Incorrect vocalizations were relabeled for analysis of output-related datasets (described below), but left unchanged for input-related analyses. Unrecognizable vowels were excluded from acoustic analysis. Verification was performed by one member of the research staff, with spot checks for ambiguous vocalizations.

2.2.6 Preprocessing

We reconstructed cortical surfaces from the MPRAGE images with FreeSurfer (Dale et al., 1999; Fischl, 2012) v5.3.0, which was also used to parcellate cortical surfaces into regions of interest according to a custom speech-centric atlas (Tourville and Guenther, 2012). All functional volumes from all runs were realigned to the first volume of the first run using the FreeSurfer Functional Analysis Stream (FsFast). A preliminary general linear model (GLM) was designed to remove linear trends and motion-related variance (using Friston's 24-parameter model Friston et al., 1996) in

the BOLD signal. Regressors were entered separately for A and B scans, and a linear ramp was sampled at the temporal mid-points of each scan for drift removal. Finally, separate constant regressors were used to remove overall mean intensity differences between A and B scans for each run. The first volume from each run (which was not timed to a specific event of interest) in all subjects was modeled with a separate parameter and discarded. Detrending was performed using custom Python scripts.

With the exception of the estimation of overall task effects (see Results), we normalized (z-scored) the activation of each voxel relative to the mean and variance of control conditions (silence and noise trials) across all runs, prior to analysis. Control trials were thereafter excluded from analysis.

2.2.6.1 Uneven scan timing correction

To capture the peaks of both input- and output-related events, we used an unevenly-spaced sparse acquisition protocol. As a result, A and B scans are influenced by different T1 saturation effects (Schmidt et al., 2008), which is problematic since events are also intrinsically linked with different experimental conditions (i.e., input or output). To correct for this, we modeled the ratio of gray-matter intensities between B and A scans on a per-slice basis.

Scan timing correction was performed using the first 13 subjects recorded under this protocol. We first applied FsFast motion correction to register all functional images to the first volume of each session. We used the `aseg` FreeSurfer segmentation to define gray-matter voxels in each subject. Figure 2.2 shows the effect of slice index on the ratio of the mean gray-matter intensity in each B scan to the mean intensity in the corresponding A scan.

We modeled the relation between slice index i and B/A ratio r as an exponential

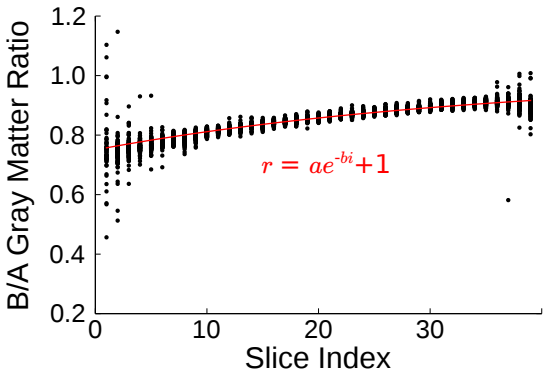


Figure 2.2: **Scan timing effects by slice.** The ratio of mean gray-matter intensities from B scans to immediately preceding A scans varies by slice. Shown are 5000 randomly selected ratios, and an exponential curve of best fit (red) describing the relation of all ratios to slice index.

function with free parameters a and b:

$$r = ae^{-bi} + 1 \tag{2.1}$$

Using the `optimize.curve_fit` function from the `scipy` (Jones et al., 2001–) package, with initial estimates $(a, b) = (-0.5, 0.1)$, we estimated $a = -0.243$ and $b = 0.0280$. In each raw (non-motion-corrected) functional volume, each slice i of the A scans was scaled by $1 - 0.243e^{-0.0280i}$ and copied into a new FsFast session, along with unmodified B scans. The resulting sessions were again motion corrected.

These same parameters were used to correct intensities for the final two subjects.

2.2.6.2 Estimation of individual event responses

We defined an **input-related** event as the estimated hemodynamic response to the auditory stimulus, and an **output-related** event as the estimated response to the initiation of the motor act (i.e., the GO signal). The specification of regressors for these events is illustrated in Figure 2.1. These regressors were used to construct

design matrices to estimate responses to individual events from the detrended scans using a GLM as described by Perrachione and Ghosh (2013), with modifications detailed below. Acquisitions in which there was at least a half-voxel (1.5mm) or greater shift from the previous volume were treated as motion outliers. Events where the magnitude of the theoretical HRF was greater than 10% of its maximum height during one of these outlier volumes were excluded from analysis. A total of 124 input-related events and 126 output-related events were excluded across subjects.

Each individual event was modeled by a canonical HRF convolved with a delta function, normalized to a maximum height of 1 (Figure 2.3A). For a scan that occurs during the course of the hemodynamic response, its contribution to the event was estimated as the mean of the HRF during its 2.5s duration. Because a volume was captured in 40 slices, we used an HRF temporal resolution (Nipype's `spm_hrf()` parameter `RT`) of 2.5s / 40 = 0.0625s.

The GLM estimation was performed piecewise by condition (Figure 2.3B), using custom Python scripts. A condition was defined as a (stimulus, event type) pair, where stimulus is a *specific syllable*, *silence*, or *noise*, and event type is either *input-linked* or *output-linked*. With 18 syllables used, this resulted in 40 conditions, between which all events may be labeled (i.e., /mɛl/$_{in}$, /mɛl/$_{out}$, /mɪl/$_{in}$, ...). When one of these conditions was modeled, all events labeled with that condition were modeled as individual events (with individual regressors), while all other events were collapsed into a single column for each of the other conditions (Figure 2.3C). Only the estimates of individual events were kept, while estimates of the collapsed conditions were discarded.

An input-linked dataset for subsequent MVPA analysis was created from the HRF estimates for each input-linked event, and an output-linked dataset was created like-

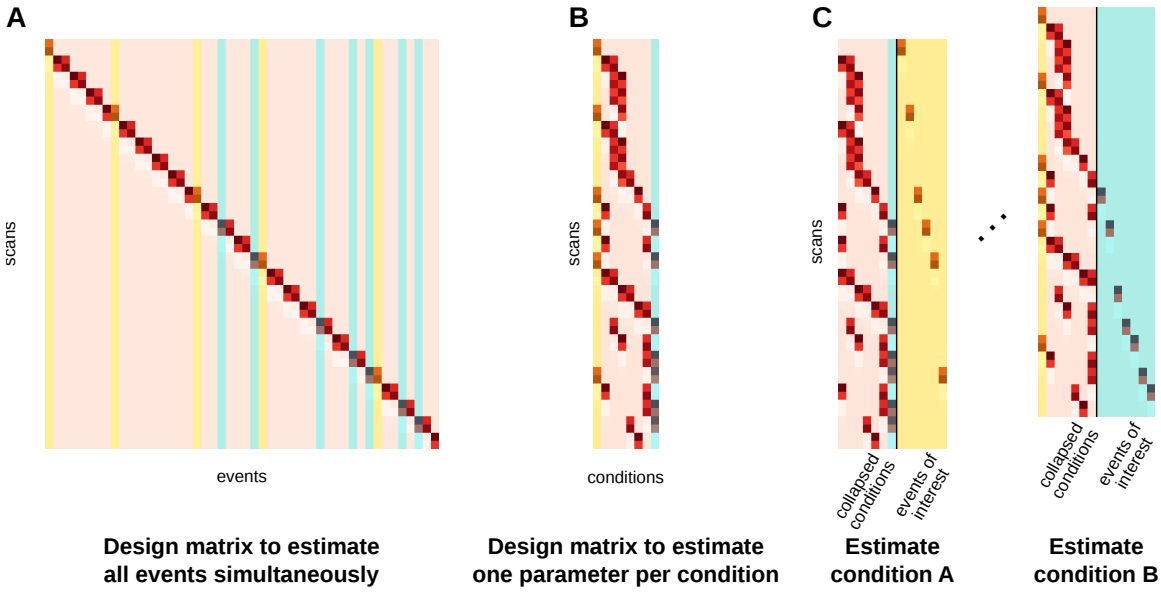


Figure 2.3: **Piecewise general linear model.** To model each event individually, events are grouped into *conditions*, two of which are shown in yellow and cyan in these schematized design matrices. The raw design matrix (A) contains a regressor for each event (stimulus/input or GO-signal/output), where each regressor is an impulse function convolved with an HRF, sampled at each scan time. The condition design matrix (B) contains a regressor for each condition, which is the sum of the event-related regressors within that condition. Condition-specific design matrices (C) contain one regressor for each *event* in the condition of interest and one for each *condition* of non-interest. The estimates for the events for each condition of interest are reconstructed to produce input- and output-related datasets with one estimate per-trial.

wise. These datasets were considered separately in further analyses.

2.2.7 Surface Searchlight

We implemented a cortical surface searchlight (e.g., Chen et al., 2011; Oosterhof et al., 2011) within the PyMVPA framework (Hanke et al., 2009), using a geodesic radius of 9mm centered on each vertex of interest, calculated using Dijkstra's algorithm (Dijkstra, 1959) on the FreeSurfer mesh halfway between the pial and white matter surfaces (Chen et al.'s "graymid" surface). To avoid resampling or excess smoothing, our analyses were performed on volumetric (voxel) data in subject-native space. At a given voxel, the nearest surface vertex to the center of that voxel was selected as the center of a 9mm disk, and all voxels intersecting that disk constituted the searchlight at that voxel. The result of a searchlight analysis creates an *accuracy map*, a volume in which each voxel contains a statistic for the searchlight centered at that voxel. For group analyses, each subject's accuracy map was projected onto their graymid surface.

The statistic of interest in our analyses was leave-one-run-out cross validation accuracy, and a linear C support vector machine (CSVM) was chosen as the classifier. For input-related analyses, the class label for each volume was the identity of the stimulus presented; for output-related datasets, the label was the identity of the speech sound produced. We performed independent searchlight analyses with full syllable (C_1VC_2 trigram), onset (C_1), vowel (V), and coda (C_2) class labels.

The PyMVPA `Balancer` mapper was used to ensure that, in each fold, the training and validation sets each contained identical numbers of volumes in each class, to reduce bias in SVM construction and validation. When input sets are unbalanced, the `Balancer` produces two random, balanced sets, each of which is classified, and

the results are averaged.

2.2.7.1 Nonparametric significance testing

For each classification result, each subject's accuracy map was registered to the FreeSurfer `fsaverage` template for group analysis. A map of *relative* accuracy was created by subtracting the mean cross-validation accuracy across vertices from each vertex, such that the resulting image had a mean of zero. One-tailed t-tests were used to assess whether individual vertices had consistently high relative accuracy across subjects (Lee et al., 2012). The map was subjected to a $p < 0.05$ (uncorrected) threshold, and we defined clusters as connected subgraphs (of supra-threshold vertices) on the surface, which were used for further inference.

To perform cluster-extent based thresholding, we generated a null distribution of chance cluster sizes. For each subject, we permuted class labels 100 times and retrained classifiers to generate random accuracy maps (see Stelzer et al., 2013); each such map was registered to the `fsaverage` template. Choosing one random accuracy map from each subject and subjecting it to the thresholding process described above produced a set of cluster sizes. In this way, we constructed 10^4 sets of empirical chance cluster sizes, providing a null-distribution for assigning cluster-level significance. Cluster-level thresholds of $p < 0.01$ were used for phoneme-level analyses. A stricter threshold of $p < 0.001$ was used for syllable-level analyses; due to the lower theoretical chance-level accuracy (1/18, or $0.0\bar{5}\%$) random fluctuations produce a larger number of small clusters, with the effect of decreasing p-values for all clusters. This threshold was used to result in a minimum cluster extent that was more similar to that used for phoneme-level analysis.

2.2.8 Acoustic Analysis

Classification accuracy is driven by correlations between class labels and patterns of voxel activations. These patterns may, for instance, be consistent with a "categorical" representation, in which a voxel responds identically (modulo noise) to all stimuli drawn from the same class, or with a "continuous" representation, in which a voxel responds to an acoustic feature, which in turn correlates with class labels. Vowels, for instance, form clusters in the F_1, F_2 formant frequency space (Peterson and Barney, 1952); a handful of voxels whose responses correlate with formants, then, could drive above-chance vowel classification.

To assess whether formant frequency related variance in the BOLD signal was driving classification accuracies, we constructed alternative design matrices for the event-estimation GLM. In these matrices, we added nuisance regressors for the first and second formants, as well as the ratio F_2/F_1, convolved by the HRF. With the resulting dataset, we performed an additional set of searchlight analyses, as above. Mapping into the `fsaverage` template and subtracting the empirical chance classification accuracy, we performed a two-sample t-test at each vertex between classification accuracies, where one sample contains the accuracies from the original datasets, and the other from this alternative dataset to determine if removing formant-related variance decreased classification accuracy (as would be expected if an ROI was "tracking" formant values).

2.3 Results

2.3.1 Behavioral results

One subject (S13) repeated syllables with an error rate of 20.74%. For all other subjects, the mean error rate was 1.61% ($\sigma = 1.35\%$), with minimum and maximum error rates of 0 and 5.40%. S13 was excluded from all subsequent analyses. One subject's vocalizations were not recorded for one run; however, this subject made no errors in the recorded runs, and was included in all further analyses with the assumption that no errors were made in this first run. Inconsistent recording quality resulted in 0-30 wholly or partially missing vocalizations, out of 352 total trials (mean=13.2, std=8.8).

2.3.2 Overall task effects

To verify the effectiveness of individual event modeling with the piecewise general linear model, we first calculated contrasts between task and control conditions for input- and output-linked event estimates. Within each dataset, a *condition* is defined here as the set of estimated coefficients (betas) associated with a particular stimulus type. In this case, task-related (syllable responses linked to either input or output events) betas and control-related (silence or noise) betas constituted the conditions. The beta estimates for a condition were taken to be the mean of the betas for all individual events in that condition.

We contrasted the average voxel-wise betas for the task and control conditions in each subject's native space, then mapped into their FreeSurfer surface representation and onto the `fsaverage` template. At each vertex in the `fsaverage` space, we performed a two-tailed, one-sample t-test across participants. Figure 2.4 shows positive t-statistics that surpass an uncorrected threshold of $p < 0.05$. This liberal thresh-

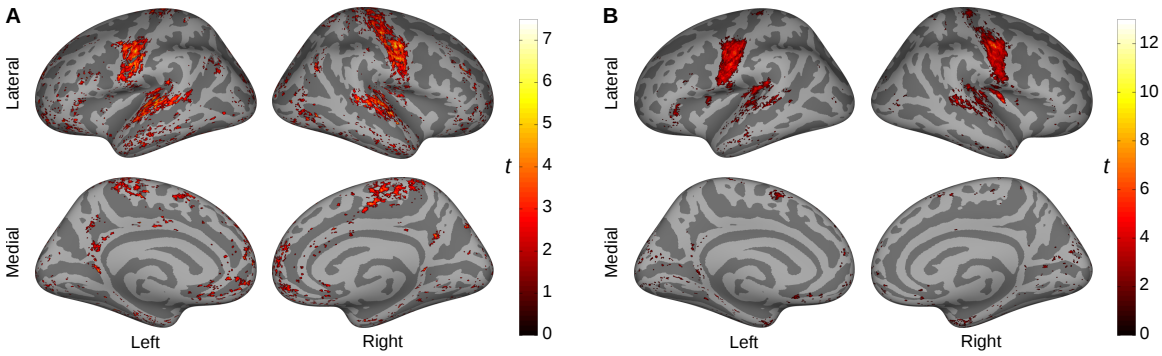


Figure 2.4: **Overall task effects.** Contrasts between mean estimates of individual event responses for task and control (silence, noise) conditions in the (A) input- and (B) output-linked hemodynamic responses. Shown are t-statistics (13 dof), thresholded at $p < 0.05$ (uncorrected).

old was used due to the unconventional approach of using estimates from individual events.

The input-related responses were primarily localized to bilateral superior temporal gyrus, planum temporale, and ventral motor cortex. Output-related responses were localized primarily to bilateral sensorimotor cortex, with fewer significant activations in the auditory cortex.

2.3.3 Speech sound information mapping

An MVPA searchlight analysis identifies "informative" regions with regard to the class labels assigned to a dataset, using cross-validation accuracy as the statistic of interest. Here we present results for input- and output-related datasets, analyzed with vowel, onset and coda labels, as well as with whole syllable labels. In Figures 2.5 and 2.9-2.11, the values rendered on inflated cortical surfaces are the mean raw (*i.e.*, not normalized by the voxel-wise average) cross-validation accuracies across subjects, thresholded, using a nonparametric cluster-level significance test, with a vertex-wise uncorrected threshold of $p < 0.05$ and cluster-wise thresholds of $p < 0.01$

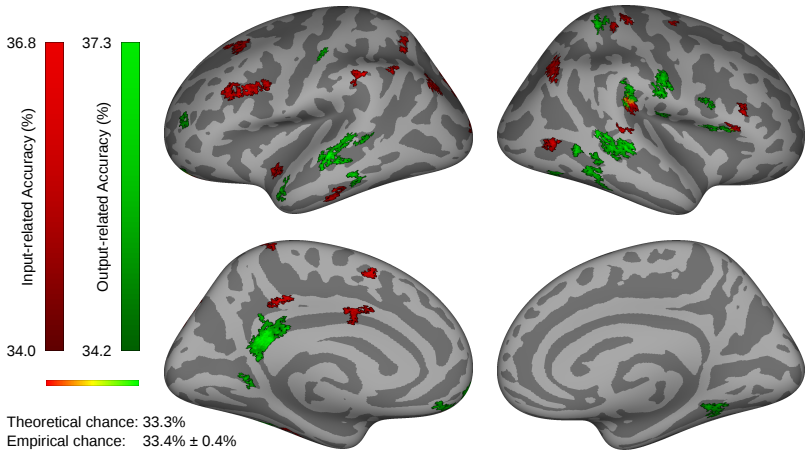


Figure 2.5: **Vowel accuracy.** Regions with significant cross-validation accuracy for decoding of vowel identity based on input- (red) and output-related (green) hemodynamic response estimates. Regions of overlap are shown in yellow, with greater input accuracy appearing more toward red and greater output accuracy appearing more toward green. For output-related analysis, the identity of the spoken vowel was used as the classification target if the vocalization differed from the presented stimulus. Results presented at $p < 0.05$ uncorrected, thresholded by cluster size ($p < 0.01$).

(for phoneme-level analyses) or $p < 0.001$ (for syllable-level analyses).

2.3.3.1 Vowels

Figure 2.5 shows the mean accuracies for classifiers trained on vowel identity, irrespective of the surrounding consonants, with a cluster-level threshold of $p < 0.01$. This and other figures (except Figure 2.6) render, in different hues, areas with significant information both based on input responses and output responses, allowing a direct comparison across stages of the task. Empirical chance accuracies, indicated in Figures 2.5, 2.9, 2.10 and 2.11, represent the across-subject average and standard deviation of the spatial means of each accuracy map. For input-linked analyses, we found a superior temporal cluster centered in right posterior superior temporal gyrus (STg), in addition to inferior parietal clusters in left supramarginal gyrus (SMg), right parietal operculum (PO), and angular gyrus (AG). In addition we found a prominent

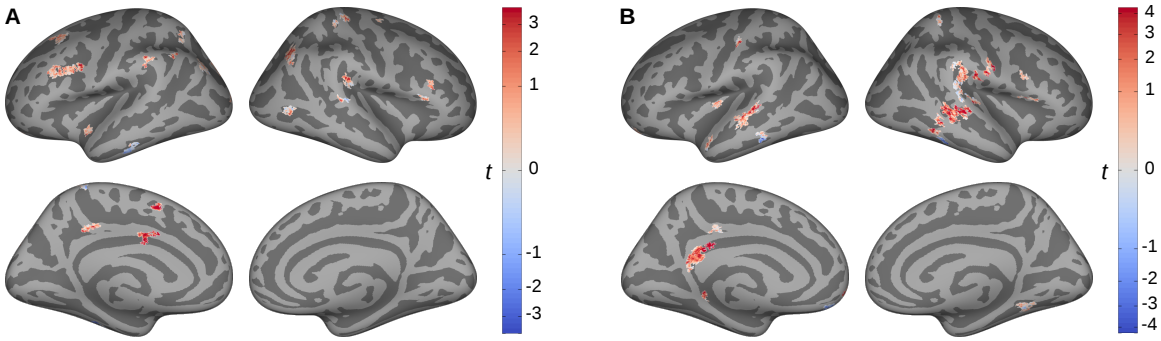


Figure 2.6: **Differences in vowel accuracy with and without acoustic regressors.** Two-sample t-statistics between relative vowel classification accuracies using event modeling with and without acoustic nuisance regressors, for (A) input- and (B) output-related responses. Maps are thresholded by ($p < 0.01$) cluster thresholds from the corresponding searchlight analyses (Figure 2.5). Positive (red) values indicate superior relative classification without nuisance regressors; negative (blue) values indicate superior classification with nuisance regressors. Color axes are scaled quadratically to enhance visual separation.

cluster centered in left inferior frontal sulcus (IFs), and two smaller clusters in right IFs and inferior frontal gyrus pars opercularis (IFo). For output-linked responses, prominent clusters of predictive information were found bilaterally in posterior superior temporal sulcus (STs) and in right PO, along with inferior frontal clusters in right IFo, frontal operculum and ventral premotor cortex (vPMC). A strong, right-lateralized ventral somatosensory cortical cluster also appears, as well as a large cluster of predictive voxels in the left posterior pericallosal sulcus / cingulate gyrus.

Locations of peak classification rates for all such clusters are listed in Tables 2.2 and 2.3. Theoretical chance accuracy for classifying three vowels was $1/3 = 0.\bar{3}$, and peak group mean accuracies ranged from 0.351 to 0.373. Note, however, that clusters in the present analysis (and all classification analyses presented) were defined by increases over the *empirical* chance accuracy rate.

In order to determine the extent to which formant frequency related information was driving classification accuracy for vowels (the most acoustically salient segments),

Input - Vowel

Coordinate (x, y, z)			Accuracy	p-value	Extent	Hemisphere	aparc	SLaparc17
-35	13	25	0.362	2.9×10^{-4}	561	lh	caudalmiddlefrontal	pIFs
-16	-87	31	0.366	1.1×10^{-3}	343	lh	superiorparietal	OC
-32	-80	29	0.362	1.2×10^{-3}	324	lh	inferiorparietal	AG
-26	14	43	0.361	1.9×10^{-3}	259	lh	caudalmiddlefrontal	pMFg
-58	-40	40	0.365	3.4×10^{-3}	185	lh	supramarginal	aSMg
-3	0	33	0.359	3.9×10^{-3}	170	lh	posteriorcingulate	midCG
-14	-42	70	0.359	3.9×10^{-3}	170	lh	paracentral	dSC
-6	-35	44	0.360	4.7×10^{-3}	151	lh	posteriorcingulate	pCG
-56	-26	-28	0.363	4.9×10^{-3}	148	lh	inferiortemporal	pITg
-15	48	-21	0.360	5.0×10^{-3}	145	lh	lateralorbitofrontal	FP
-40	-8	-14	0.357	5.2×10^{-3}	142	lh	insula	PP
-50	-57	37	0.364	5.3×10^{-3}	139	lh	inferiorparietal	AG
-41	-29	-23	0.359	6.9×10^{-3}	114	lh	fusiform	pTF
-26	-58	51	0.360	7.9×10^{-3}	103	lh	superiorparietal	SPL
-20	-99	9	0.356	8.8×10^{-3}	94	lh	lateraloccipital	OC
-7	8	52	0.362	9.3×10^{-3}	90	lh	superiorfrontal	preSMA
-27	-58	45	0.366	9.8×10^{-3}	86	lh	superiorparietal	SPL
33	-68	39	0.368	2.6×10^{-4}	588	rh	inferiorparietal	AG
44	-34	24	0.363	6.4×10^{-4}	422	rh	supramarginal	PO
29	-30	64	0.363	2.7×10^{-3}	212	rh	postcentral	dSC
43	-64	2	0.363	5.8×10^{-3}	130	rh	lateraloccipital	MTO
29	-36	57	0.361	7.5×10^{-3}	107	rh	postcentral	dSC
65	-32	2	0.358	7.8×10^{-3}	104	rh	superiortemporal	pdSTs
50	22	7	0.353	8.3×10^{-3}	99	rh	parstriangularis	vIFo
47	27	18	0.359	1.0×10^{-2}	85	rh	parsopercularis	pIFs
24	-11	54	0.360	1.0×10^{-2}	85	rh	precentral	mdPMC

Table 2.2: Significant clusters for vowel-classification in input-linked dataset. Columns: (a) Coordinates of peak accuracy, in MNI305 space; (b) mean cross-validation accuracy, across subjects; (c) empirical cluster-wise p-value; (d) extent (# of contiguous vertices); (e) hemisphere; (f) Desikan et al. (2006) atlas label; (g) Tourville and Guenther (2012) speech-related atlas label.

we trained classifiers on events modeled with and without acoustic nuisance regressors specifying the first and second formant frequencies (and their ratio) of vowels heard or produced. The formant values extracted for all stimuli as well as a random subset of subjects' productions are shown in Figures 2.7-2.8. Vertex-wise two-sample t-tests were conducted between the relative vowel classification rates resulting from searchlight analyses for these two models.

Output - Vowel (vocalized)

Coordinate (x, y, z)			Accuracy	p-value	Extent	Hemisphere	aparc	SLaparc17
-54	-22	-5	0.373	6.4×10^{-5}	921	lh	superiortemporal	pdSTs
-4	-48	18	0.370	1.0×10^{-4}	801	lh	isthmuscingulate	pCG
-53	8	-20	0.363	3.6×10^{-3}	182	lh	superiortemporal	adSTs
-5	49	-24	0.366	4.4×10^{-3}	160	lh	medialorbitofrontal	FP
-8	62	-13	0.367	5.7×10^{-3}	134	lh	frontalpole	FP
-20	-51	-3	0.360	5.9×10^{-3}	130	lh	lingual	LG
-65	-33	-17	0.361	6.5×10^{-3}	121	lh	middletemporal	pMTg
-48	-18	50	0.363	6.5×10^{-3}	121	lh	postcentral	dSC
-34	51	15	0.361	7.3×10^{-3}	111	lh	rostralmiddlefrontal	FP
-52	-37	0	0.359	7.8×10^{-3}	105	lh	bankssts	pdSTs
-18	46	-16	0.363	8.1×10^{-3}	102	lh	lateralorbitofrontal	FP
-41	-43	-22	0.366	8.2×10^{-3}	101	lh	fusiform	pTF
47	-29	-2	0.367	5.1×10^{-5}	989	rh	superiortemporal	pdSTs
64	-14	25	0.369	1.8×10^{-4}	688	rh	postcentral	vSC
54	-31	30	0.362	5.2×10^{-4}	472	rh	supramarginal	PO
29	-53	-6	0.360	1.6×10^{-3}	290	rh	lingual	LG
29	-44	62	0.368	1.7×10^{-3}	279	rh	superiorparietal	SPL
59	-48	-20	0.365	2.0×10^{-3}	255	rh	inferiortemporal	pITg
39	9	10	0.361	5.6×10^{-3}	136	rh	parsopercularis	pFO
44	-14	20	0.359	6.7×10^{-3}	119	rh	postcentral	pCO
40	-67	-16	0.363	7.4×10^{-3}	110	rh	fusiform	TOF
46	8	18	0.357	7.9×10^{-3}	104	rh	parsopercularis	dIFo
44	19	7	0.362	9.2×10^{-3}	92	rh	parsopercularis	vIFo
64	-46	-6	0.356	9.4×10^{-3}	90	rh	middletemporal	pMTg
56	-54	-3	0.365	9.4×10^{-3}	90	rh	middletemporal	MTO

Table 2.3: Significant clusters for vowel-classification in output-linked dataset. Columns: (a) Coordinates of peak accuracy, in MNI305 space; (b) mean cross-validation accuracy, across subjects; (c) empirical cluster-wise p-value; (d) extent (# of contiguous vertices); (e) hemisphere; (f) Desikan et al. (2006) atlas label; (g) Tourville and Guenther (2012) speech-related atlas label.

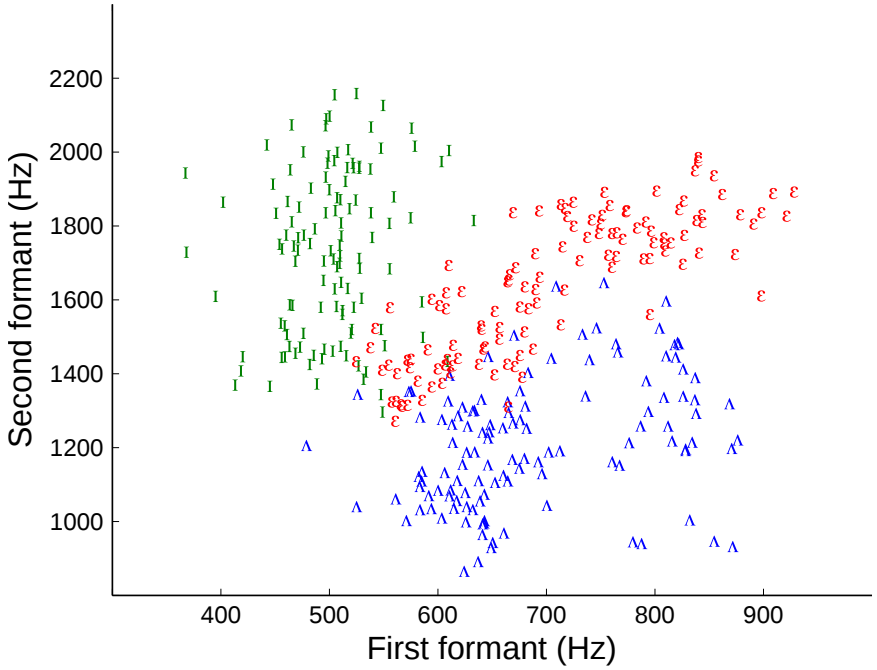


Figure 2.7: **Formant frequencies of stimuli.** Formant frequencies were estimated from the approximate mid-points of vowels using custom PRAAT scripts (see Methods). Shown are the first two formant estimates from all 360 stimuli used, labeled by vowel. Formants from syllables containing /ɪ/ are shown in green; /ɛ/ in red; and /ʌ/ in blue.

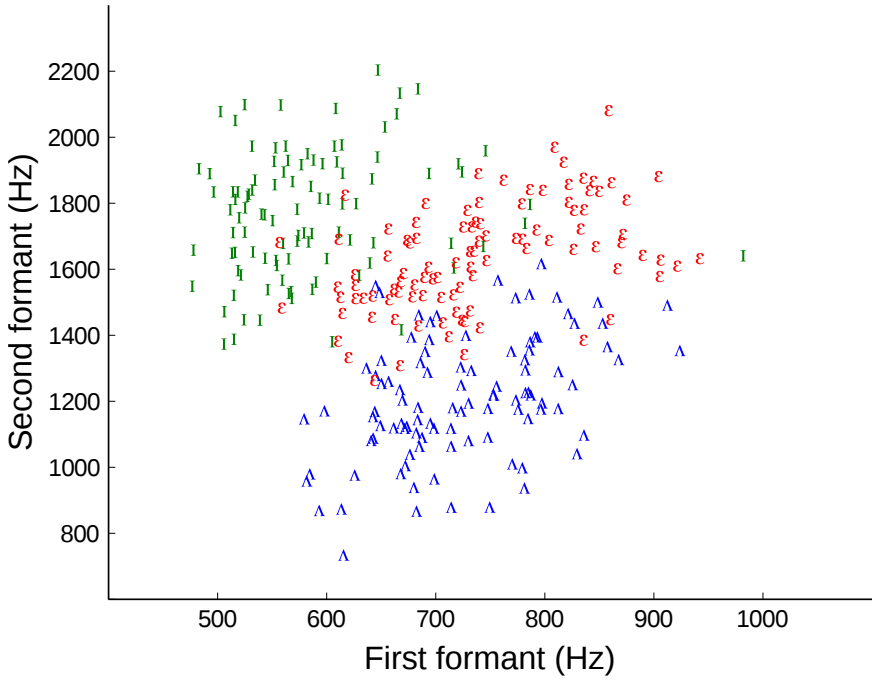


Figure 2.8: **Formant frequencies of subjects' vocalizations.** Subject syllable productions were extracted and labeled manually. Formant frequencies were estimated from the approximate mid-points of vowels using custom PRAAT scripts (see Methods). Shown are the first two formants estimates from 300 vocalizations, randomly selected from across all subjects. Formants from syllables containing the vowel /ɪ/ are shown in green; /ɛ/ in red; and /ʌ/ in blue.

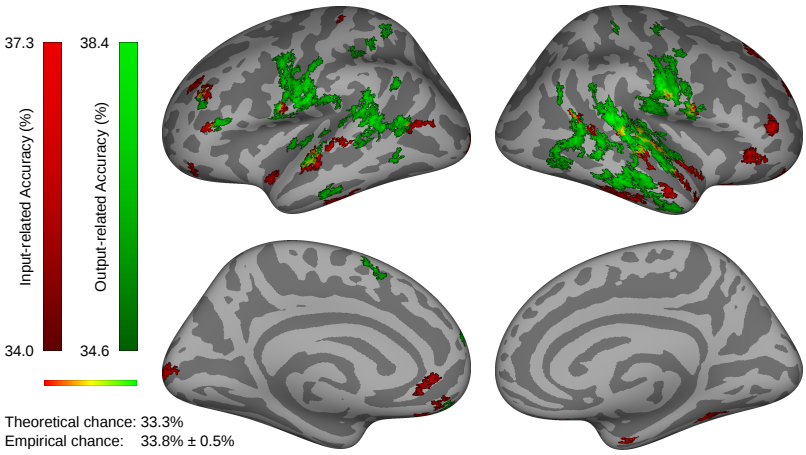


Figure 2.9: **Onset consonant accuracy.** Regions with significant cross-validation accuracy for decoding onset consonant identity based on input- (red) and output-related (green) hemodynamic response estimates. Regions of overlap are shown in yellow, with greater input accuracy appearing more red and greater output accuracy appearing more green. For output-related analysis, the identity of the spoken consonant was used as the classification target, when the vocalization differed from the presented stimulus. Results presented at $p < 0.05$ uncorrected, thresholded by cluster size ($p < 0.01$).

Figure 2.6 shows the resulting t-statistics, masked by the clusters identified in the input- and output-linked vowel searchlight analyses. In both input- and output-related analyses, nearly all regions showed positive t-statistics, indicating that removing systematic variations related to formants resulted in *decreased* classification accuracy relative to empirical chance accuracy, in regions identified as containing information about vowel identities. However, no regions reached statistical significance when applying a false discovery rate correction.

2.3.3.2 Consonants

Searchlight classification using consonants as class labels (performed separately for onset and coda positions) revealed a pattern of areas largely distinct from that found for vowels. Figure 2.9 shows mean cross-validation accuracies for classifiers trained

on onset identity, with a cluster threshold of $p < 0.01$. The input-linked analysis shows predominantly superior temporal clusters, bilaterally, in addition to a few small clusters in inferior frontal regions. We found predictive information in right inferior frontal gyrus (IFg) pars orbitalis and anterior STg. Clusters with informative activation patterns were also found in left anterior planum polare (PP), as in the vowel analysis. Further findings included superior temporal sensitivity centered in STs bilaterally, in right-lateralized planum temporale (PT) and Heschl's gyrus (H). In inferior frontal regions, we found a left-lateralized cluster centered in IFs (anterior to that observed for vowel identity) and right-lateralized vPMC.

For the corresponding output-linked analyses, ventral somatosensory and motor cortices provided strong predictive information. In superior temporal regions, we found a left posterior cluster spanning posterior STs/STg/PP and smaller clusters centered in anterior STs, Heschl's gyrus and posterior insula, as well as a cluster spanning much of the right STg, into PT/H/PP. Finally, we found a left lateralized cluster in ventral IFs/dorsal IFg pars triangularis (IFt).

Locations of peak classification rates for all such clusters are listed in Tables 2.4-2.5. Theoretical chance accuracy for classifying three consonants was again $0.\bar{3}$, and peak group mean accuracies ranged from 0.354 to 0.384.

Input - Onset

Coordinate (x, y, z)			Accuracy	p-value	Extent	Hemisphere	aparc	SLaparc17
-10	-99	3	0.369	4.2×10^{-4}	495	lh	pericalcarine	OC
-50	-18	-12	0.365	6.8×10^{-4}	413	lh	superiortemporal	pdSTs
-56	-28	-31	0.365	1.7×10^{-3}	277	lh	inferiortemporal	pITg
-44	-3	-19	0.362	3.2×10^{-3}	192	lh	superiortemporal	PP
-43	34	26	0.362	3.5×10^{-3}	184	lh	rostralmiddlefrontal	aMFg
-43	-68	11	0.361	3.5×10^{-3}	183	lh	inferiorparietal	OC
-59	-2	12	0.356	3.7×10^{-3}	177	lh	precentral	vMC
-25	35	26	0.360	4.0×10^{-3}	167	lh	rostralmiddlefrontal	aMFg
-11	39	-22	0.365	4.6×10^{-3}	154	lh	lateralorbitofrontal	FP
-5	41	-4	0.359	5.0×10^{-3}	146	lh	rostralanteriorcingulate	aCG
-56	-35	0	0.360	5.0×10^{-3}	146	lh	bankssts	pdSTs
-60	-23	-4	0.365	5.0×10^{-3}	145	lh	superiortemporal	pdSTs
-25	38	-12	0.358	6.6×10^{-3}	118	lh	lateralorbitofrontal	FP
-27	-30	64	0.357	7.5×10^{-3}	107	lh	postcentral	dSC
-45	35	8	0.361	8.3×10^{-3}	99	lh	parstriangularis	aIFs
-3	45	-21	0.358	9.4×10^{-3}	89	lh	medialorbitofrontal	FMC
53	-33	-28	0.373	4.7×10^{-4}	475	rh	inferiortemporal	pITg
57	-13	3	0.371	1.2×10^{-3}	330	rh	superiortemporal	H
44	30	-7	0.367	1.4×10^{-3}	305	rh	parstriangularis	FOC
49	-7	-10	0.365	2.2×10^{-3}	240	rh	superiortemporal	PP
39	40	6	0.366	2.3×10^{-3}	235	rh	rostralmiddlefrontal	FP
60	-18	-3	0.364	2.5×10^{-3}	223	rh	superiortemporal	pdSTs
59	5	-9	0.360	3.0×10^{-3}	202	rh	superiortemporal	aSTg
18	49	-17	0.370	3.4×10^{-3}	185	rh	lateralorbitofrontal	FP
27	-43	-18	0.366	3.6×10^{-3}	180	rh	fusiform	pTF
14	20	-21	0.369	3.6×10^{-3}	179	rh	lateralorbitofrontal	FOC
17	15	-24	0.368	3.8×10^{-3}	175	rh	lateralorbitofrontal	FOC
21	32	44	0.362	4.0×10^{-3}	169	rh	superiorfrontal	SFg
53	-48	7	0.358	4.4×10^{-3}	157	rh	bankssts	MTO
58	-5	-22	0.361	4.7×10^{-3}	152	rh	middletemporal	avSTs
59	3	7	0.361	4.8×10^{-3}	149	rh	precentral	vPMC
60	-8	19	0.359	5.6×10^{-3}	133	rh	postcentral	vSC
23	54	26	0.362	5.7×10^{-3}	132	rh	rostralmiddlefrontal	FP
66	-30	1	0.367	7.5×10^{-3}	107	rh	superiortemporal	pdSTs
25	-3	-36	0.364	7.7×10^{-3}	105	rh	entorhinal	aPH
44	18	-28	0.363	7.8×10^{-3}	104	rh	superiortemporal	TP
52	-55	15	0.357	9.4×10^{-3}	89	rh	inferiorparietal	AG

Table 2.4: Peak classification accuracies for input-related beta estimates, classified on the identity of the onset consonant of each stimulus.

Output - Onset (vocalized)

Coordinate (x, y, z)			Accuracy	p-value	Extent	Hemisphere	aparc	SLaparc17
-57	-3	32	0.377	1.2×10^{-7}	3060	lh	precentral	vMC
-45	-58	12	0.379	3.6×10^{-6}	1780	lh	inferiorparietal	MTO
-58	-10	-6	0.373	9.3×10^{-4}	372	lh	superiortemporal	adSTs
-34	-41	40	0.368	2.4×10^{-3}	232	lh	superiorparietal	SPL
-33	-50	64	0.370	2.9×10^{-3}	208	lh	superiorparietal	SPL
-47	28	11	0.370	3.5×10^{-3}	186	lh	parstriangularis	aIFs
-55	-23	-27	0.369	3.6×10^{-3}	184	lh	inferiortemporal	pITg
-51	-60	28	0.374	4.5×10^{-3}	159	lh	inferiorparietal	AG
-44	-22	-2	0.374	4.5×10^{-3}	159	lh	superiortemporal	PP
-37	-45	49	0.364	4.9×10^{-3}	150	lh	superiorparietal	SPL
-41	40	27	0.366	5.0×10^{-3}	147	lh	rostralmiddlefrontal	aMFg
-6	10	57	0.368	5.2×10^{-3}	142	lh	superiorfrontal	preSMA
-54	-38	2	0.369	6.2×10^{-3}	125	lh	bankssts	pdSTs
-51	-37	0	0.363	6.4×10^{-3}	123	lh	bankssts	pdSTs
-6	51	-25	0.367	7.1×10^{-3}	113	lh	medialorbitofrontal	FP
-7	62	21	0.368	7.4×10^{-3}	109	lh	superiorfrontal	FP
-55	-61	-3	0.368	7.9×10^{-3}	104	lh	middletemporal	MTO
-37	-19	-4	0.365	8.6×10^{-3}	97	lh	insula	pINS
-48	-26	6	0.367	9.8×10^{-3}	87	lh	transversetemporal	H
62	-8	32	0.384	1.2×10^{-7}	3195	rh	postcentral	vSC
62	-38	10	0.382	1.2×10^{-7}	3826	rh	bankssts	pSTg
55	-36	-17	0.375	2.3×10^{-5}	1212	rh	inferiortemporal	pMTg
50	-58	0	0.373	3.6×10^{-5}	1075	rh	middletemporal	MTO
64	-42	-3	0.376	1.5×10^{-4}	721	rh	middletemporal	pvSTs
59	4	6	0.374	7.6×10^{-4}	405	rh	precentral	vPMC
48	-3	-32	0.372	1.6×10^{-3}	292	rh	middletemporal	aITg
40	-17	63	0.365	4.1×10^{-3}	169	rh	precentral	dMC
51	-62	24	0.368	5.7×10^{-3}	134	rh	inferiorparietal	AG
51	-1	49	0.368	6.4×10^{-3}	123	rh	precentral	midPMC
35	-36	62	0.366	8.2×10^{-3}	101	rh	postcentral	dSC

Table 2.5: Peak classification accuracies for input-related beta estimates, classified on the identity of the onset consonant of each vocalization.

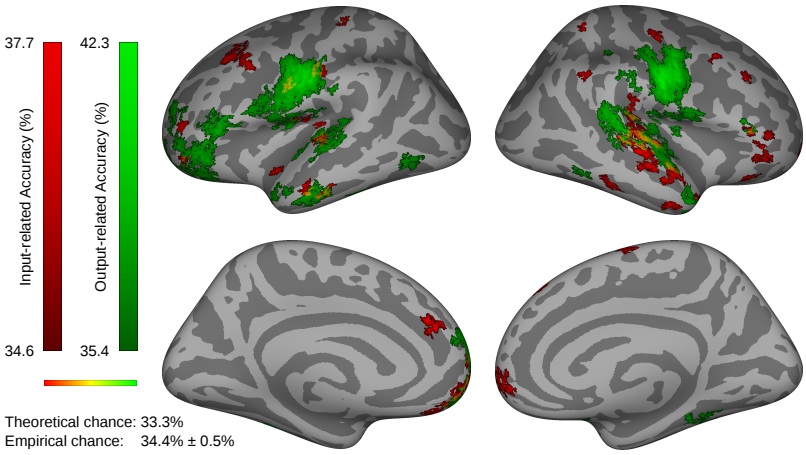


Figure 2.10: **Coda consonant accuracy.** Regions with significant cross-validation accuracy for decoding coda consonant identity based on input- (red) and output-related (green) hemodynamic response estimates. Regions of overlap are shown in yellow, with greater input accuracy appearing more red and greater output accuracy appearing more green. For output-related analysis, the identity of the spoken consonant was used as the classification target, when the vocalization differed from the presented stimulus. Results presented at $p < 0.05$ uncorrected, thresholded by cluster size ($p < 0.01$).

We next examined the information associated with prediction of coda consonant identity. Mean cross-validation accuracies for classifiers trained on coda identity are shown in Figure 2.10. For input-linked analyses, we found superior temporal sensitivity in left Heschl's gyrus and PP, a large right-hemisphere cluster spanning mid-STs/STg, Heschl's gyrus and PT, and a further cluster centered in right anterior STg. Right lateralized inferior parietal clusters were found in AG and PO, and right-lateralized inferior frontal clusters were found in IFt.

For the corresponding output-linked analyses, ventral sensorimotor cortices were prominent, bilaterally, with especially high group mean classification accuracies. As seen for classification of onsets, superior temporal lobe sensitivity was localized to posterior STg in the left hemisphere, and in posterior and anterior STg in the right hemisphere. Bilateral clusters were also found in SMg and IFt.

Locations of peak classification rates for all such clusters are listed in Tables 2.6-2.7. Theoretical chance accuracy for classifying three consonants was again $0.\bar{3}$. Peak group mean accuracies for the input-linked analysis ranged from 0.363 to 0.377, while those for the output-linked analysis ranged from 0.370 to 0.423.

Input - Coda

Coordinate (x, y, z)			Accuracy	p-value	Extent	Hemisphere	aparc	SLaparc17
-29	22	40	0.377	1.8×10^{-4}	654	lh	caudalmiddlefrontal	aMFg
-5	54	-16	0.375	1.2×10^{-3}	323	lh	medialorbitofrontal	FP
-51	-20	7	0.376	2.2×10^{-3}	241	lh	transversetemporal	H
-6	43	-24	0.370	2.6×10^{-3}	218	lh	medialorbitofrontal	FMC
-9	34	27	0.371	3.6×10^{-3}	181	lh	superiorfrontal	SFg
-52	-6	23	0.367	4.0×10^{-3}	168	lh	precentral	vMC
-38	-15	20	0.365	5.0×10^{-3}	146	lh	insula	pCO
-32	-32	69	0.367	5.3×10^{-3}	139	lh	postcentral	dSC
-58	-18	42	0.366	6.1×10^{-3}	126	lh	postcentral	vSC
-59	-12	-26	0.370	6.1×10^{-3}	125	lh	middletemporal	aMTg
-38	-29	13	0.371	6.3×10^{-3}	123	lh	transversetemporal	H
-45	-8	-14	0.369	6.8×10^{-3}	116	lh	superiortemporal	PP
-46	-10	12	0.366	6.9×10^{-3}	115	lh	postcentral	pCO
-7	60	6	0.367	7.2×10^{-3}	111	lh	superiorfrontal	FP
-39	51	9	0.373	7.4×10^{-3}	109	lh	rostralmiddlefrontal	FP
-56	-20	36	0.363	7.5×10^{-3}	107	lh	postcentral	vSC
-55	-21	-26	0.367	8.0×10^{-3}	102	lh	middletemporal	pITg
-36	52	-10	0.367	9.8×10^{-3}	86	lh	parsorbitalis	FP
52	-15	6	0.376	4.2×10^{-6}	1659	rh	transversetemporal	H
56	4	-16	0.375	7.1×10^{-4}	404	rh	superiortemporal	aSTg
8	51	-5	0.371	1.2×10^{-3}	320	rh	medialorbitofrontal	FP
48	40	-13	0.371	1.8×10^{-3}	269	rh	parsorbitalis	FP
36	-15	52	0.373	1.9×10^{-3}	262	rh	precentral	dMC
48	-19	-12	0.368	2.9×10^{-3}	206	rh	superiortemporal	pdSTs
12	-5	71	0.367	3.0×10^{-3}	201	rh	superiorfrontal	mdPMC
51	-6	-30	0.374	4.0×10^{-3}	169	rh	middletemporal	aMTg
56	-50	38	0.367	4.4×10^{-3}	157	rh	inferiorparietal	AG
54	-37	-19	0.367	5.2×10^{-3}	141	rh	inferiortemporal	pITg
46	-30	24	0.369	5.7×10^{-3}	132	rh	supramarginal	PO
43	19	-28	0.363	6.6×10^{-3}	118	rh	superiortemporal	TP
9	39	52	0.370	6.8×10^{-3}	116	rh	superiorfrontal	SFg
39	27	32	0.370	7.4×10^{-3}	108	rh	rostralmiddlefrontal	aMFg
8	67	-3	0.368	7.7×10^{-3}	105	rh	frontalpole	FP
26	29	-13	0.370	8.3×10^{-3}	99	rh	lateralorbitofrontal	FOC
43	39	3	0.365	8.7×10^{-3}	95	rh	parstriangularis	FP
43	16	47	0.373	8.8×10^{-3}	94	rh	caudalmiddlefrontal	pMFg
34	-51	64	0.371	9.0×10^{-3}	93	rh	superiorparietal	SPL
54	30	3	0.365	1.0×10^{-2}	85	rh	parstriangularis	dIFt

Table 2.6: Peak classification accuracies for input-related beta estimates, classified on the identity of the coda consonant of each stimulus.

Output - Coda (vocalized)

Coordinate (x, y, z)			Accuracy	p-value	Extent	Hemisphere	aparc	SLaparc17
-53	-12	30	0.414	$< 1 \times 10^{-7}$	5813	lh	postcentral	vSC
-52	28	3	0.388	4.8×10^{-7}	2476	lh	parstriangularis	vIFt
-64	-17	-1	0.380	1.1×10^{-5}	1435	lh	superiortemporal	pSTg
-15	59	-8	0.381	2.5×10^{-4}	614	lh	rostralmiddlefrontal	FP
-10	66	12	0.380	2.8×10^{-4}	588	lh	superiorfrontal	FP
-60	-13	-25	0.385	3.5×10^{-4}	548	lh	middletemporal	aMTg
-27	48	15	0.377	1.0×10^{-3}	354	lh	rostralmiddlefrontal	FP
-46	14	-28	0.379	1.4×10^{-3}	308	lh	superiortemporal	TP
-40	-40	-24	0.381	1.6×10^{-3}	293	lh	fusiform	pTF
-47	-62	-8	0.377	1.8×10^{-3}	276	lh	inferiortemporal	ITO
-50	-8	-3	0.373	6.7×10^{-3}	119	lh	superiortemporal	PP
-38	31	16	0.378	7.2×10^{-3}	112	lh	rostralmiddlefrontal	aIFs
-57	-0	-26	0.371	8.0×10^{-3}	103	lh	middletemporal	aMTg
-62	-27	23	0.378	9.9×10^{-3}	86	lh	supramarginal	aSMg
53	-12	33	0.423	6.0×10^{-8}	4511	rh	postcentral	vSC
59	-16	4	0.401	1.2×10^{-7}	3387	rh	superiortemporal	PT
57	1	7	0.382	9.5×10^{-5}	823	rh	precentral	aCO
44	13	-32	0.393	8.3×10^{-4}	391	rh	superiortemporal	TP
55	-27	27	0.373	1.1×10^{-3}	346	rh	supramarginal	PO
34	-36	-19	0.381	1.5×10^{-3}	301	rh	fusiform	pTF
44	-33	12	0.376	1.8×10^{-3}	272	rh	superiortemporal	PT
44	-17	18	0.379	2.9×10^{-3}	211	rh	supramarginal	pCO
36	-9	11	0.375	4.9×10^{-3}	149	rh	insula	aINS
18	12	-17	0.387	5.9×10^{-3}	131	rh	lateralorbitofrontal	FOC
55	-53	-20	0.376	6.9×10^{-3}	116	rh	inferiortemporal	pITg
34	-36	39	0.373	7.2×10^{-3}	112	rh	supramarginal	dSC
52	28	13	0.384	7.4×10^{-3}	110	rh	parstriangularis	dIFt
37	6	-38	0.383	7.5×10^{-3}	108	rh	inferiortemporal	TP
35	-20	16	0.370	9.8×10^{-3}	87	rh	insula	pINS
43	-50	-20	0.374	9.9×10^{-3}	86	rh	fusiform	pTF

Table 2.7: Peak classification accuracies for input-related beta estimates, classified on the identity of the coda consonant of each vocalization.

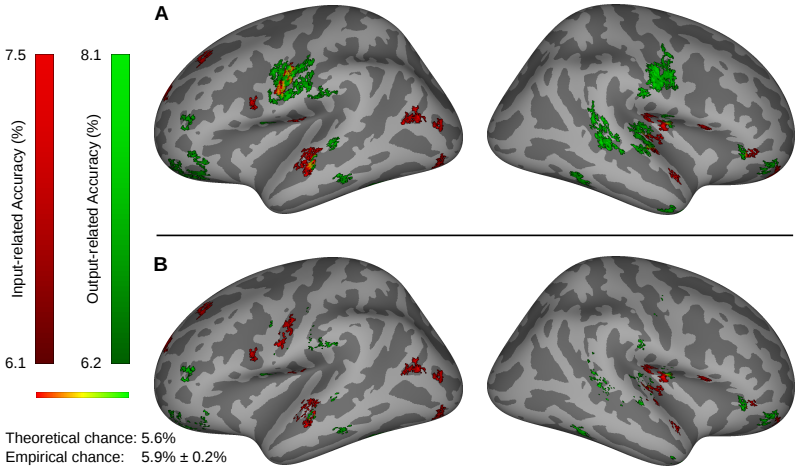


Figure 2.11: **Syllable accuracy.** (**A**) Regions with significant cross-validation accuracy for decoding of syllable identity based on input- (red) and output-related (green) hemodynamic response estimates. Regions of overlap are shown in yellow, with greater input accuracy appearing more red and greater output accuracy appearing more green. For output-related analysis, the identity of the spoken syllable was used as the classification target, when this differed from the presented stimulus. Results presented at $p < 0.05$ uncorrected, thresholded by cluster size ($p < 0.001$). Medial surfaces reveal no significant results. (**B**) An additional mask has been applied, hiding decoding results in vertices where significant decoding of constituent phonemes was found. Phoneme-level clusters were thresholded at $p < 0.01$, as shown in Figures 2.5, 2.9 and 2.10. Input- and output-related masks were constructed separately.

2.3.3.3 Syllables

Finally, we examined the information associated with prediction of whole-syllable identity. Figure 2.11A shows mean cross-validation accuracies for classifiers trained on syllable identity, with a cluster-level threshold of $p < 0.001$. Input-linked response estimates in left STs/STg, right anterior STg, posterior insula (pINS) and Heschl's gyrus provided significant predictive power for the syllable heard by the subject. We also found inferior frontal clusters in the left vPMC and vMC, as well as right frontal operculum and IFg pars orbitalis. For output-linked responses, predictive voxel patterns were localized in the left posterior STg, mid-STs, and right hemisphere Heschl's gyrus and posterior STs. Bilaterally, we found large, ventral sensorimotor

clusters and smaller anterior SMg clusters. No significant clusters were observed on the medial surface (not shown).

Figure 2.11B shows the same values as in Figure 2.11A, masked to reveal significant vertices that do not appear in onset, vowel, or coda results. Input-linked clusters remain largely unaltered in left vPMC and right pINS, while anterior superior temporal clusters bilaterally and left vMC were found to overlap with onset and coda clusters, to greater or lesser extents. All large output-linked syllable clusters were found to overlap with onset or coda clusters.

Locations of peak classification rates for all such clusters are listed in Tables 2.8-2.9. Theoretical chance accuracy for an 18 class classification problem was $1/18 = 0.0\bar{5}$, and peak group mean accuracies ranged from 0.07 to 0.083.

Input - Syllable

Coordinate (x, y, z)			Accuracy	p-value	Extent	Hemisphere	aparc	SLaparc17
-60	-13	-6	0.075	1.0×10^{-6}	507	lh	superiortemporal	adSTs
-14	56	25	0.074	1.1×10^{-5}	312	lh	superiorfrontal	FP
-43	-71	13	0.072	6.4×10^{-5}	194	lh	inferiorparietal	OC
-43	-12	30	0.072	8.2×10^{-5}	179	lh	precentral	vMC
-55	-5	22	0.074	8.2×10^{-5}	179	lh	precentral	vMC
-21	29	37	0.073	1.5×10^{-4}	146	lh	superiorfrontal	SFg
-38	-88	-16	0.072	1.9×10^{-4}	135	lh	lateraloccipital	OC
-41	-24	-23	0.072	3.0×10^{-4}	114	lh	fusiform	pTF
-37	-15	19	0.070	4.8×10^{-4}	96	lh	insula	pINS
-44	5	19	0.071	5.4×10^{-4}	91	lh	parsopercularis	vPMC
-40	-79	9	0.073	5.4×10^{-4}	91	lh	lateraloccipital	OC
-6	46	-24	0.071	5.9×10^{-4}	88	lh	medialorbitofrontal	FMC
44	-16	20	0.071	7.0×10^{-5}	188	rh	postcentral	pCO
35	-11	16	0.073	7.0×10^{-5}	188	rh	insula	pINS
23	50	-14	0.071	1.7×10^{-4}	142	rh	rostralmiddlefrontal	FP
54	8	-13	0.074	4.1×10^{-4}	102	rh	superiortemporal	TP
36	-19	5	0.071	5.9×10^{-4}	88	rh	insula	pINS
51	-10	1	0.070	6.2×10^{-4}	86	rh	superiortemporal	H
49	33	-13	0.071	7.7×10^{-4}	79	rh	parsorbitalis	FOC
42	12	6	0.071	7.7×10^{-4}	79	rh	parsopercularis	pFO
45	-19	5	0.071	7.7×10^{-4}	79	rh	transversetemporal	H

Table 2.8: Peak classification accuracies for output-related beta estimates, classified on the identity of each presented syllable.

Output - Syllable (vocalized)

Coordinate (x, y, z)			Accuracy	p-value	Extent	Hemisphere	aparc	SLaparc17
-48	-12	27	0.079	$< 1 \times 10^{-7}$	1160	lh	postcentral	vSC
-59	-9	29	0.075	7.0×10^{-6}	352	lh	postcentral	vSC
-35	31	-15	0.074	8.5×10^{-6}	334	lh	lateralorbitofrontal	FOC
-31	47	-14	0.072	9.1×10^{-5}	173	lh	lateralorbitofrontal	FP
-60	-30	18	0.072	1.1×10^{-4}	163	lh	supramarginal	PO
-58	-32	-1	0.075	1.8×10^{-4}	137	lh	superiortemporal	pdSTs
-22	36	-14	0.077	2.1×10^{-4}	131	lh	lateralorbitofrontal	FP
-52	-19	-9	0.073	3.9×10^{-4}	103	lh	superiortemporal	pdSTs
-9	56	-23	0.072	4.3×10^{-4}	100	lh	lateralorbitofrontal	FP
-39	41	7	0.074	4.3×10^{-4}	100	lh	rostralmiddlefrontal	FP
-65	-32	-16	0.073	5.3×10^{-4}	92	lh	middletemporal	pMTg
-43	-47	-15	0.078	6.0×10^{-4}	87	lh	fusiform	ITO
-45	5	5	0.074	6.2×10^{-4}	86	lh	parsopercularis	aCO
-60	5	24	0.074	8.6×10^{-4}	75	lh	precentral	vPMC
47	-12	29	0.083	$< 1 \times 10^{-7}$	1254	rh	postcentral	vSC
48	-22	9	0.078	1.2×10^{-7}	761	rh	transversetemporal	H
47	-37	4	0.077	1.1×10^{-6}	528	rh	bankssts	pdSTs
53	-40	11	0.077	9.9×10^{-5}	168	rh	bankssts	pdSTs
33	50	-13	0.075	1.2×10^{-4}	159	rh	parsorbitalis	FP
52	-50	-19	0.075	1.5×10^{-4}	147	rh	inferiortemporal	ITO
62	-38	9	0.076	1.6×10^{-4}	142	rh	bankssts	pSTg
61	-19	25	0.074	2.3×10^{-4}	127	rh	supramarginal	vSC
46	-2	-34	0.079	2.4×10^{-4}	125	rh	inferiortemporal	aITg
39	-6	17	0.071	5.1×10^{-4}	93	rh	postcentral	pCO
50	-7	47	0.076	6.2×10^{-4}	86	rh	precentral	midMC
47	-10	12	0.073	6.8×10^{-4}	83	rh	postcentral	pCO
47	30	-11	0.075	8.3×10^{-4}	76	rh	parsorbitalis	FOC
36	10	-39	0.078	8.9×10^{-4}	74	rh	temporalpole	TP

Table 2.9: Peak classification accuracies for output-related beta estimates, classified on the identity of each vocalized syllable.

2.4 Discussion

In this study, we applied cortical surface searchlight-based MVPA to estimated BOLD responses during a simple syllable repetition task. We sought to determine brain areas whose response patterns offered significant predictive power about the specific speech sounds heard and repeated, and to dissociate responses that were linked to the stimulus (i.e., input-linked) and those associated with the vocal response (i.e., output-linked). Below, we first discuss the overall pattern of results observed relative to our initial hypotheses, comparing and contrasting areas that contained speech sound related information related to the stimulus or to vocal output across phoneme and syllable units. This is followed by a general summary and synthesis of results and discussion of more technical aspects of the study.

2.4.1 Vowels

We anticipated that MVPA performed on vowel identity would provide the strongest and most interpretable results. This is because vowels are more acoustically salient and have longer durations than their consonant counterparts. In addition, the full set of vowels that a listener must have the ability to represent is smaller than the set of consonants. Although we used the same number of distinct vowels and consonants (three each) in our study, the more limited vowel set in American English might be expected to result in a higher signal-to-noise ratio in vowel-related activity patterns. Therefore, we consider the results in Figure 2.5 to be the most indicative of the neural representation of phonemes.

We hypothesized that left inferior frontal sulcus (IFs) would show sensitivity to phonemes during the output phase of the task, based on previous work resulting in the GODIVA neurocomputational model (Bohland et al., 2010), which proposed that this

region serves as a site for planning and selecting abstract sequences of phonemes (i.e., a *phonological output buffer*, Jacquemot et al., 2007). This region, stretching into dorsal inferior frontal gyrus (IFg) pars triangularis, has been shown to distinguish between-phoneme, but not within-phoneme, differences in voice-onset time (VOT) (Myers et al., 2009), as well as increase in activation in the planning of syllables of increasing phonological/phonetic complexity (Bohland and Guenther, 2006). Somewhat surprisingly, we discovered a large cluster of sensitivity to vowel identity in left IFs during the *input* phase of the task. We interpret this result as indicating that participants accessed the articulatory rehearsal and speech output systems immediately upon hearing and encoding the syllable. This strategy makes sense because participants were aware that they would need to remember and immediately speak the heard syllable upon receiving a GO signal.

We also therefore feel that it is important to revise our characterization of the "input-related" datasets to note that many responses may be more appropriately associated with working memory or the maintenance of a speech plan. Phonological working memory has previously been proposed to depend on the left inferior parietal lobule (IPL) (e.g., Paulesu et al., 1993; Celsis et al., 1999; Raizada and Poldrack, 2007; Buchsbaum and D'Esposito, 2008), and we discovered significant clusters in supramarginal gyrus (SMg) and angular gyrus (AG) for predicting vowels based on input-related responses. Raizada and Poldrack (2007) identified left SMg as performing categorical phonemic processing in a univariate analysis of responses to stimuli along the /ba/-/da/ continuum. A multivariate analysis of the same dataset, in which events were labeled with the subject's perceived category, however, failed to detect stimulus identity information in the same region (Lee et al., 2012). It is possible that the working memory demands of the current study, in comparison to Lee et al. (2012),

account for the appearance of vowel information in this analysis. An additional input-related cluster in the left rostral supplementary motor area (pre-SMA) coincides with a peak activation likelihood estimate (ALE) found in a meta-analysis contrasting 189 working memory tasks (a subset of which were phonological tasks) against control conditions (Rottschy et al., 2012). The left IFs and SMg clusters observed here also lie near peaks in the "working memory network" identified by that same contrast.

We also predicted that the left posterior superior temporal sulcus (STs) would provide phonological information during the input phase of the task, and instead found bilateral posterior STs clusters during the *output* phase. A strong residual auditory response to the stimulus is unlikely, as the "output-related" hemodynamic response peaks approximately 13 seconds after stimulus presentation. This information, we believe, is more likely to represent some form of auditory target (e.g., Guenther et al., 2006; Tourville et al., 2008) for the acoustically salient vowel portion of the spoken syllable. A lack of input-related information may indicate that any categorical vowel representations are absent from or not strongly activated in the STs in the early stage of this specific task, or such representations are insufficiently distinct at the spatial resolution of this study. In contrast, Formisano et al. (2008) found significant vowel classification in left STg and right STs in a passive listening task. In addition to the demand to generate a motor program, this study has additional methodological features that may explain the failure to replicate their result. First, a successful classification in this study requires that neighboring vowels in formant space be detected across several consonant contexts, while their stimuli were pure vowels with highly distinct formant profiles (/a/, /i/ and /u/). Second, we used a searchlight analysis in contrast to the recursive feature elimination method employed by Formisano and colleagues, which imposes a locality constraint on informative voxels.

It is notable that clusters that had predictive vowel information for the input- and output-linked response estimates are largely non-overlapping (see Figure 2.5). One common cluster, however, was observed near the right hemisphere homolog of area Spt, but not in the left hemisphere. Spt has been proposed to serve as a critical auditory-motor interface (Hickok et al., 2011), so its role in predicting vowel identity in both phases of the task supports this view. The present result is also consistent with a proposed role of the right planum temporale in converting auditory inputs into phonological representations (Deschamps and Tremblay, 2014). The rightward lateralization of this cluster (and the bilateral localization of STs clusters) might reflect a right-hemisphere preference for vowel processing (Britton et al., 2009).

A number of areas that predicted vowel identity during either the input or output portions of the task were not anticipated. Based on the GODIVA model (Bohland et al., 2010), which builds upon the frame-content theory of speech production (MacNeilage, 1998), the pre-SMA, which had input-linked vowel information, is suggested to encode abstract syllabic frames without regard for phonemic content. Jonas (1981) and Ziegler et al. (1997) each concluded, based primarily on clinical studies, that the medial premotor areas are unlikely to code for specific speech sounds, but rather may be involved more generally in the sequencing and initiation of speech. For the stimuli in our study, all CVC and sequencing demands were alike. However, a number of studies have identified pre-SMA as a site of response selection in sentence, word and nonspeech facial gestures (Alario et al., 2006; Tremblay and Gracco, 2006, 2009, 2010; Tremblay and Small, 2011), and Carreiras et al. (2006, 2009) previously found differential pre-SMA response to high- and low-frequency words. Additionally, Adank (2012) has shown, based on ALE meta-analyses, that pre-SMA activity increases with the difficulty of speech comprehension tasks. If the difficulty of correctly perceiving,

encoding, or selecting each syllable in phonological working memory differed across vowel classes, then this could be driving the above-chance accuracy in pre-SMA. Further study is warranted to more directly address these issues.

A surprisingly prominent right ventral angular gyrus (AG) cluster was also observed in the input-linked analysis. Left AG has classically been associated with grapheme-phoneme correspondences (Hynd and Hynd, 1984), and recent studies suggest that right AG may have a role in orthographic comprehension (Mei et al., 2014) and phonological / orthographic mappings (Bonte et al., 2014). We speculate that if the right AG provides information about vowel identity, it may be supported by orthographic associations with the vowels in the stimuli, although participants did not report using visual strategies in this task. Based on a thorough review of neuroimaging literature, Price (2012) assigned bilateral angular gyrus a role in semantic processing, but noted that descriptions of this region's role in comprehension are still necessarily vague and insufficient. We note here that the semantic content available in syllables used was not uniform across vowels (6 of 6 /ɪ/ syllables, 4 of 6 /ɛ/ and 1 of 6 /ʌ/ syllables could be perceived as words or names). It is conceivable that semantically-related processes were selectively evoked for these stimuli, resulting in a signal whose discriminability artifactually correlated with vowel identity.

The large locus for output-related vowel information observed in left posterior cingulate (pCG) was unexpected since this region is not traditionally associated with speech or language. Functional connectivity studies have associated both pCG and bilateral STg with the default mode network (Martuzzi et al., 2010), which may provide a functional pathway for speech sound related information. It should be noted, however, that significant vowel-related information was detected in bilateral STs, but not STg, for output-linked responses. Myers (2007) showed a posterior cingulate

response to category-atypical VOT in a phonetic discrimination task, while Menenti et al. (2012) found a repetition suppression effect for words, as opposed to semantic or syntactic structure. Finally, working memory tasks using visually presented letters have also yielded effects in the posterior cingulate, including decreased activation with load (Tomasi et al., 2007) and activation while selecting an action among remembered targets (Hester et al., 2007). Taken together, these findings suggest a somewhat general role for posterior cingulate / precuneus in tasks requiring linguistic processing and/or working memory; the specific role of this region in our task is unclear and warrants for further investigation.

2.4.1.1 Acoustic analysis: phonetic or phonological representations?

Interpretation of MVPA studies remains an open problem, particularly because results may diverge from traditional univariate analyses, and classification success may be driven by many factors (Etzel et al., 2013; Davis and Poldrack, 2013; Todd et al., 2013; Davis et al., 2014). Our efforts to localize information corresponding to phonemes provides a window into this problem. In the case of vowels, there is a strong relationship between formant frequencies — acoustic properties of the speech sound that can vary continuously — and the identity of the vowel. Classification rates provide no obvious way to distinguish a classification based upon a neural representation of formants (an acoustic-phonetic representation) or a more abstract (phonological) representation of the vowel. We therefore tested the effect on vowel classification accuracy of removing univariate, voxel-wise BOLD responses that covary with formant frequencies specific to the sound heard or produced (see Figure 2.6). The idea was that, if an area maintains an abstract, categorical code for speech sounds, its classification accuracy should not be reduced by removing formant-related variance. If an

area's response is instead sensitive to formant frequencies of the sound, then its basis for classifying vowel sounds may be diminished by removing such variance.

While no results reached group-level statistical significance, our findings suggest there may be a small decrease in classification performance when formant-related information is removed from each voxel. In particular, the pre-SMA, cingulate and posterior portion of the IFs clusters showed the strongest effects in our input-related analysis, while left posterior cingulate and right posterior STs showed the strongest effects in our output-related analysis. Areas that showed less reduction included left anterior IFs, insula, SMg, right posterior STg and IFg in input-related analysis, as well as left inferior temporal sulcus, central operculum, and right posterior STg/area Spt. The relative lack of reduction in accuracy would indicate representations that may be more abstract or phonological, as opposed to areas tracking phonetic/acoustic/articulatory details. Further work using tools such as representational similarity analysis (Kriegeskorte et al., 2008) may help to refine our understanding of the nature of these representations (see, for example, its use by Evans and Davis, 2015, to test between different levels of representation in speech perception).

It may be also be that direct use of formant profiles is inappropriate, given that it does not account for talker normalization (see, *e.g.*, Johnson, 1990). Several subjects spontaneously reported identifying a limited number of talkers during debriefing, which may provide sufficient context for talker normalization to take place (Ladefoged and Broadbent, 1957) and talker pitch has been shown to assist vowel classification even in the absence of context (Halberstam and Raphael, 2004). An intermediate auditory representation of vowels that is talker-independent but nonetheless reflects acoustic variation cannot be ruled out by this analysis.

2.4.2 Consonants

Consonants are acoustically and motorically dissimilar to vowels, and so provide an opportunity to observe the classification of phonemes whose acoustic and phonetic representations can be expected to substantially differ from those of vowels. Accordingly, the accuracy maps produced by classifying trials by the consonants in each stimulus (Figures 2.9 and 2.10) differ greatly from those produced by classifying by vowels. It should be noted that, unlike other studies that used stimuli with the same vowel but a consonant contrast (e.g., Myers et al., 2009; Lee et al., 2012), classifiers here had to generalize across vowel contexts to be successful (see also Zhang et al., 2016).

The accuracy maps for onset and coda consonants are broadly similar in both input- and output-related responses. Most prominently, information predictive of both onset and coda consonants was present bilaterally in large segments of the ventral somatosensory (vSC) and motor cortices (vMC) in output-related responses. The distinct articulations required for each consonant are the most likely source of information detected here, consistent with descriptions of an articulatory somatotopic map in ventral sensorimotor cortex (Bouchard et al., 2013; Conant et al., 2014). Notably, we did not observe similar output-related results for vowels. Articulatory (and incoming somatosensory) information corresponding to different vowels may be more difficult to differentiate because vowels differ more subtly in precise tongue placement (cf. different places of articulation for consonants). Such subtle variations in articulations are likely to result in less distinguishable signals for vowels in primary motor and somatosensory cortices, particularly with the current resolution of fMRI data.

Input-related analyses revealed much smaller clusters of consonant-level informa-

tion in bilateral motor and premotor cortices than the corresponding output-related analyses. In the basic contrasts of both input and output event estimates vs. baseline (Figure 2.4), each showed strong, bilateral responses in vMC, though, though the output signal was stronger. Thus, it appears the motor / premotor cortex was engaged in the input-portion of the task, but that activation patterns were overall less predictive of consonant identity than during output, when explicit motor programs were enacted. The overlap between input and output related consonant predictors along the sensorimotor cortex is relatively small but notable. Though precise localization differed slightly for onsets and codas, both input- and output-related responses in small portions of the ventral precentral gyrus and/or sulcus, near the border of motor and premotor cortices, predicted consonant identity. These clusters appear to be consistent with the suggestion of an abstract, non-acoustic coding of consonant sounds in the ventral somatomotor regions (Evans and Davis, 2015).

Consonant information was also found broadly in the temporal cortex, bilaterally, where vowel information was more spatially confined, particularly to the superior temporal sulcus (Figure 2.5). Predictive clusters based on input and output event estimates overlapped mainly in the superior temporal gyrus, with left anterior STg and right middle to posterior STg clusters for onsets, and bilateral mid STg clusters for coda consonants. The wide profile of predictive responses across the superior temporal areas does not suggest a single area in the STg that encodes consonants in an abstract manner, consistent with recent work suggesting distributed, feature-based representations in left STg (Mesgarani et al., 2014; Leonard and Chang, 2014).

Although the onset and coda accuracy maps are broadly similar, a number of differences suggest that information relevant to consonant identity may not be represented independent of position within a syllable. However, it is not possible to

distinguish between the effect of serial position on representation and the effects of primacy/recency on BOLD response in this analysis. Additional experiments will be required to elucidate important issues related to serial order within syllables, words, and sentences.

We also note that, while we attempted to minimize head movements and included regressors to account for motion in event estimates, it seems possible that some of the large, diffuse output-related results for consonants (particularly in the inferior temporal and ventral frontal pole areas) might be artifactually related to small movements. Since subjects were overtly articulating, it is possible that some consonants (especially final consonants) led to small motion-related artifacts that drove the discrimination of consonants.

2.4.3 Syllables

In addition to phoneme-level analyses, this study design permits us to discover regions whose responses correlate with the identity of the whole syllable (see Figure 2.11). Some findings, most obviously ventral Rolandic cortex in output-related responses, recapitulate the findings of the consonant analyses. However, a number of clusters are distinct from any of those found in the phoneme-level analyses (Figure 2.11B), and may indicate a representation that cannot be decomposed into smaller units.

We hypothesized that the left ventral premotor cortex would encode a syllabic representation at the time of speech output based on its proposed role as a *Speech Sound Map* (Guenther et al., 2006). A cluster in this region was shown to significantly predict the syllable heard, but clusters that predicted the syllable based on output-related events were more posterior in and around the central sulcus (and overlapped with consonant predictors). As was the case with left IFs for vowels, we suspect

that the fact that this cluster was *input-related* is likely due to the specific task requirements, which may have allowed subjects to activate a syllabic program for speech output immediately upon hearing the stimulus. It is important to note that left vPMC did not show phoneme-level predictive information for consonants or vowels in the present study; furthermore this region was the *only* region suggested to encode entire planned syllables in a repetition suppression study conducted by Peeva et al. (2010).

Additional input-related clusters that predicted syllable identity but did not predict vowels or consonants were found in the left motor cortex along the ventral central sulcus. Recent work by Evans and Davis (2015) found left vMC to be sensitive to syllable identity, phonemic content and phoneme ordering (CV vs VC) in a speech perception task, but not to acoustic variations, such as speaker identity and acoustic degradation. We can corroborate, but not fully replicate, this finding with the vMC cluster found here; phoneme content and their order naturally support syllable classification, and the inter- and intra-speaker variations in our stimuli require some degree of abstraction from acoustic detail, but we did not present acoustically degraded stimuli.

Although additional small, distributed clusters remain for both input- and output-related prediction of syllables (Figure 2.11B), due to their relative size and proximity to phoneme-level clusters, we reserve the possibility that significant syllable classifications may be supported by phoneme-level information.

It is worth noting the absence of syllable-level information in the SMA/pre-SMA (or on the medial surface altogether), in the context of frame-content theory (MacNeilage, 1998) and the specific proposals of the GODIVA model (Bohland et al., 2010). In GODIVA, for CVC syllables such as those in this task, the frame rep-

resentation in pre-SMA would consist of "instructions" to select a consonant, then a vowel, and then a consonant, while the phonemes themselves are encoded in left IFs. Classification for syllable identity detects response patterns that are sufficiently different between classes. If pre-SMA encodes abstract frames without content, then we would not expect predictive information to be present. Thus, this lack of syllable-level results is consistent with GODIVA predictions. Additionally, these results are consistent with findings that left SMA has reduced activity when producing learned words over phonotactically illegal words (Segawa et al., 2015), under the assumption that all biphones in this study are sufficiently practiced to avoid differential demands on abstract sequencing.

2.4.4 Summary of main hypotheses

One of our major hypotheses was that output-related responses in the left inferior frontal sulcus would provide significant information about phonemic identity, based on our suggestion that this region serves as a phonological output buffer. We found that left IFs patterns predicted vowel (but not onset or coda consonant) identity during the *input* portion of the syllable repetition task, suggesting perhaps the automatic recruitment of working memory and output-planning representations upon hearing the stimulus. This area did not predict the whole syllable, and thus our results support a role in phonological content at a sub-syllabic level. The lack of predictive information for consonants was unexpected, but may reflect their reduced salience, increased competition due to a larger consonant alphabet compared to vowels, or a finer-grained representation that could not be recovered at the present fMRI resolution.

We also predicted that activity in the left ventral premotor cortex would pre-

dict whole syllable identity, based on the idea that this region stores sensory-motor programs for well-learned syllables or words. Indeed, a cluster in left vPMC contained significant information about whole syllable identity but not about individual phonemes. This cluster, like IFs however, appeared for input-linked rather than output-linked responses.

On the other hand, bilateral superior temporal sulcus patterns were found to be predictive for output-related classification of vowels. We had anticipated, based on previous studies and the dual pathways model, that this region would encode phonemic content related to the auditory stimulus, but no such effect was observed. Instead, one interpretation of our results is that the strong STs clusters observed for vowel prediction reflect the activation of speech sound targets for production.

Finally, based on the dual pathways model and previous related studies, we hypothesized that area Spt and/or other portions of the planum temporale, would predict speech sounds during both input and output components of the task. This was, to an extent, confirmed, by overlapping clusters that predicted vowel identity at the two different task periods in the right hemisphere posterior Sylvian fissure.

Overall the pattern of results supported the main areas we hypothesized to be involved in representation of speech content, though our assumptions about separation of input- and output-related representations may need to be revised. A follow-up study is in progress designed to better enable separation of these representations.

2.4.5 Methodological considerations

2.4.5.1 Delayed syllable repetition task

The design of the delayed repetition task used here afforded several key features for this study. By using multiple recordings of multiple speakers, we increase ecological

validity over synthesized stimuli, which are often used in categorization and discrimination studies. The naturally-occurring variation in the presented speech sounds required participants to map many acoustic stimuli onto their learned (and presumably categorical / abstract) internal memory representations, and ultimately their own motor programs for a given syllable. Further, a delayed repetition task is well-suited for examining both input- and output-related responses to the same stimuli, requiring both perception and production while allowing some temporal separation between the task components. Most previous fMRI tasks have focused on either perceptual or production processes, and our study represents, to our knowledge, the first effort to map predictive information related to speech sounds across both types of processes in the same experiment. The task shares some similarities with nonword repetition (NWR) tasks that have been used in clinical neuropsychology to assess phonological working memory in developing children and patients with aphasia (e.g., Gathercole, 1995; Jefferies et al., 2006), and to a lesser extent in neuroimaging studies with healthy subjects (McGettigan et al., 2011). In contrast to NWR, which typically uses multi-syllabic nonwords, our task used single syllable stimuli, which allowed for clear identification of sounds of interest for classification on each trial.

2.4.5.2 Sparse design and hemodynamic modeling

The experimental protocol was designed to allow as many trials as possible in order to provide sufficient training data for classifiers trained on multiple linguistic classes of interest. For this reason, we used a fast, event-related design with sparse volume acquisition (Perrachione and Ghosh, 2013), which allowed the subject to both hear and produce the stimulus during periods of relative quiet (Eden et al., 1999; Edmister et al., 1999; Hall et al., 1999). The decision to place both volume acquisitions during

the maintenance phase of each trial, such that they were timed to capture the peak response to the stimulus onset and the GO signal (from the previous trial), resulted in a pattern of unevenly timed sparse acquisitions. As a result, the single-shot EPI volumes were taken at different points on the T1 relaxation curve, and thus there were distinct slice-timing effects between A scans and B scans.

In this sparse, aperiodic paradigm, each volume acquired is inherently linked to a different event of interest, in contrast to clustered volume acquisition (CVA) designs (e.g., Zaehle et al., 2007; Schmidt et al., 2008), in which multiple volumes are acquired (sequentially) for each event of interest to reduce noise. As a consequence, HRF modeling was necessary to separate the overlapping signals. The correction detailed in the methods (Section 2.2.6.1) results in A scans with the same approximate slice-timing effects as B scans, yielding similar mean amplitudes across all trials. This permitted us to use both A and B scan data together in general linear models to estimate response amplitudes for both input- and output-related events.

2.4.5.3 Input- and output-related response estimates

The piecewise-by-condition GLMs used to estimate individual event responses (following the modeling described above) drew inspiration from the method presented by Mumford et al. (2012) to reduce collinearity in design matrices for rapid, event-related designs. In their method, an n-column design matrix was converted into n 2-column design matrices, where the second column is the sum of all other columns of the original matrix. One way to interpret this method is that all events are assumed to have an equal mean response, and variation in individual responses is treated as error; each GLM then estimates the variation attributable to the given individual event. Here we required response estimates for individual events, which were used to

train and test classifiers. Our approach was to constrain all but one condition at a time (i.e., collapsing all events of each other type into single columns in the design matrix), and let the parameter estimates for all individual events in the condition of interest vary (see Figure 2.3).

The estimates of individual event responses were averaged and contrasted with baseline response estimates to provide basic univariate speech contrasts (Figure 2.4). These contrasts reveal increased hemodynamic responses primarily localized to superior temporal and ventral sensorimotor cortices, with a relative shift from auditory to somatosensory cortex in the output-linked contrast (Figure 2.4B). In addition to expected activations in auditory cortex when hearing a syllable, listening to speech also activates motor and premotor areas (Wilson et al., 2004; Pulvermüller et al., 2006; Meister et al., 2007; D'Ausilio et al., 2009), as was observed here. It is noteworthy, though, that despite the large cluster of input-related activation in this basic contrast, only small portions of this cluster significantly predicted the identity of phonemes or syllables, highlighting the differences between analytic approaches. Furthermore, as is observed here, output-linked events could be expected to show an overall stronger motor response and a moderate, self-induced suppression of the auditory responses to hearing one's own voice (Numminen et al., 1999; Houde et al., 2002; Flinker et al., 2010). These overall task effect maps thus provide reason for confidence in the individual event estimates, despite the fast temporal structure used in our protocol.

2.4.5.4 Multi-voxel response patterns

In turn, the individual event estimates form the basis for the multi-voxel pattern analyses that formed the core of this study. The MVPA approach affords certain advantages over univariate approaches. Because we were interested in responses that

predict multiple features embedded within syllables, we expected that the noise introduced by the features of non-interest (*e.g.*, all combinations of vowel and coda segments) would likely dominate the single-voxel responses to the feature of interest (*e.g.*, /t/ in the onset position) in the analyses presented here. Using multiple voxels affords greater degrees of freedom for revealing associated response patterns. Furthermore, we would not expect individually informative voxels to align precisely across subjects; searchlight techniques introduce an implicit smoothing in the results of the analysis, facilitating greater inter-subject alignment.

In contrast to using individual event estimates, many MVPA studies (e.g., Chen et al., 2011; Oosterhof et al., 2011) use a single response estimate per-condition, per-run as the basis of classification. In so doing, these studies treat the variation from each trial as noise, and classification reflects the consistency of the mean response *across runs* to classes of events. Per-event classification instead treats the variation from each trial as signal, and classification reflects the generalization across *distributions* of responses to classes of events. This latter strategy produces cross-validation accuracies that tend to center around theoretical chance (*i.e.* 0.33 for 3 classes or 0.056 for 18), which are lower than commonly reported (for many stimulus classes) for the former strategy. To avoid assumptions about "chance" accuracy, and to account for spatial variations in signal, we used computationally intensive non-parametric tests to determine whether accuracies were significantly above empirical chance. It should be noted that the goal of our approach is not to "read minds," but to detect predictive information about speech sounds. Therefore, while the neural representations of speech sounds are almost certainly at a finer scale than fMRI currently affords, limiting possible classifier accuracy, statistically significant (but otherwise unimpressive) cross-validation accuracies provide such a measure of class-specific information.

2.4.6 Limitations and future directions

It is important to note that although we focused our analyses on the cerebral cortex, the cerebellum and other subcortical structures are likely to also play roles in the representation of speech sounds and warrant further investigation. MVPA may be applied volumetrically to probe these areas, though it is critical to constrain searchlights anatomically (as was done here for the cortex) to ensure interpretability of the roles of individual areas / nuclei.

In this study, the stimuli were discrete CVC syllables, with variation derived from multiple recordings and speakers, which is in contrast with a number of studies that instead used finely varying synthetic stimuli, which may more directly allow addressing questions related to categorical processing of speech. However, such stimuli could easily be employed within a repetition paradigm as used here, which might allow refined testing of phonological vs. phonetic levels of representation. CVC syllables were used to allow comparisons of classifiers trained on the same consonants in different serial positions. We expect using the more common CV frame would produce similar results, but varying syllable types might permit additional testing related to the representation of abstract frames.

A substantial challenge in this and related work is the problem of disentangling input- and output-linked responses during simple repetition, when the stimulus and vocal production have the same class. This leads to difficulty, for instance, in interpreting if an input-linked response in frontal cortex is important for perception or if it simply reflects preparation of the sound for production. In the following study, we address this confound using a design that sometimes breaks the symmetry between input and output (i.e., the subject produces a different syllable or word than the one heard; see also Cogan et al., 2014).

Chapter 3

Decoupling input and output representations of words and nonwords in a speech repetition task

3.1 Introduction

In the previous experiment, subjects listened to and reproduced a single syllable during each trial, introducing ambiguity as to whether regions whose response profiles correlate with the phonetic content are responding to the auditory stimulus or the vocal target. Because temporal and frontal speech areas are engaged in both speech perception and production, and representations in memory may evolve over time, neither region nor the temporally defined *input* or *output* event is sufficient to resolve this ambiguity.

In particular, activity in the left posterior inferior frontal sulcus (pIFs) correlated with the perceived vowel at input, which may be either a necessary component of speech perception or that of a rapidly generated motor plan. Similarly, activity in the bilateral posterior superior temporal sulcus (pSTs) correlated with the vocalized vowel at output, which we interpreted as an auditory target, but we cannot logically eliminate auditory memory as an explanation. Mid-to-posterior STs (and STg) has been shown to be responsive to both perceived and produced speech Paus (1996); Hickok and Poeppel (2000); Okada and Hickok (2006b); Hickok and Poeppel (2007) over a variety of non-speech controls, so we sought to clarify this activation.

In this chapter, I present a study designed to dissociate the phonological content of the auditory stimulus and vocal target in a subset of trials. In this study, subjects were visually presented with two (non)word syllables simultaneously, then aurally presented with one of the syllables. A visual cue then informed subjects either to repeat the heard syllable (repeat trials) or produce the unheard, visually presented syllable (change trials). Following a delay, a further visual cue instructed subjects to produce the planned syllable. As in the previous task, described in Chapter 2, in each trial subjects perceived an aurally presented syllable and spoke a planned syllable after a delay. However, in half of all trials (pseudorandomly ordered), the planned speech target was cued visually.

On any given trial, by presenting all linguistic materials (both visual and auditory) before informing subjects of the trial type (repeat or change), we required subjects to prepare to repeat either syllable. This guaranteed that the same auditory stimulus would be processed the same under both conditions. Thus the task had three phases of interest in which to detect the correlates of the auditory stimulus syllable or the vocal target syllable: auditory perception, motor plan preparation (following the cue informing the trial type), and the overt speech production. In change trials, the mismatched phonological content of the auditory stimulus syllable and the vocal target syllable served as a tag to indicate whether information detected in a brain area derives from auditory processing or vocal preparation.

Half of all stimuli in this task were words, and half nonwords, balanced across repeat and change trials for a 2×2 factorial design. In addition to dissociating perceptual and vocal plan content, we were able to probe differences in processing of words and nonwords, which have been shown to be processed differentially in repetition tasks (Binder et al., 2005; Raettig and Kotz, 2008; Saur et al., 2008). Where

word repetition may rely on lexical and semantic processing pathways (see Binder et al., 2009), nonwords must be processed phonologically / phonetically (Hickok and Poeppel, 2004; Jacquemot and Scott, 2006).

Finally, this design permitted us to perform principled contrasts of searchlight analyses, building on the methods presented in Chapter 2. For example, by contrasting classification accuracy of the auditory stimulus and the vocal target, we were able to distinguish regions involved in perception and generating a speech motor plan; by contrasting accuracy of the same analysis at two time points, we were able to identify regions whose representations became more or less prominent over time. Thus, we were able to reconstruct some of the dynamic processes engaged by the task.

3.2 Materials and methods

3.2.1 Participants

21 right-handed native American English speakers participated in this study. Two participants were excluded for excessive motion, and one chose to discontinue participation in the study before a minimum number of runs were completed. Of the remaining 18 subjects, 10 were female (ages 18-31; $\mu = 21.9$, $\sigma = 4.2$) and 8 were male (ages 21-36; $\mu = 27.6$, $\sigma = 5.6$). No participants reported any history of speech, language, or hearing disorders. All participants gave informed consent under the protocol approved by the Institutional Review Board of Boston University. Two participants chose to discontinue participation after 4 functional runs were completed. In all other participants, 8 functional runs were completed.

3.2.2 Task design

The task was designed to require subjects to produce CVC syllables with and without a direct acoustic model. To achieve this in an event-related design, each trial entailed the visual presentation of two syllables, and the auditory presentation of one of those syllables, followed by a cue indicating whether the subject was to prepare to repeat the auditory stimulus or produce the unheard visual stimulus. Thus, throughout the stimulus phase of each trial, subjects must be prepared to produce either syllable, and cannot pre-commit to different strategies.

A trial began with the visual presentation of two words or two nonwords, containing two different vowels. After 2s, the (non)words were replaced with a white fixation cross. At 2.25s, one of the two (non)words was presented aurally. At 3.25s, either a green or yellow rectangle cue was presented around the fixation cross, which served as the task instruction cue. At 9s, the fixation cross turned orange (the GO cue),

cueing subjects to produce a syllable. Each trial began 13.5s after the start of the previous trial.

On a *repeat* trial, subjects were presented with a green rectangle cue, indicating they were to produce the syllable they heard. On a *change* trial, subjects were presented with a yellow rectangle cue, indicating they were to produce the syllable they read but did not hear. Control trials were identical to task trials, except the aurally-presented sound was a speech-shaped noise stimulus (see Section 3.2.3).

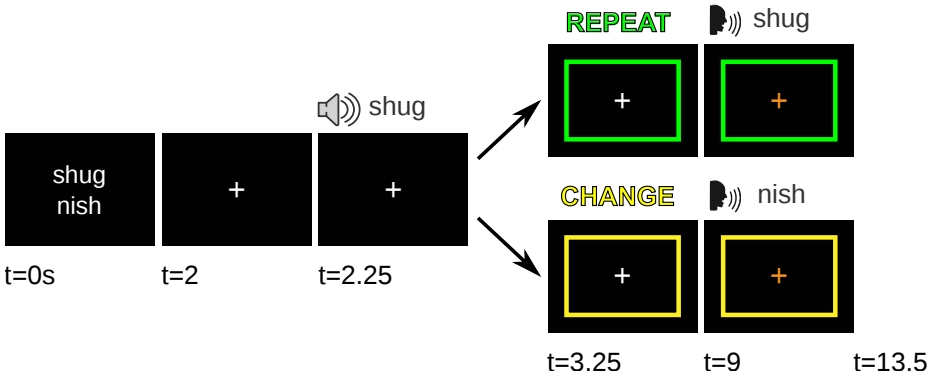


Figure 3.1: **Stimulus paradigm.** A schema of the progression of a task trial. At $t = 0$s, two words or two nonwords are presented visually for 2s, then replaced by a white fixation cross. At $t = 2.25$s, one of the two syllables is presented auditorially. At $t = 3.25$s, a rectangular repeat/change (green/yellow) cue is presented, indicating whether the subject is to repeat the auditory stimulus or change to the unheard, visually presented stimulus. After a delay, at $t = 9$s, the fixation cross changes color to orange, prompting subjects to speak the planned syllable. The trial ends at $t = 13.5$s.

Figure 3.1 depicts the course of two example trials with the same visual and auditory stimuli. The trials diverge at the presentation of the repeat/change cue (rectangle).

Each run consisted of 36 task trials – each of 18 stimulus syllables, balanced across *repeat* and *change* trials – and 6 control trials, half with a green task instruction cue and half with a yellow. As half of stimuli were words, word and nonword trials were

counter-balanced with repeat and change trials. A run lasted a total of 9 min. 34s.

3.2.3 Materials

Stimuli were consonant-vowel-consonant (CVC) syllables, including 9 words and 9 nonwords, 3 of each containing each of the vowels: /æ/, /ɪ/, and /ʌ/. The words fall into three semantic categories – animal, body part, vehicle – balanced across vowels, although this factor is not analyzed here. See Table 3.1 for a full list of syllables.

Vowel	**Animals**	**Body parts**	**Vehicles**	**Nonwords**
/æ/	yak	back	cab	yag fath tham
/ɪ/	pig	shin	ship	nish yig thip
/ʌ/	bug	thumb	bus	shug fup nus

Table 3.1: The vowels /æ/, /ɪ/, and /ʌ/ were selected to construct 18 word and nonword CVC stimuli. The words were chosen to fall into three semantic categories: animal, body part, and vehicle. Syllables are spelled as they were displayed to participants. For nonwords, subjects were instructed to use phonetic pronunciations, and that "th" was to be pronounced /θ/ (see Appendix A.2).

All stimuli were selected from the corpus of 5,765 CVC syllables compiled by Storkel (2013). Our overall goal was to choose syllables with phonotactic transition probabilities and phonological neighborhood densities (*e.g.*, Vitevitch and Luce, 1999) that were close to the average across CVC syllables.

Specifically, we looked at the Z-scores for the positional segment sum, the biphone sum, and the number of phonological neighbors, and chose syllables containing the vowels /æ/, /ɪ/, and /ʌ/ with standardized values close to zero. Selection was constrained by several elements of the experimental design, including a desire to vary the

neighboring consonants across syllables, and to choose three words (one using each vowel) from each of three semantic categories (animals, body parts, and vehicles). Thus, some variability in phonotactic probabilities and densities was present across stimuli, but we attempted to manually minimize this variation.

A table showing the z-scores for these phonotactic variables (relative to the Storkel (2013) corpus) for all stimuli used is shown in Table 3.2.

Syllable	**Segment sum**	**Biphone sum**	**Neighborhood density**
yak	-0.09	-0.13	-0.82
bug	-0.81	-0.31	0.76
pig	1.18	0.38	-0.16
back	0.87	0.96	1.56
thumb	-1.09	-0.19	-0.69
shin	1.26	0.82	-0.03
cab	1.18	1.66	0.23
bus	0.54	0.18	-0.03
ship	-0.04	-0.09	-0.16
yag	-0.08	0.01	0.60
fath	0.57	0.35	0.12
tham	0.62	0.59	0.28
shug	-0.97	-0.37	0.92
fup	0.32	-0.26	-0.37
nus	0.76	0.62	0.92
yig	0.30	0.18	-0.37
nish	0.44	-0.02	0.12
thip	0.72	0.82	0.76

Table 3.2: **Phonotactic variables (z-scored) for selected syllables** Within the design constraints of the experiment, CVC syllables were selected to have close-to-average positional segment sum (`HML_S_Sum`), biphone sum (`HML_B_Sum`) and neighborhood density (`HML_N_Nbors`), relative to the Storkel (2013) corpus. Shown are the z-scored values for each of the selected syllables. Z-scores were calculated for word and nonword datasets, separately.

Two male and two female native English speakers recorded the stimuli, and one to five recordings of each syllable per speaker were used to allow for additional acoustic variation in the auditory tokens that subjects heard. Speech-shaped control stim-

uli were generated by amplitude modulating pink noise – noise with a $1/f$ power spectrum – by the Hilbert envelopes of the original stimuli.

3.2.4 MR-data acquisition

All measurements were performed using a 3T Philips Achieva MRI scanner with a 32 channel head coil at the Boston University Center for Biomedical Imaging. T1-weighted anatomical images were acquired for anatomical reference and coregistration with functional data ($0.98\times0.98\times1.2$ mm^3 voxels, 150 sagittal slices, 256×254 matrix, repetition time = 6.8 ms, echo time = 3.1 ms, P reduction (AP) SENSE factor = 1.5, S reduction (RL) SENSE factor = 2). Functional volumes consisted of 43 echo-planar transverse slices (3mm thickness), acquired in ascending order, with no gap ($3.03\times3.14\times3$ mm voxels, 76×73 matrix, acquisition time = 2250ms, repetition time = 3375ms, echo time = 35ms, flip angle = 90°, P reduction SENSE factor = 3). Functional volumes were acquired in a sparse acquisition paradigm (Figure 3.2). Two additional volumes were acquired at the end of each run in order to capture residual hemodynamic activity in response to previous experimental events.

3.2.5 Behavioral assessment

All subjects' vocalizations were verified against the presented stimuli. 5 errors were possible: (1) distractor errors, in which subjects repeated the wrong stimulus (see Section 3.2.5.1); (2) speech errors, in which subjects produced a syllable that was not presented; (3) timing errors, in which subjects produced speech partially or wholly overlapping with a volume acquisition (see Section 3.2.5.2); (4) speech failure, in which subjects incorrectly remained silent; (5) control failure, in which subjects produced a syllable on a control trial.

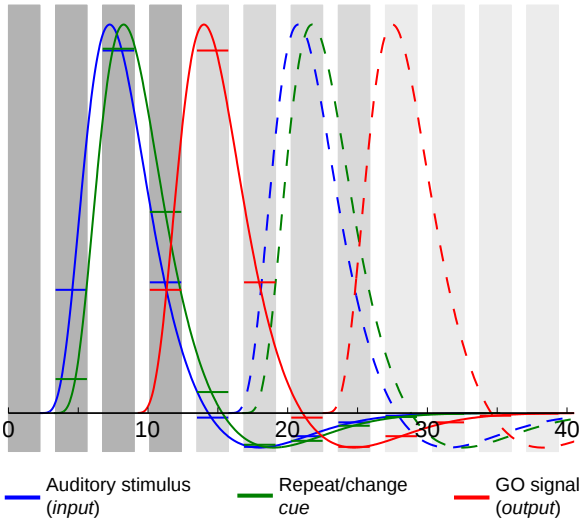


Figure 3.2: **Schema of sparse acquisition paradigm.** The solid blue, green, and red curves indicate the expected hemodynamic response functions (HRFs) associated with the presentation of the stimulus at $t = 2.25$s, the repeat/change cue at $t = 3.25$s, and the GO signal at $t = 9$s, respectively. Scans, represented by gray boxes, are acquired with TA = 2250ms, TR = 3375ms, and each 13.5s trial is indicated by 4 scans of the same shade. The dashed lines indicate the expected responses to the events of a second trial, starting at $t = 13.5$s. Horizontal lines indicate the mean values of theoretical HRFs across the duration of a scan, approximating the regressors representing the three events in a general linear model.

Speech errors, speech failures, and control failures were excluded from multivariate analyses. Trials with apparent speech failures were further examined to ensure that subjects were not speaking at the wrong time, though timing errors cannot be ruled out if vocalizations were entirely masked by scanner noise.

3.2.5.1 Recoding

If subjects produced the distractor syllable instead of the target, the the vocalized syllable and trial type were recoded to match subject behavior. For example, if a subject read *nish* and *nus*, heard *nus*, and repeated *nus*, the trial was coded as a *repeat* trial with a vocalization class of *nus*.

Recoding was applied to MVPA analyses only, and not considered for univariate

contrasts.

3.2.5.2 Timing errors

If a vocalization partially or wholly overlapped with a volume acquisition, as determined by auditory inspection of recordings and/or visual inspection of acoustic waveforms, that volume was marked as a motion outlier. See Section 3.2.6.2 for handling of motion outliers.

If the discernible phonemes matched the target or distractor, then the vocalization was coded as successful or recoded as a distractor error (see Section 3.2.5.1).

3.2.5.3 Class labels

We considered two class labelings for trials: *aud* - the vowel class of the auditorially presented stimulus; *voc* - the vowel class of the actually spoken syllable. In the latter case, a failure to speak was treated as an error, and not as a control trial.

3.2.6 Preprocessing

We reconstructed cortical surfaces from the T1-weighted structural images with FreeSurfer (Dale et al., 1999; Fischl, 2012) v5.3.0. The preprocessing pipeline for functional data, schematized in Figure 3.3, used the FreeSurfer Functional Analysis Stream (FsFast) and the FMRIB Software Library (FSL) (Smith et al., 2004; Jenkinson et al., 2012).

All functional volumes from all runs were realigned to the first volume of the first run in FsFast. Further preprocessing of functional data splits into two streams: univariate analyses and cortical multivariate analyses. For univariate analyses, FSFAST was used to register functional volumes to the `fsaverage` template. For multivariate

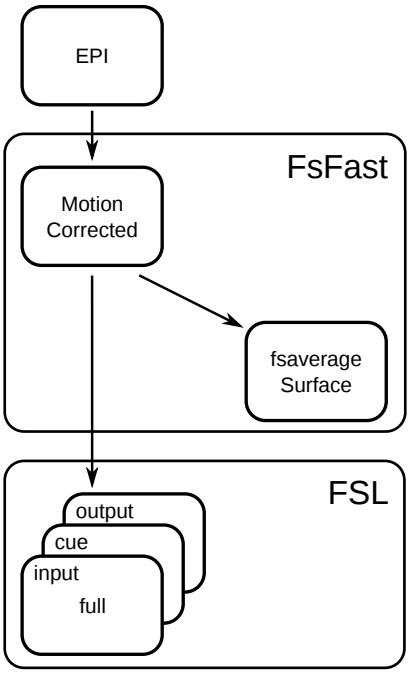


Figure 3.3: **Preprocessing pipeline.** (1) EPI volumes were imported into FsFast directory structure and realigned to the first volume of the first run, in each subject's native space. (2) Realigned volumes were resampled into the fsaverage cortical surface space, using 5mm FWHM smoothing. These datasets were used for univariate analyses. (3) Datasets were constructed for events timed to the auditory stimulus (*input*), repeat/change *cue*, and GO signal (*output*) in the subject-native space, using FSL. These datasets were used for multivariate analyses.

analyses, responses to individual events were estimated in the subject-native space with an FSL pipeline managed by Nipype (Gorgolewski et al., 2016).

3.2.6.1 Datasets for multivariate analyses

Response estimates for a given event type (*i.e.*, input, cue or output) across all trials were compiled into separate "datasets", one volume per trial, with a corresponding set of class labels (see Section 3.2.5.3). Input- and output-related datasets were constructed by simultaneously modeling the responses to the auditory stimulus presentation and the GO cue of each trial. Cue-related datasets were constructed by

modeling a single event per-trial, timed to the onset of the rectangular repeat or change cue.

Events were modeled as impulses of 0s duration, using the `SpecifySparseModel` (Perrachione and Ghosh, 2013) function of Nipype. Design matrices were constructed using FSL's `feat_model` and event estimation was performed with FSL's `film_gls`.

3.2.6.2 Motion outliers

We detected motion motion and intensity artifacts using RapidArt (Gorgolewski et al., 2016), with a norm threshold of 1.5mm and an intensity Z-threshold of 3. These artifacts were labeled as outlier volumes, in addition to volumes identified as overlapping with speech (see Section 3.2.5.2).

Outlier volumes were not excluded from GLM estimation (c.f. Siegel et al., 2014). Events where the magnitude of the theoretical HRF was >10% of its maximum height during one of these outlier volumes were marked for exclusion from multivariate analysis.

3.2.7 Univariate analysis

Univariate analyses were performed using FsFast on the cortical surfaces, in the `fsaverage` space. First-level analysis used 5mm FWHM smoothing. Second-level analyses are performed with `mri_glmfit`, and thresholded with `mri_glmfit-sim` at an uncorrected $p < 0.01$ and a cluster size threshold $p < 0.05$ (Bonferroni correction factor of 3, for separate analysis of left hemisphere, right hemisphere and subcortical responses (not presented here)).

Input- and output-related betas were estimated using the same GLM, and could be compared to one another. Because cue-related events were more-closely timed to

input-related than output-related events (see Figure 3.2) estimating cue-related betas in the same GLM would reduce evident activation at input, harming comparison. Therefore, cue-related responses were estimated separately, and could only be directly compared to other cue-related responses.

3.2.8 Multivariate analyses

Cortical searchlight analyses (Chen et al., 2011; Oosterhof et al., 2011) were performed within the PyMVPA (Hanke et al., 2009; Halchenko et al., 2015) framework. For each voxel halfway between the pial and white matter surfaces, a searchlight wass formed from the set of voxels intersecting the 9mm radius disk centered at the surface vertex nearest to that voxel. Each searchlight formed a feature vector for a classification analysis, and the result is a volume in which each voxel contains a statistic for the searchlight centered at that voxel.

Classification was performed using C support vector machines in a leave-one-run-out cross-validation configuration, and raw predictions were recorded for each run and concatenated, for subsequent analysis. To account for training set imbalances, each analysis was performed 10 times, down-sampling the training set of each fold so that each class label was equally represented in training.

Table 3.3 describes the searchlight analyses we considered in this study. Classification accuracy may be calculated for all trials, or separately for subsets of trials in which the stimuli were *words* or *nonwords*, or in which subjects *repeated* the auditory stimulus or *changed* to the unheard visual stimulus. In all cases, classifiers were trained on every trial. Each analysis may be described functionally; if a classifier was trained over the input, cue or output dataset, it was denoted as $I()$, $Q()$, or $O()$, respectively, with a superscript a or v indicating whether it was trained on

Dataset	Class label	Test sets	Symbol
input	aud	all	$I^a(all)$
		word	$I^a(word)$
		nonword	$I^a(nonword)$
cue	aud	all	$Q^a(all)$
		repeat	$Q^a(repeat)$
		change	$Q^a(change)$
	voc	all	$Q^v(all)$
		repeat	$Q^v(repeat)$
		change	$Q^v(change)$
output	voc	all	$O^v(all)$
		word	$O^v(word)$
		nonword	$O^v(nonword)$

Table 3.3: **Searchlight analyses and sub-analyses.** Searchlight analyses were performed over input, cue and output datasets; input and cue datasets were analyzed with the auditory stimulus class label, and cue and output datasets were analyzed with the vocal target class label. Input and output classification accuracy was calculated separately for trials with word and nonword stimuli; cue classification accuracies were calculated separately for repeat and change trials. Each (sub-)analysis may be described in functional notation (see text for details).

aud or *voc* class labels (reflecting the stimulus heard or produced, respectively; see Section 3.2.5.3). The parameter indicates the subset of trials described by the accuracy statistic. So, $Q^a(change)$ would indicate classifiers trained on the cue dataset to classify the vowel heard, with accuracy measured over the subset of change trials.

3.2.8.1 Multivariate contrasts

A pair of searchlight analyses may be contrasted in order to highlight a difference in information when running the same analysis on two datasets or different analyses on the same dataset. To perform each contrast, accuracy maps were subtracted within each subject and the differences z-scored using the spatial mean and standard deviation. A two-tailed t-test was used to determine regions which consistently show greater accuracy in one analysis than the other, across subjects.

Contrast	Description
$I^a(word) - I^a(nonword)$ $O^v(word) - O^v(nonword)$	Word and nonword specific responses to auditory stimuli and vocal targets
$Q^a(change) - I^a(change)$	Change in response to auditory stimulus following change cue
$Q^a(change) - Q^v(change)$	Distinct responses to auditory stimulus and vocal target following change cue
$O^v(all) - Q^v(all)$	Distinct responses to motor plan and motor act

Table 3.4: **Searchlight contrasts.** Searchlight accuracy maps (see Table 3.3) were contrasted to identify differences in information content across datasets, class labels, or test subsets.

Table 3.4 lists a series of contrasts of interest between two separate searchlight analyses.

3.2.8.2 Nonparametric significance testing

To perform cluster-extent based thresholding, we generated a null distribution of chance cluster sizes, adapting the technique described in Section 2.2.7.1 to consider testing on subsets of trials.

For each subject, class labels were permuted 100 times, and classifiers were retrained. Datasets were down-sampled to balance class labels in the training set, and classification accuracy rates were stored separately for repeat and change testing subsets. Trial types were not permuted. Classification accuracy for all trials is taken to be an unweighted average of repeat and change accuracies. Pending separate computation of chance classification accuracies for word and nonword trials, in all cases repeat accuracy is substituted for word and change accuracy is substituted for nonword.

Chance relative accuracy maps are computed by removing the spatial mean (see Section 3.2.8.3) from a randomly selected permutation from each subject, and per-

forming a one-sided t-test. A vertex-wise p-value threshold is applied to the resulting map; cluster extents are defined to be the number of connected vertices on the cortical surface. Null distributions are calculated from the cluster extents of 1000 chance accuracy maps.

Searchlight contrasts Searchlight analyses are contrasted with a two-sided t-test of differences of accuracies. For two analyses A and B, a random permutation of A is selected for each subject, and a random permutation of B is subtracted from it. The difference of spatial means of each (see Section 3.2.8.3) is subtracted, and the result is divided by the spatial standard deviation, resulting in a random z-score map for each subject. Applying a two-sided t-test to each vertex and thresholding at $p < 0.05$ (uncorrected), we constructed a set of chance cluster sizes. We repeat this method 1000 times to construct a null distribution.

3.2.8.3 A note on cortical spatial transformations

To minimize smoothing and oversampling, all classification analyses are performed in the subject-native, volumetric space. Group-level analyses must be performed in some common space, in this case the `fsaverage` surface. Although the spatial transformation between these spaces is non-linear, the transformation of values is linear, *i.e.*, the value at each vertex is a linear combination of the values at its contributing voxels, and thus the transformed sum of two maps is equal to the sum of the transformed maps.

Thus, it should be noted that summary statistics such as spatial means must be *calculated* in the subject-native space to avoid distortion, but they may be added to or multiplied by a map of values in `fsaverage` space without issue.

3.3 Results

3.3.1 Behavioral results

All subjects' vocalizations were manually inspected for correct task performance. Figure 3.4 shows the number of trials for each type of recoding (see Section 3.2.5). A mean of 7.2 ($\sigma = 5.6$) trials were recoded for each subject. A "repeat" designation indicates the subject repeated the auditory stimulus when instructed to produce the unheard syllable; a "change" designation indicates the subject produced the unheard syllable when instructed to repeat the auditory stimulus. Among subjects with "repeat" or "change" recodings, no preference was found for either recoding (two-sided t-test; $p = 0.84$).

One subject's (S20) vocalizations were not recorded for the final 35 trials of the session. Due to the low error rate in recorded trials, correct responses were assumed.

Figure 3.5 shows the number of trials excluded from each dataset due to motion, on account of motion artifacts detected in the data ("motion") or manually detected speech during a scan ("speech"). Trials affected by both are marked "motion", and not counted twice. See Section 3.2.6.2 for outlier attribution details.

3.3.2 Univariate results

Univariate contrasts were performed using two FsFast pipelines: the evenly-spaced input- and output-linked responses could be estimated with a single GLM, while the cue-linked response must be estimated separately. Each pipeline produced left and right hemisphere datasets on the `fsaverage` surfaces. All results are thresholded at $p < 0.01$, and the resulting clusters are subjected to size thresholding based on non-parametric Monte Carlo simulations. All results presented here have an additional Bonferroni correction factor of 3, for two hemispheres and subcortical space

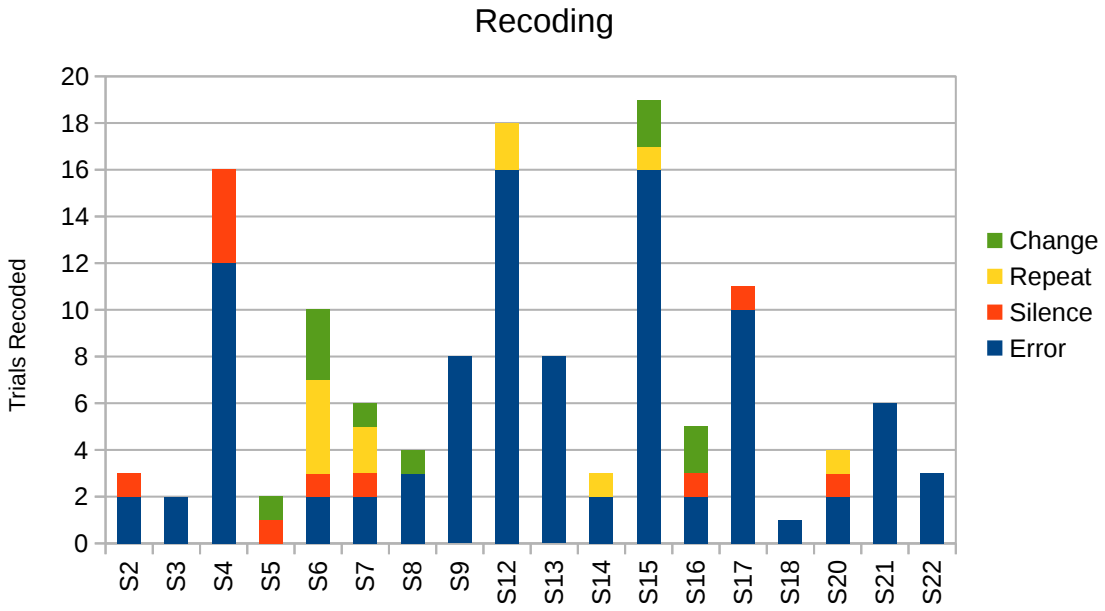


Figure 3.4: Number of recoded trials, broken down by class. Error indicates subjects produced an invalid syllable. Silence indicates subjects did not produce an audible syllable. Repeat indicates subjects incorrectly repeated the auditory stimulus. Change indicates subjects incorrectly produced the unheard, visually presented syllable. Trials coded as "error" or "silence" are considered invalid and excluded from classification analyses using vocal targets as labels; trials coded as "repeat" or "change" are used in classification analyses with the actually spoken syllable.

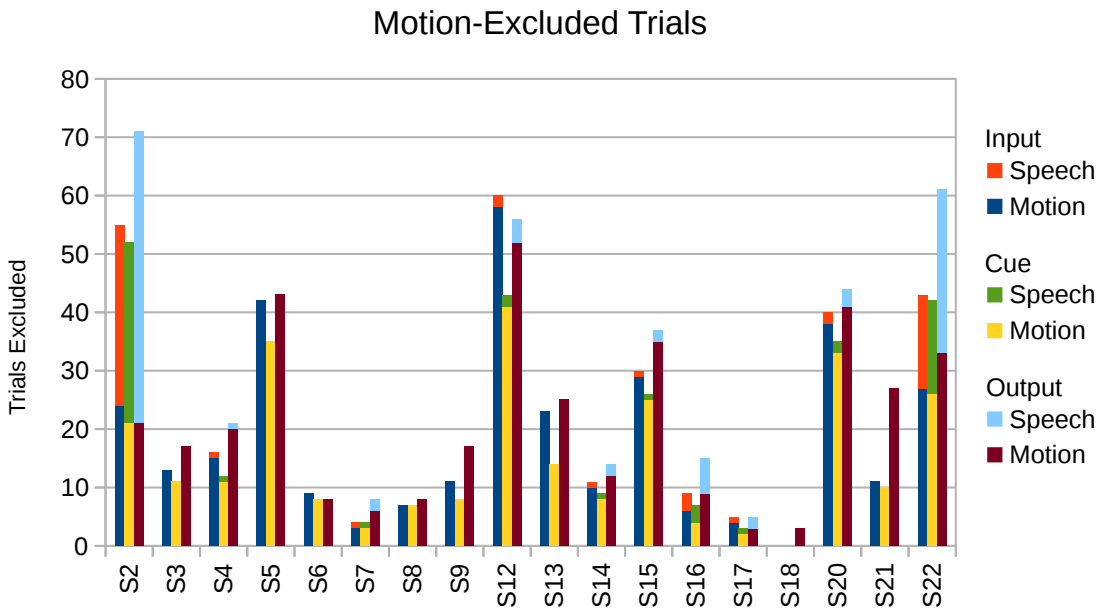
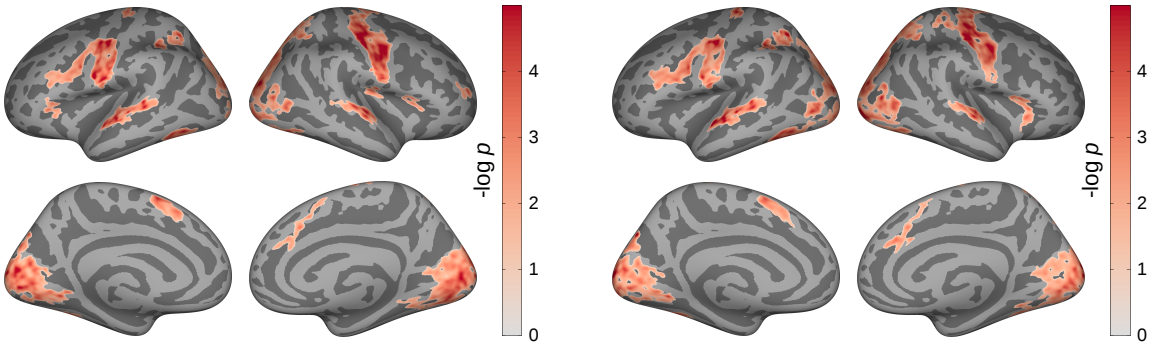
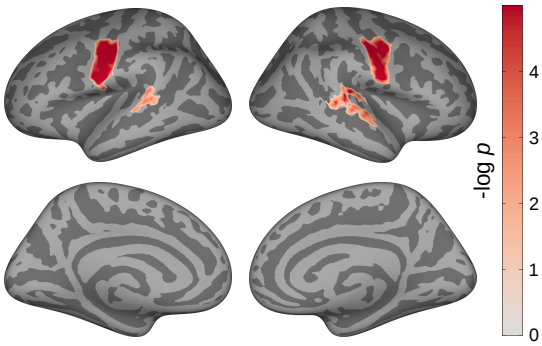


Figure 3.5: Number of trials excluded from each dataset (input, cue and output) due to head motion. Trials excluded due to motion artifacts detected by RapidArt are marked "motion". Trials excluded due to speech during a scan are marked "speech". A trial excluded for both reasons is marked "motion". Subjects S6 and S14 participated in 176 total trials; all other subjects participated in 352 trials.



(a) Input events

(b) Cue events



(c) Output events

Figure 3.6: **Main effect of task.** Contrasts between task trials (with syllable stimuli) and control trials (with pink noise stimuli) in the (a) input- (b) cue- and (c) output-linked hemodynamic responses. Shown are *significance* statistics, *i.e.* $-\log_{10} p$, from a one-tailed t-test (17 dof), thresholded at $p < 0.01$ and cluster-wise corrected at $p < 0.05$. Cluster thresholds are Bonferroni corrected by a factor of 3, for separate analyses of two hemispheres and sub-cortical regions.

(subcortical results not presented here).

Figure 3.6 contrasts beta estimates associated with task trials, in which subjects heard a spokens syllable, and control trials, in which subjects heard a pink-noise stimulus, timed to the stimulus onset (input-linked), cue onset (cue-linked) or GO signal onset (output-linked). Responses at input and cue are very similar, predominantly engaging superior temporal gyrus, somatosensory and motor (Rolandic) cortex and visual cortex, bilaterally, as well as left ventral premotor cortex (vPMC), posterior inferior frontal sulcus (pIFs) and superior parietal regions. At output, bilateral ven-

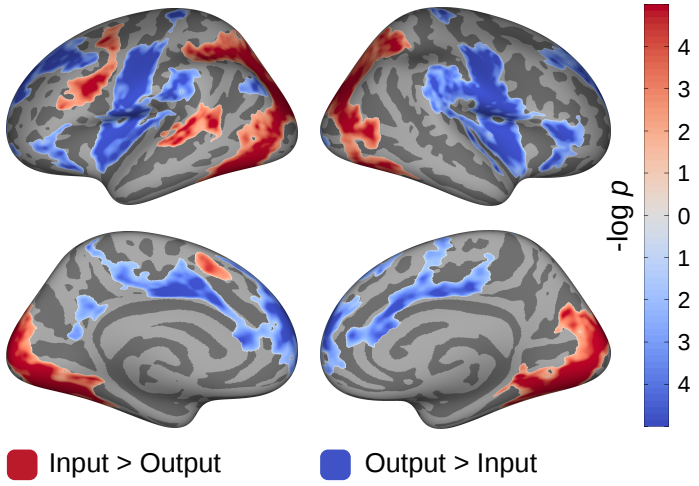
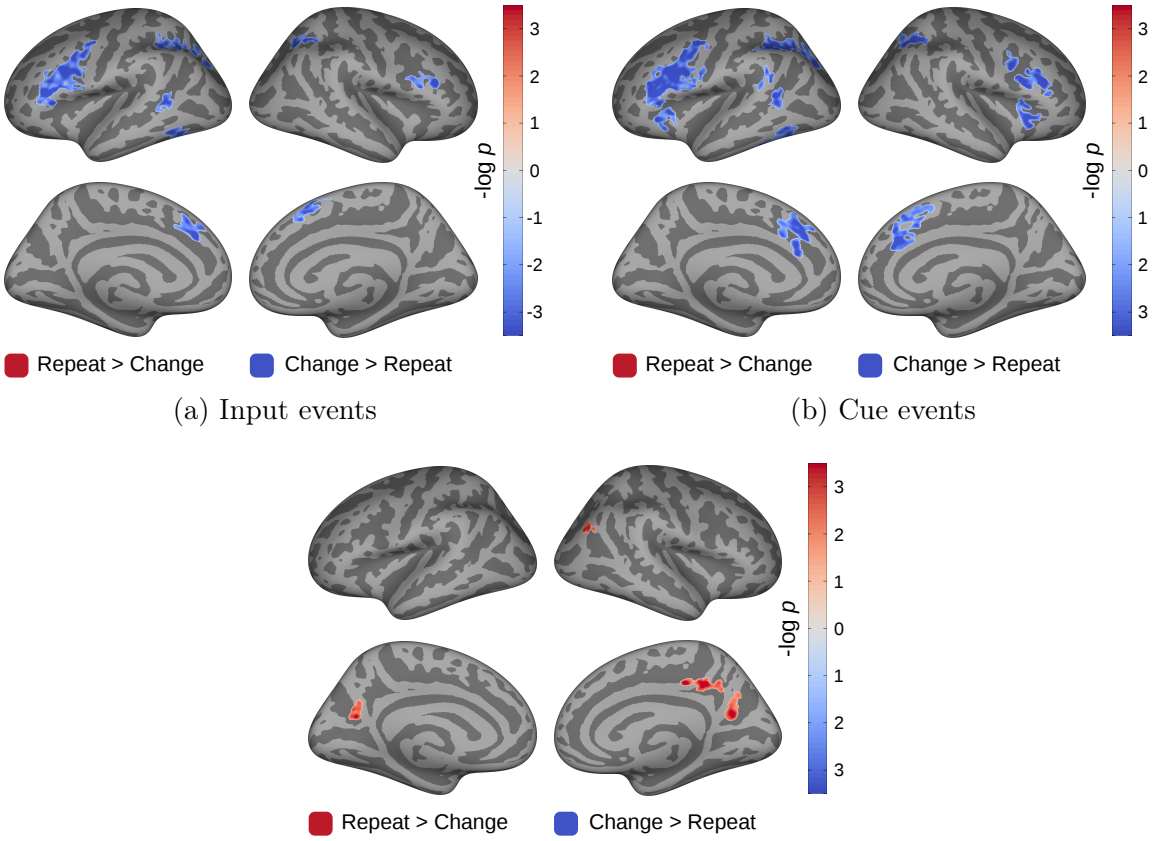


Figure 3.7: **Distinct responses to input and output.** Contrasts between input- and output-linked hemodynamic responses during all task trials. Shown are *significance* statistics, *i.e.* $-\log_{10} p$, from a two-tailed t-test (17 dof), thresholded at $p < 0.01$ and cluster-wise corrected at $p < 0.05$. Cluster thresholds are Bonferroni corrected by a factor of 3, for separate analyses of two hemispheres and sub-cortical regions.

tral Rolandic cortex is significantly more active during task trials than control trials, as is, to a lesser extent, bilateral superior temporal cortex.

Figure 3.7 contrasts input- and output-linked beta estimates during task trials only. Input-linked responses are significantly stronger in the left hemisphere in the superior temporal gyrus (STg) and posterior superior temporal sulculs (pSTs), as well as inferior frontal sulcus (IFs), medial frontal gyrus/middle-dorsal premotor cortex (MFg/mdPMC) and the anterior portion of the supplemental motor area (pre-SMA). In addition, input-linked responses in occipital cortex, stretching bilaterally into parietal and inferior temporal cortex, are stronger than output-linked responses. Output-linked responses are significantly stronger, bilaterally, in ventral Rolandic cortex, supramarginal gyrus (SMg), insula and cingulate gyrus.

Figure 3.8 contrasts beta estimates for repeat and change trials (averaged across other experimental factors), for each of the input, cue and output datasets, with



(a) Input events

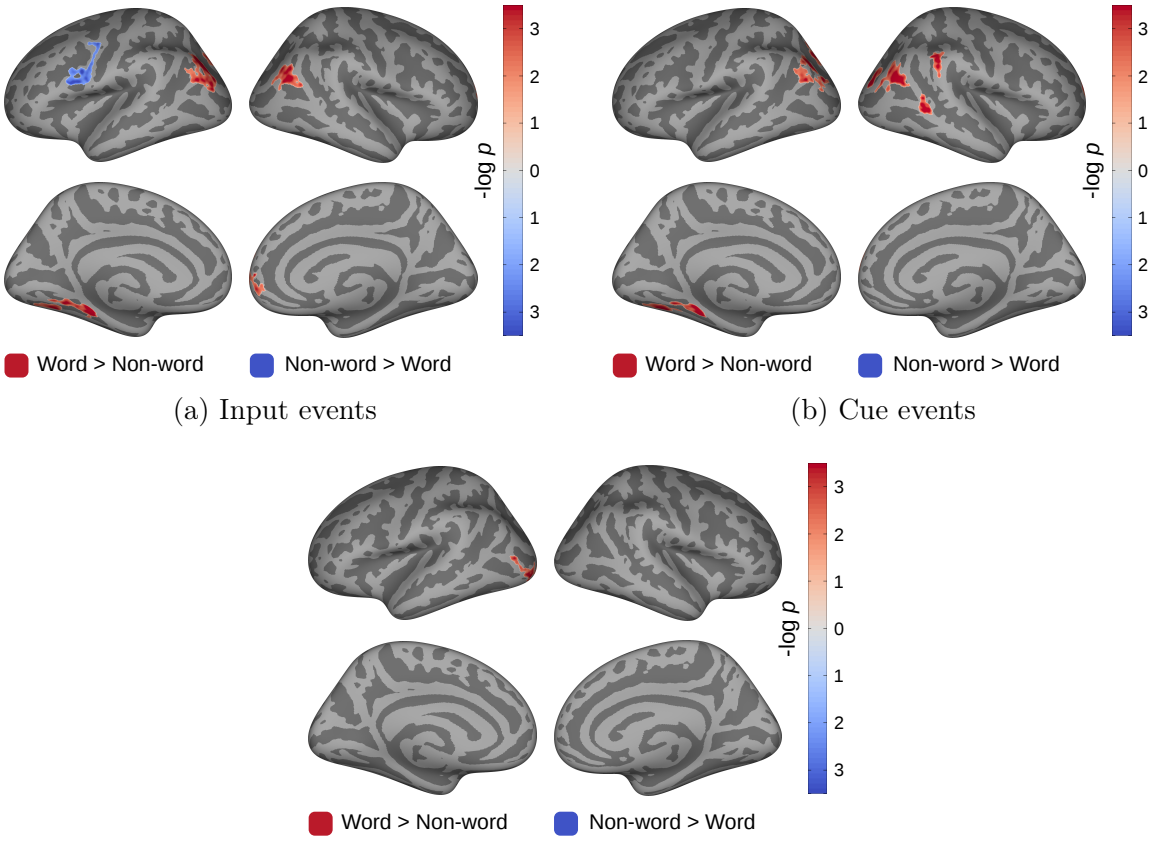
(b) Cue events

(c) Output events

Figure 3.8: **Response to *repeat* vs *change* trials.** Contrasts between repeat trials and change trials in the (a) input- (b) cue- and (c) output-linked hemodynamic responses. Shown are *significance* statistics, *i.e.* $-\log_{10} p$, from a two-tailed t-test (17 dof), thresholded at $p < 0.01$ and cluster-wise corrected at $p < 0.05$. Cluster thresholds are Bonferroni corrected by a factor of 3, for separate analyses of two hemispheres and sub-cortical regions.

regions appearing red where there is a greater response on repeat trials and blue where there is a greater response on change trials. At input and cue, only regions with stronger responses to change trials reach significance, predominantly left ventral premotor cortex, inferior frontal sulcus and dorsal inferior frontal gyrus. Left posterior STs, right dorsal IFg, bilateral intraparietal sulcus and bilateral medial prefrontal areas show smaller significant clusters. In addition, the cue contrast reveals bilateral orbitofrontal cortex responding more to change trials than repeat. At output, stronger responses to repeat trials than change trials were found in right posterior angular gyrus, right subparietal sulcus, and bilateral parieto-occiptal sulcus.

Figure 3.9 contrasts beta estimates for word and nonword trials (averaged across other experimental factors), for each of the input, cue and output datasets. Regions appear red where words induce a stronger response and blue where nonwords induce a stronger response. At input, words cause a stronger response than nonwords in bilateral posterior AG and left collateral sulcus, while nonwords produce a stronger response in ventral premotor cortex, extending into posterior IFg. At cue, stronger responses to words appear in bilateral posterior AG, right posterior surpamarginal gyrus, and right posterior middle temporal gyrus. At output, a stronger response to words appears in visual cortex.



(a) Input events

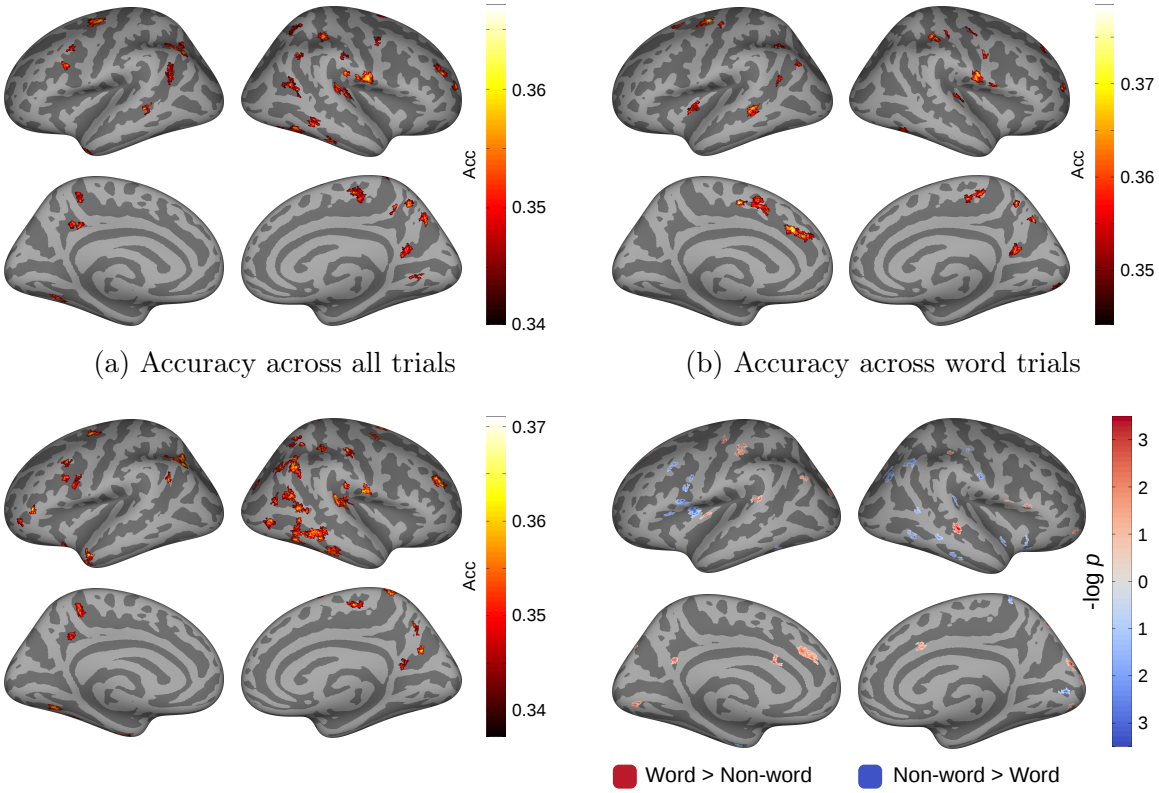
(b) Cue events

(c) Output events

Figure 3.9: **Response to word vs nonword trials.** Contrasts between word trials and nonword trials in the (a) input- (b) cue- and (c) output-linked hemodynamic responses. Shown are *significance* statistics, *i.e.* $-\log_{10} p$, from a two-tailed t-test (17 dof), thresholded at $p < 0.01$ and cluster-wise corrected at $p < 0.05$. Cluster thresholds are Bonferroni corrected by a factor of 3, for separate analyses of two hemispheres and sub-cortical regions.

3.3.3 Multivariate results

3.3.3.1 Vowel information in input datasets



(a) Accuracy across all trials

(b) Accuracy across word trials

(c) Accuracy across nonword trials

(d) Word vs nonword accuracy contrast

Figure 3.10: **Auditory stimulus information in word/nonword contexts.** Mean classification accuracies across subjects for searchlight analyses trained on auditory stimulus vowel identity over all input-related responses and tested on (a) all, (b) word, and (c) nonword trials. (d) A two-tailed comparison of classification accuracy over word and nonword trials. All figures are thresholded at a vertex-wise threshold of $p < 0.05$ and a cluster-wise threshold of $p < 0.01$.

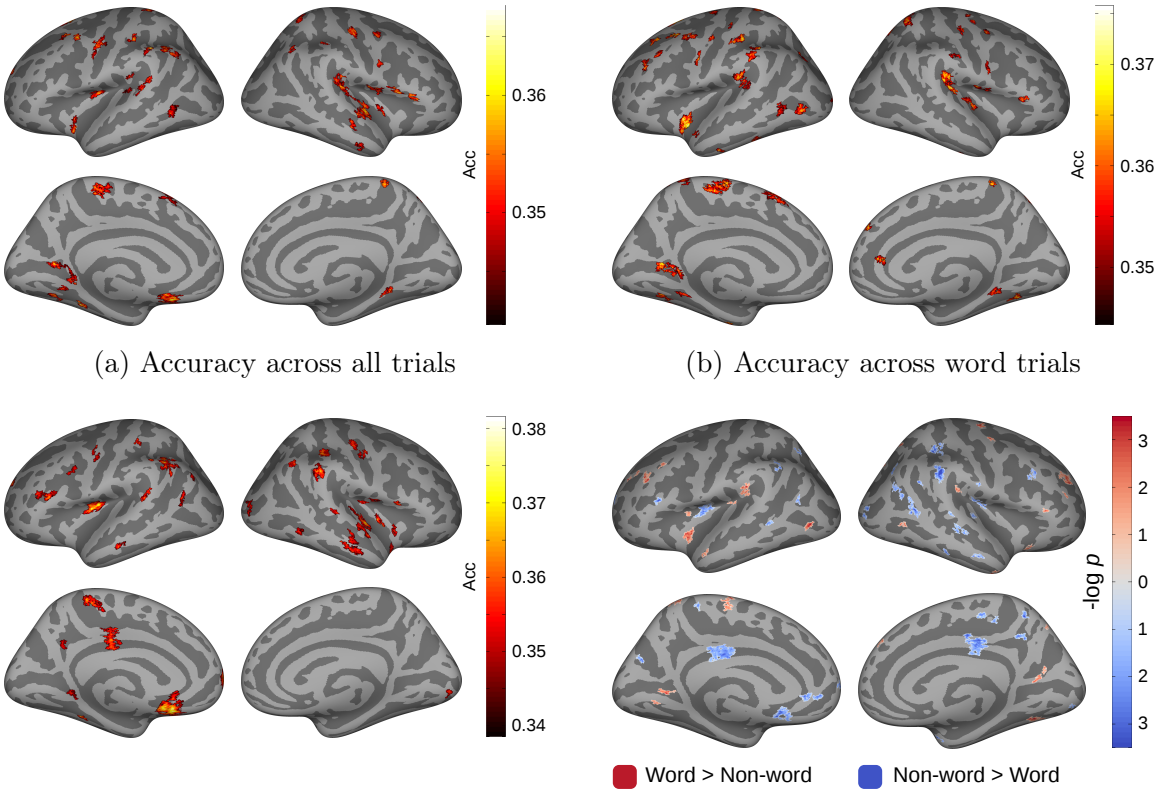
Figure 3.10a shows classification results for searchlight analyses trained on auditory stimulus vowel identity over input-related datasets. In the left hemisphere, we find correlates with stimulus identity in posterior IFs and inferior middle frontal gyrus (pMFg), posterior STs, angular gyrus and the superior part of the precentral sulcus. Calculating accuracy only across word trials (Figure 3.10b) results in a larger

pSTs cluster with higher raw accuracy, but sub-threshold classification in pIFs, while the pIFs and pMFg clusters remain significant on nonword trials (Figure 3.10c), but the pSTs cluster disappears. Significant classification of nonwords is also found in an additional ventral premotor cluster. Contrasting word and nonword accuracies directly (Figure 3.10d) a significant preference for vowel identity in words is found in pSTg/PT, posterior AG and ventral posterior insula, and a preference for nonwords is found in pMFg, vPMC, vIFo and dorsal anterior insula.

In the right hemisphere, a large posterior central operculum cluster is found in word and nonword trials. A right PT cluster appears in all subsets of trials, and it is larger and with significantly higher accuracies in nonword trials than in word trials. Larger nonword clusters are found in right inferior parietal and posterior temporal regions, including posterior angular gyrus. A significant preference for words is found in right pSTs and vIFo, while a significant preference for nonwords is found in PT and AG.

3.3.3.2 Vowel information in output datasets

Figure 3.11a shows mean classification accuracies for searchlight analyses trained on vocal target vowel identity over output-related estimates. We find correlates, in bilateral PT/pSTg, insula, AG and posterior MFg, in addition to left-lateralized SMg and right-lateralized planum polare/aSTg, aSTs, and vIFo. Testing classification only on word trials (Figure 3.11b), larger clusters with higher mean accuracies are observed in bilateral PT, as well as left SMg, inferior insula and posterior MFg; similar clusters are found in right vPMC and IFg. Non-word specific accuracy (Figure 3.11c) shows larger clusters in left superior insula, right SMg and PP/aSTg. Non-words are also represented in left IFs.



(a) Accuracy across all trials

(b) Accuracy across word trials

(c) Accuracy across nonword trials

(d) Word vs nonword accuracy contrast

Figure 3.11: **Vocal target information at output in word/nonword contexts.** Mean classification accuracies across subjects for searchlight analyses trained on vocal target vowel identity over all output-related estimates and tested over (a) all, (b) repeat, and (c) change estimates. All figures are thresholded at a vertex-wise threshold of $p < 0.05$ and a cluster-wise threshold of $p < 0.01$.

Figure 3.11d contrasts classification accuracies on word (red) and nonword (blue) trials. The separation of inferior/superior insula between word and nonword information, respectively, is evident. Posterior PT/parietal operculum shows a bilateral preference for words, and right SMg and pSTs show a preference for nonwords.

3.3.3.3 Vowel information in cue datasets

The cue informs subjects of the trial type (repeat or change), and thus which (non)-word they will need to produce. Figure 3.12 shows average classification accuracies

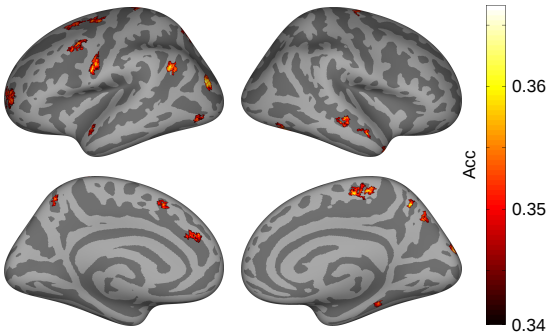
when training on the auditory stimulus vowel identity over all cue trials, and testing over (a) all, (b) repeat, or (c) change trials. Notably, the left superior IFs/MFg cluster appears when testing on repeat trials, but not change, and bilateral STs clusters contain significant predictive information when testing on change trials, but not repeat trials. Further, ventral motor cortex predicts the auditory stimulus in both repeat and change trials. Other clusters are found in left superior BA6, angular gyrus (AG) and pre-SMA across all trials and change trials, while a large medial motor/somatosensory cluster appears for all trials and repeat trials. Right anterior supramarginal gyrus (aSMg) predicts the auditory stimulus in repeat trials, and right ventral somatosensory cortex predicts the auditory stimulus in change trials.

Figure 3.13 shows the same classification, when responses were labeled with the vocal target. The ventral motor cortex cluster is absent when testing on any subset of trials, while the superior IFs/MFg cluster is present when testing on any subset. Additionally, testing only on change trials reveals clusters in left vPMC and posterior PT, and the STs cluster is absent.

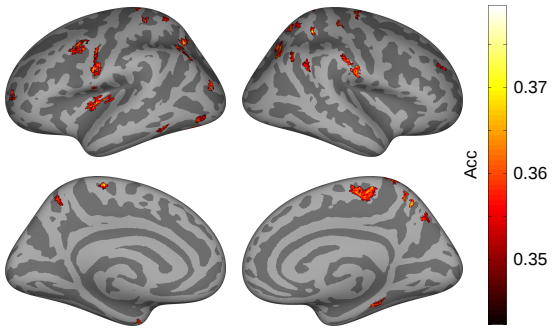
Right inferior parietal cortex, particularly supramarginal gyrus (SMg) and angular gyrus (AG) contains clusters predictive of vocalized vowels in repeat trials, and to a lesser extent in change trials, while right ventral somatosensory cortex contains clusters predictive of the auditory stimulus in change trials. Additionally, a large cluster in the dorsal postcentral gyrus predicted vocalized vowels across any subset of trials.

3.3.3.4 Informational contrasts

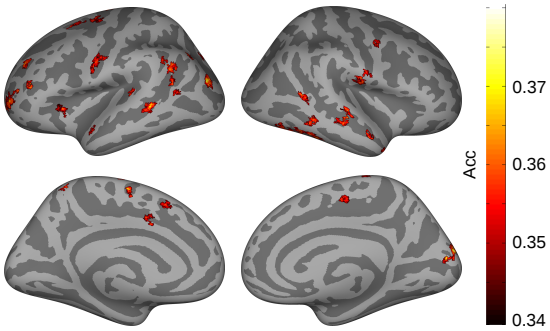
We next consider differences that occur specifically in change trials, in which subjects are required to speak a different syllable than the one they heard. Figure 3.14a shows



(a) Accuracy across all trials

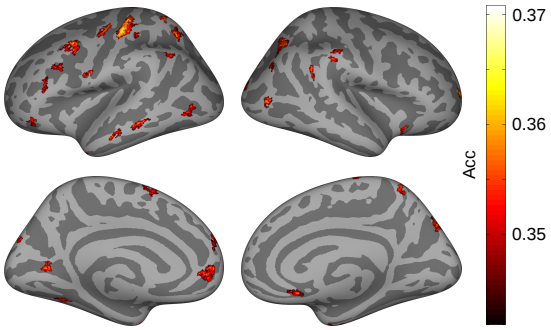


(b) Accuracy across repeat trials

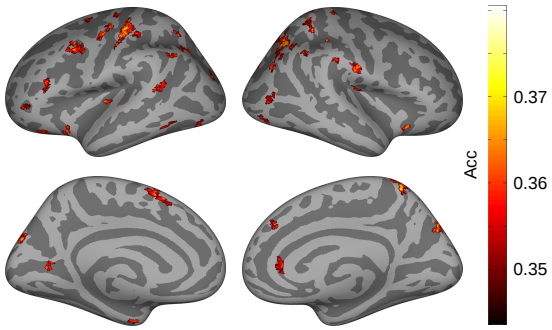


(c) Accuracy across change trials

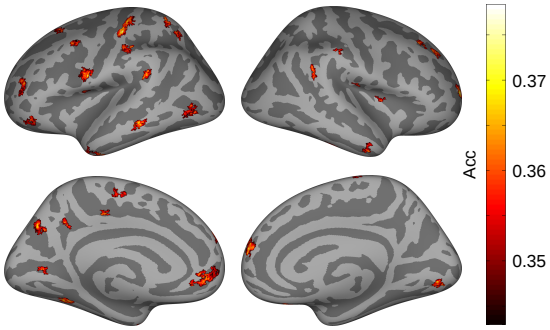
Figure 3.12: **Auditory stimulus information at cue in repeat and change trials.** Mean classification accuracies across subjects for searchlight analyses trained on auditory stimulus vowel identity over all cue-related estimates and tested over (a) all, (b) repeat, and (c) change estimates. All figures are thresholded at a vertex-wise threshold of $p < 0.05$ and a cluster-wise threshold of $p < 0.01$.



(a) Accuracy across all trials

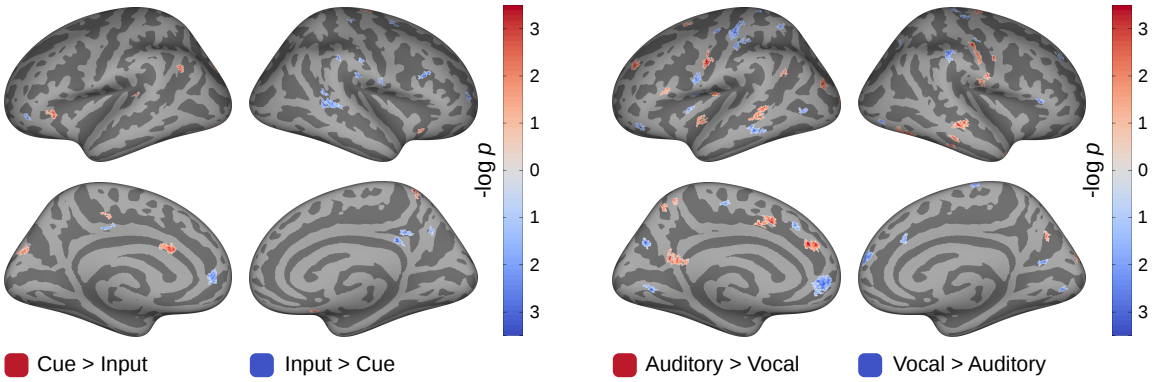


(b) Accuracy across repeat trials



(c) Accuracy across change trials

Figure 3.13: **Vocal target information at cue in repeat and change trials.** Mean classification accuracies across subjects for searchlight analyses trained on vocal target vowel identity over all cue-related estimates and tested over (a) all, (b) repeat, and (c) change estimates. All figures are thresholded at a vertex-wise threshold of $p < 0.05$ and a cluster-wise threshold of $p < 0.01$.



(a) Auditory stimulus at cue and input

(b) Stimulus and vocal target at cue

Figure 3.14: **MVPA Contrasts: Auditory stimuli and vocal targets at cue.** (a) A contrast of classification of auditory stimulus vowel identity over cue and input estimates, testing only on change estimates. (b) A contrast of classification of auditory stimulus and vocal target vowel identities over cue estimates, tesitng only on change estimates. Shown are significance statistics, *i.e.*, $-\log_{10}(p)$, from a two-tailed t-test (17 dof). All figures are thresholded at a vertex-wise threshold of $p < 0.05$ and a cluster-wise threshold of $p < 0.01$.

differences in classification accuracy of auditory stimulus vowel identity at input (blue) and at cue (red), testing only on change trials. Following the change cue, subjects know they may discard the auditory stimulus and prepare to speak the syllable that was read but not heard. Most notably, right hemisphere clusters in pSTs, pIFs, SMg and pCO are found to have higher accuracy at the time of auditory input than at the time of the cue. In contrast, left PT and AG, and bilateral orbitofrontal cortex show stronger predictive information about the vowel heard at cue than at input. Figure 3.14b shows differences in classification accuracies at the time of the cue for trials labeled with the auditory stimulus vowel identity (red) compared to the vocal target vowel identity (blue), testing again only on change trials. Auditory stimuli are significantly better classified in bilateral pSTs and motor cortex, as well as left IFo and AG. Vowel vocal targets are better classified in left vPMC and pMTg, and right SMg and vIFt.

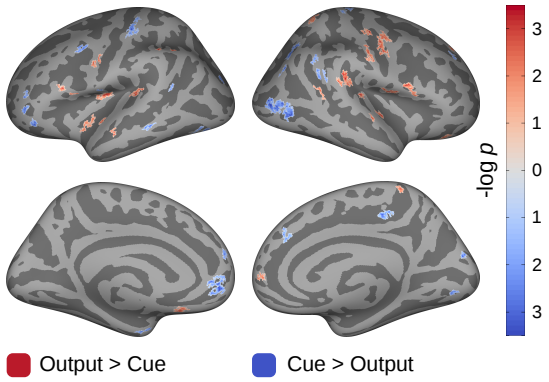


Figure 3.15: **MVPA Contrast: Vocal targets at output and cue.** Comparisons of classification accuracies for searchlight analyses trained on vocal target vowel identity for output (red) and cue (blue) datasets. Shown are significance statistics, *i.e.*, $-\log_{10}(p)$, from a two-tailed t-test (17 dof), thresholded at a vertex-wise threshold of $p < 0.05$ and a cluster-wise threshold of $p < 0.01$.

Finally, we consider the evolution of responses corresponding with the vocal plan between the times at which subjects know (cue) and produce (output) the vocal target. Figure 3.15 shows difference in classification of vocal target vowel identity over output- (red) and cue-related responses (blue). More reliable correlates of the vocal target are found at cue in left posterior MFg (adjacent to the inferior frontal sulcus), pMTg and dorsal somatosensory cortex, and in right angular gyrus and middle occipital gyrus. At output, more reliable correlates are found in bilateral PT, IFo, and Rolandic cortex, as well as left insula and anterior STs.

3.4 Discussion

In this study, we applied cortical surface searchlight-based MVPA to estimated BOLD responses during a delayed repetition task in which subjects selected a response from visually and aurally presented (non)words. We estimated hemodynamic responses to individual events, timed to three stages of the task: aural stimulus presentation (input), task-selection cue (cue), and GO signal presentation (output). We analyzed these datasets for correlates of the phonological content of the auditory stimulus and vocal target, namely the vowel identity, in order to dissociate the processing of auditory inputs and planned speech outputs. Below, we first discuss the input and output stages of the task, with a focus on the differential response to word and nonword syllables. We then consider the cue event, and the process of selecting a response from two items in working memory. This is followed by a broader synthesis of results in the context of current theories of speech repetition and verbal / phonological working memory.

3.4.1 Input-related responses

Within each trial, subjects were presented visually with two (non)words at $t = 0$s, and aurally with one of these (non)words at $t = 2.25$s. The input event of each trial was timed to this aural presentation. Which syllable was to be spoken was unknown at this point in the trial, and thus estimated responses were expected to be most reflective of hearing a particular syllable.

Figure 3.6a shows significant task-related activation at input in bilateral superior temporal gyrus (STg) and Rolandic cortex, and left posterior superior temporal sulcus (pSTs), posterior inferior frontal sulcus (pIFs), as well as the anterior portion of the supplementary motor area (pre-SMA). Contrasting with task-related responses at

output (Figure 3.7), these responses are comparatively strong in left pSTs/pSTg and pIFs into middle frontal gyrus (MFg), and in pre-SMA. Note that these frontal regions, typically associated with speech planning and production (*e.g.*, Bohland et al., 2010), are active in the absence of definitive knowledge of the syllable to be spoken, suggesting either a role in sensory perception or an automatic activation of the speech output system after hearing the syllable (although, see Section 3.4.6).

Further, contrasting responses when the stimuli are both words to those when the stimuli are both nonwords (Figure 3.9a) reveals a stronger response to nonwords in left mid-/ventral premotor cortex. At a relaxed threshold ($p < 0.05$ uncorrected; cluster-wise threshold $p < 0.05$), left anterior supramarginal gyrus (SMg) also shows a stronger response to nonwords. Overall, this is consistent with the view that reading and listening to nonwords makes greater demands of the dorsal stream (Saur et al., 2008). Additionally, stronger responses to words than nonwords were found in bilateral angular gyrus (AG), which has been implicated in lexical and semantic processing of written and spoken words (Binder et al., 2005; Raettig and Kotz, 2008; Binder et al., 2009).

Clusters in pIFs and pSTs were found to be significantly predictive of the auditory vowel identity, across all trials (Figure 3.10a), suggesting their involvement directly in perceiving the auditory stimulus or in encoding the speech sounds heard in short-term memory to enable the subject to successfully complete the remainder of the task. Testing classification accuracy on specific subsets of trials (words vs. nonwords; Figure 3.10b,c) provided additional insight into the differential roles of these areas for lexical and non-lexical stimuli. Accuracy across subsets of trials shows that pSTs remains significantly predictive in words, while pIFs remains significantly predictive in nonwords, but not vice versa. Figure 3.10d shows neither difference

reaches significance when contrasting accuracies directly, though posterior middle frontal gyrus (pMFg) and vPMC clusters show a nonword preference, corroborating a greater frontal representation for nonwords, while a right pSTs cluster homologous to the left pSTs cluster is found to have a preference for encoding vowels within words.

3.4.2 Output-related responses

Within each trial, subjects were cued to speak at $t = 9$s, and output events were timed to this visual cue. At this point in the trial, subjects had prepared to produce a single syllable, and there was no requirement to actively maintain a trace of the other syllable.

Figure 3.6c shows significant task-related activation at output in bilateral pSTg and Rolandic cortex; contrasting with task-related responses at input (Figure 3.7), these responses are comparatively strong in bilateral Rolandic cortex, extending posteriorly into the supramarginal gyri / parietal opercula, and medially into the insula. Additionally, superior orbitofrontal cortex (slightly right lateralized) and medial regions – most notably cingulate cortex – have increased activation at output over input. There were no major differences in word or nonword trials at the output stage of the task (Figure 3.9c), consistent with expectations, as the motor execution of phonotactically legal syllables is unlikely to be affected by their lexical status.

Regions predictive of vocal target vowel identity at output (Figure 3.11) are primarily superior temporal, inferior parietal and insular, although a pMFg cluster appears for words and a slightly more ventral pIFs cluster appears for nonwords. Bilateral posterior PT (area Spt) is strongly predictive of words, while nonwords induce small, bilateral clusters in pSTg. A cluster in mid-motor cortex is found not far from the articulatory maps for lip, tongue and respiration (Takai et al., 2010; Bouchard

et al., 2013).

The bilateral superior temporal clusters are statistically significant when classification accuracy is assessed in change trials alone (not shown), for which no auditory memory of the vocalized syllable is available. Thus, the right aSTg/PP cluster and the bilateral pSTg clusters are most likely to represent auditory targets or expectations of the speech act or auditory responses to self-produced speech.

The emergence of an inferior/superior preference within left insula for words and nonwords, respectively, is worth noting and, to our knowledge, not previously reported. A recent structural connectivity analysis of the insula suggests that the anterior inferior portion of the insula is highly connected in the ipsilateral hemisphere with both anterior superior temporal lobe and frontal regions including IFg pars triangularis (Ghaziri et al., 2015), and thus may sit at the interface of the temporal and frontal portions of the ventral stream, though no functional significance is proposed. The superior region is highly connected with ipsilateral motor and premotor cortex, and anterior insula has been previously implicated in apraxia of speech (Dronkers, 1996). Additionally, anterior insula has been suggested to be engaged by novel or complex utterances (Ackermann and Riecker, 2004; Sörös et al., 2006; Baldo et al., 2011). While nonwords have increased novelty in comparison to words, we did not see a greater univariate response to nonwords (Figure 3.9c).

3.4.3 Cue-related responses

Within each trial, subjects were presented with a green or yellow rectangular cue at $t = 3.25$s, informing them which syllable was to be spoken. A green cue indicated subjects were to repeat the auditory stimulus (repeat trial), while a yellow cue indicated subjects were to repeat the unheard, visually presented stimulus (change trial).

The cue event of each trial was timed to this visual presentation.

Figure 3.8b shows a generally stronger response at cue for change trials than repeat trials, reflecting a greater cognitive demand during change trials. Although repeat and change trials were evenly balanced, and subjects would not have benefited by using an active strategy of "pre-loading" the heard stimulus into a speech motor plan, this suggests that, prior to the cue, frontal speech areas nonetheless load the auditory stimulus instead of the unheard visual stimulus. In particular, bilateral vMC, vPMC, IFs and anterior insula were engaged more in response to a change cue than a repeat cue, as were left posterior STs and SMg/PT, likely reflecting processes involved in replacing the target vocal plan from the auditory stimulus to the unheard visual stimulus. Interestingly, the strongly left lateralized pSTs cluster that shows greater activation on change trials overlaps with the posterior peak of a cluster in Figure 3.7, where the estimated response was significantly greater to input than to output, indicating that it is preferentially re-activated on change trials; this response was also left lateralized.

If it is the case that, following a change cue, subjects extinguish representations related to the auditory stimulus from frontal cortex and load a speech motor plan for the visual stimulus, we should expect to find statistically significant representations of both the auditory stimulus and the vocal target at this stage of the task. Regions predictive of the auditory stimulus at the cue may either be slower to refresh their representation or more explicitly involved in auditory perception than in generating a speech motor plan.

Multivariate pattern analysis results help us to understand how the observed activations relate to the two relevant syllables presented in the trial. Let us first consider classification accuracies for cue events labeled with auditory stimulus vowel identity

(Figure 3.12). Left ventral motor cortex (vMC) predicts the auditory stimulus with relatively high accuracy in both repeat and change trials. This region likely reflects the articulatory configuration of the tongue (Takai et al., 2010; Bouchard et al., 2013), one of the primary determinants of vowel sounds. An early, motoric representation of the identity of the perceived vowel identity is consistent with proposals that motor engagement is a necessary component of speech perception (D'Ausilio et al., 2009; Pulvermüller and Fadiga, 2010). In contrast to the cluster in vMC, the pMFg cluster evident across all trials (Figure 3.12a) reaches significance in repeat, but not change, trials, and the left anterior AG and right mid-/aSTs clusters are found in change, but not repeat trials. A left pSTs cluster found specifically in change trials (Figure 3.12c) is in the same location that showed significant auditory vowel prediction at the input stage of the task (Figure 3.10a). These results would suggest that, when cued to repeat the auditory stimulus, there is little need to access auditory cortical representations of the heard syllable, relying instead on frontal representations that have already (and perhaps automatically) been engaged by hearing the syllable; when cued to change to the unheard stimulus, auditory representations of the auditory stimulus appear to be reactivated.

Next, consider classification accuracies for cue events labeled with vocal target vowel identity (Figure 3.13). The vMC cluster that predicted the heard vowel is entirely absent, but a vPMC cluster has replaced it, which is more apparent during change trials. Figure 3.14b, which directly contrasts the classification accuracy maps, confirms both clusters to have significant differences in accuracy for classifiers trained on the two class labels. Near the inferior frontal sulcus, the cluster in pMFg across trial types is consistent with that seen in Figure 3.12b, suggesting either that this region codes for a planned vocal target, or otherwise has an activation pattern that

covaries with the vocal target, and is quickly refreshed when the required speech output changes. It is also worth noting the large cluster in dorsal postcentral gyrus that significantly predicts the vocal target vowel across all trials, although no functional significance to speech is known for this region.

These results, together, lead to the hypothesis that the superior pIFs/pMFg cluster is involved in the preparation or maintenance of a planned speech act, while vPMC, found in Chapter 2 to predict the identity of a repeated syllable at input, is here involved in the preparation of a new vocal target. The observation that vMC maintains a stronger representation of the auditory stimulus may indicate that, at this stage in the task, this area has not refreshed its representation to encode the vocal target, or that this region is engaged by perception (as described in the motor theory of speech perception; Liberman et al., 1967; Wilson et al., 2004; Pulvermüller et al., 2006); the absence of such a cluster at input (Figure 3.10) permits either interpretation.

3.4.4 Reliance on the dorsal stream for words and nonwords

The task described in this study was designed to contrast the treatment of word and nonword speech sounds during the required perception and production components of a repetition task in healthy adults, as well as dissociate the content of the perceived and produced speech sounds. Speech repetition is thought to rely on the dorsal stream, which provides a sensorimotor interface for translating between auditory and articulatory representations of speech sounds (via area Spt, which is proposed to link posterior temporal lobe areas and posterior frontal lobe areas; Hickok and Poeppel, 2004; Buchsbaum and D'Esposito, 2008; Buchsbaum et al., 2011). Additionally, the dorsal stream is obligatorily engaged by the repetition of nonword speech sounds, due to the lexico-semantic specificity of the ventral stream. This hypothesis is supported

by evidence showing preferential usage of dorsal stream areas in sublexical repetition tasks, while ventral stream areas are preferentially used in tasks requiring comprehension (*e.g.*, Saur et al., 2008). In the present study, univariate analysis (Figure 3.9) confirms an increased engagement of the posterior frontal areas considered to be part of the dorsal stream at the input portion of the trial when the task required repetition of a nonword rather than a word. Further, a multivariate contrast (Figure 3.10b-d) suggests that the phonological content – specifically, the vowel identity – of nonwords may be more prominently encoded in mid- and ventral premotor regions at the input portion of the task than for words. Whether nonwords are preferentially or differentially represented is unclear from these results, but as there is no theoretical reason (such as inter-stream competition) for words to be handled more slowly or less ably by the dorsal stream, this may be an effect of chunking, or the abstraction of common sequences of phonemes into their own phonological units. The increased activation of premotor cortex for nonwords may thus reflect a larger number of chunks needed to process a nonword. This increase in premotor engagement may also result from a greater dependency on phonological working memory for nonwords than words (Baddeley, 1992), which has been argued to be supported by speech perception and production areas (Jacquemot and Scott, 2006; Perrachione et al., 2017).

When a change cue is presented, it is unclear to what extent the dorsal stream should be expected to be engaged. On the one hand, the vocal target on change trials corresponds to a visually presented (non)word, and nonword reading may be spared in persons with conduction aphasia, in whom nonword repetition is profoundly impaired (*e.g.*, Jacquemot et al., 2007). However, delayed nonword production relies on phonological working memory, which may in turn rely on the dorsal stream (Jacquemot and Scott, 2006, also see below). In addition to vPMC and IFg, univariate contrasts show

left pSTs more active at the input and cue during change trials than repeat trials (Figure 3.8a,b). pSTs is considered to be the auditory phonological input to the dorsal stream (Hickok and Poeppel, 2007). Although Spt does not show enhanced activation for change trials (though note nearby activation clusters in left pSTg in Figure 3.8b), it does appear to contain information predictive of the vocal target (Figure 3.13c), along with vPMC (BA6). This series of regions (pSTs-Spt-vPMC) is consistent with dorsal stream models, which would suggest a role in preparing an articulatory representation of the unheard syllable from a phonological representation in pSTs (Hickok, 2012). Our overall hypothesis, based on these data, is that the dorsal stream is automatically engaged in translating the heard syllable to a frontal representation appropriate for speech motor output; when the change cue is received, additional circuitry is engaged to reactivate the phonological representation of the heard syllable in pSTs, and refresh the frontal representation in order to encode the other (non-heard) syllable.

At output, the dorsal stream would be expected to play a modulatory role, generating an auditory expectation for error detection based on frontal speech motor output representations (Guenther, 1994; Guenther et al., 2006; Hickok et al., 2011; Hickok, 2012). There was no expected difference in the overall response to words or nonwords, which was consistent with the results of our univariate analysis (Figure 3.9c). Interestingly, however, bilateral area Spt / parietal operculum shows a cluster with significantly better prediction of vowel identity in words over nonwords at output. This suggests the articulatory-auditory transformation is more consistent or more relied-upon for words than nonwords with the same vowel. A worthwhile follow-up analysis would be to compare the consistency of subjects' vocalizations between words and nonwords.

3.4.5 Phonological working memory

One goal of this task paradigm was to preempt, if possible, the automatic loading of a motor plan for the aurally presented syllable by providing a visually presented distractor syllable and a 50% probability that the task would require producing that distractor, rather than the heard syllable (which would always be the case in a simple repetition paradigm, such as that presented in Chapter 2). That is, two (non)words were visually presented simultaneously in order to instantiate two items in working memory; when one was presented aurally, it was "tagged", but not known to be the vocal target. However, it is clear from the observed increased neural demands, as well as the ability to decode the vowel heard from frontal structures, even based on the earliest response estimates following a change cue that the auditory stimulus was loaded in spite of this design, consistent with Correia et al. (2015), who showed activation of bilateral inferior frontal areas even in a passive listening task, finding clusters predictive of consonant features. Recent conceptions of phonological working memory (pWM), in contrast to the "phonological store" of Baddeley (1992), propose that pWM is supported directly by the processes and structures of speech perception and production systems (Jacquemot and Scott, 2006; Buchsbaum and D'Esposito, 2008; Majerus, 2013). From this perspective, if the two visually presented syllables were in phonological working memory, they were being cycled between an auditory phonological representation in posterior superior temporal cortex and an articulatory phonological representation in posterior inferior frontal cortex. Thus, while simultaneous visual presentation was intended to have an inhibitory effect, preventing either motor plan from being loaded, it may instead have had a reinforcing effect on the aurally presented syllable.

Majerus (2013) proposes that the phonological input/output buffers may be supplemented with a bilateral fronto-parietal network including BA6/9 frontally and BA7/40 parietally when multiple (non)words need to be maintained, with the left hemisphere network responsible for attentionally-mediated maintenance. In these regions, we see two interesting pieces of evidence. The first is the increased engagement of the intraparietal sulci in change trials over repeat trials at the time of the cue (Figure 3.8b), which is consistent with the use of this fronto-parietal network to select the target syllable from working memory. Majerus proposes that this maintenance control network sustains frontal and temporal representations, which may help explain the previously noted presence of pSTs activation and absence of Spt activation. The second piece of evidence is the recurring cluster(s) at the boundary of pIFs, pMFg and the precentral sulcus (*e.g.*, as observed in Figures 3.10a,c, 3.11a, 3.12a,b, and 3.13a-c). These clusters are seen at input, predictive of the auditory stimulus; at cue, predictive of the vocal target; and again at output, predictive of the vocal target. If these clusters can be considered a unit, despite some spatial inconsistency, they might be indicative of an abstract speech plan, which by default represents the heard speech sound, but may be quickly replaced with an alternative speech sound. However, they are inconsistent with an attentional working memory interpretation. Attentional processes are not expected to directly represent the objects of attention, and there is no task-imposed need for multiple-item working memory to remain engaged beyond the cue stage. Further work is required to establish the specific functions of these regions in speech and phonological working memory tasks.

3.4.6 A caveat on interpreting input- and cue-related responses

Recall that the input event (auditory presentation onset) occurs at $t = 2.25$s after trial start, and the cue event (task instruction cue presentation) occurs at $t = 3.25$s. In single trial estimation, distinguishing events 1s apart is statistically difficult. As is visible in Figure 3.6a and b, the main effect of task over control trials is highly similar in both datasets. Thus, it may be that some apparently input-related effects are processes initiated at cue, while some apparently cue-related effects are continuing responses to the input event. This is not, however, to say that no distinctions may be drawn. In particular, differences between repeat and change trials can only make sense after the cue, prior to which trials are indistinguishable to the subject.

3.4.7 Limitations and future directions

A limitation of the task design is the difficulty of directly comparing cue-related responses to those at input and output because they had to be modeled separately in order to obtain stable, robust event estimates. A potential modification of the task design would be to evenly space visual cues, as well as speech perception and production cues, *e.g.*, visual word presentation, auditory word presentation, repeat/change cue presentation, and GO signal presentation at even intervals, immediately following a volume acquisition. This preserves the relative quiet for speech perception and production and permits responses corresponding to reading and performing the cognitive switch to be factored out of the responses to input and output. Note, however, that this would decrease the number of scans per modeled event, and reduce the time available for subjects to store the vocal target without a distractor.

In this task, we observed that frontal areas appeared to encode a motor output plan for the syllable heard in all trials. In order to inhibit automatic preparation of a motor

plan, consider an otherwise identical paradigm in which the aurally presented syllable is not first presented visually may prove more effective; this would give the unheard syllable exclusive access to the speech production system prior to the preferentially treated auditory stimulus (again, see Correia et al., 2015), as well as eliminate any priming effects. Decreasing the probability of repeating the auditory stimulus may be an additional tool to decrease the utility of automatic loading.

In addition to dissociating the auditory stimulus from the vocal target to distinguish frontal responses to speech perception and planned production, it would be useful to distinguish, in temporal cortex, the auditory consequences of self-produced speech from an auditory target, for on-line error correction (Guenther et al., 2006; Hickok et al., 2011). One possible modification to this task is to (on some trials) use an auditory mask to interfere with subjects' self-perception.

Notwithstanding its limitations, this study provides a rich dataset that invites a number of questions not explored in this dissertation. In addition to testing classifiers on subsets of the training data, it is possible also to train classifiers on subsets of each dataset; one may then ask how well a classifier trained over repeat trials predicts vowels in change trials, or vice versa. Another question invited by these data is whether words and nonwords are treated differently under repeat and change contexts; an ANOVA adapted to MVPA classification accuracy would provide valuable insight into questions of dorsal stream access during change trials. Finally, the words in this study were counter-balanced by vowel and semantic category; classification analyses on semantic categories may provide insight into an alternative, non-phonological word representation.

Chapter 4

Modeling categorical and analog signals in fMRI datasets

4.1 Introduction

Multivariate pattern analysis is a relatively young class of analytical techniques, and interpretation of its results remains notoriously problematic (Anderson and Oates, 2010; Coutanche, 2013; Etzel et al., 2013). One obstacle to interpretation is the opacity of models constructed from the training data, such that determining specific "informative" voxels and what relation they have with the classification labels is non-trivial. This problem is compounded by searchlight techniques (Kriegeskorte et al., 2006; Oosterhof et al., 2011; Chen et al., 2011), which generate thousands of models to produce their results, and also by any attempt to compare or combine results across subjects, rendering model inspection an intractable avenue for resolving questions of interpretation.

A more fundamental issue is that of attribution. In this work, classification of vowel identity in a CVC syllable context is considered to be evidence of a potential phonological representation of vowels. However, a vowel may be straightforwardly represented as a discrete class or as a pair of formants (Peterson and Barney, 1952); linear classifiers would have little trouble matching a categorical class label to either underlying representation (Hillenbrand et al., 1995). One approach for resolving ambiguous classifications is simply to test alternative class labels (Naselaris and Kay, 2015). This treats the labeling itself as the interpretation, and the relative perfor-

mance of classification on different label sets provides a measure of plausibility.

This strategy is comparable to that employed by representational similarity analysis (RSA; Kriegeskorte et al., 2008; Nili et al., 2014), another multivariate method aimed at comparing models, and used in Evans and Davis (2015) to detect a regional preference for phonological representations of consonants. RSA entails computation of "representational dissimilarity matrices" (RDMs), stimulus-stimulus distance matrices encoding some measure of difference of response to every pair of stimuli. For example, in Evans and Davis (2015), at each searchlight ROI, a 36×36 RDM was constructed from correlation distances between mean responses to six syllables under six acoustic conditions. Model RDMs were binary similar-dissimilar indicators under different assumptions, *e.g.* one RDM indicated same/different syllable identity while another indicated same/different consonants. By correlating each searchlight RDM with each model RDM, the similarity of neural responses to simple correlative models may be compared to establish more likely representations in different regions. The RSA technique has also been used to compare BOLD responses to computational models, most notably comparing human inferior temporal responses to a battery of pictures to a large number of neuroscientifically motivated models of image processing (Kriegeskorte, 2009).

While these model selection methods are useful in interpreting measured data, little work has aimed to understand how different hypothetical neural signals give rise to BOLD patterns, and how specific classifiers perform on these patterns.

This chapter approaches the question from a signal-processing perspective: given a known representation, to what extent can measures of a generic (and ill-defined) notion of "information" detect it? This approach does not try to recover a representation, but seeks to characterize limits on detection for different representations.

4.1.1 neuRosim

`neuRosim` (Welvaert et al., 2011) is an fMRI simulation framework intended to standardize simulations, and facilitate the replication and comparison of simulation studies. In particular, `neuRosim` is designed to generate datasets that mimic the spatial and temporal properties of signals of interest and noise sources in fMRI data. An fMRI dataset is a time-series of three-dimensional volumes, or grids of voxels; the temporal profile of a modeled signal is convolved temporally with a hemodynamic response function and across voxels according to a spatial kernel. Following is a partial overview of the functions provided, their properties, and their use in this analysis.

4.1.1.1 Signal generation

`neuRosim` provides functions for defining the time course of signals and their spatial extents. Here we list the functions used in this study.

Spatial extent definition `neuRosim::simprepSpatial` defines the spatial extent of modeled signals. For manually or programmatically selected voxels, the needed parameters are the number of voxels (`regions`) and a list of voxel coordinates (`coord`).

Time course definition `neuRosim::simprepTemporal` allows the definition of multiple related time-courses, to correspond to the activation of multiple regions of interest, or to define a single time course for all regions. This function is parameterized by run length (`totaltime`), a list of event `onsets`, a corresponding list of `durations` (a single value, if no variation), and TR, all in seconds. `effectsize` is a list (or single value) of maximum heights of HRF responses, `hrf` selects the HRF model from a single gamma (Boynton et al., 1996), double gamma (Friston et al., 1998), or balloon (Buxton et al., 1998), and model sampling resolution in seconds

(accuracy).

Volume construction neuRosim::simVOLfmri accepts the output of simprepSpatial and simprepTemporal, as well as spatial (dim) and temporal (nscan) dimensions to produce 4D time-series. Repetition time (TR) must be specified, and noise may be added (see below).

4.1.1.2 Noise generation

Noise refers to both measurement error and signals of non-interest. Spurious correlations of noise with signals of interest can lead to systematic errors; hence, much effort has gone into characterizing prominent sources of noise in fMRI data, and neuRosim provides functions to account for measurement, physiological, and temporally and spatially auto-correlated noise.

Each noise generator used in this study is described below along with its R function name and an equation describing its behavior. All R functions take dim and nscan parameters, describing its dimensions and the number of volumes to generate; functions with temporal dependence take a TR parameter. Each equation describes the noise generator as a function of time and space – here only two dimensions (x and y) – and is parameterized with arguments exposed by neuRosim.

A parameter common to all noise generators is σ, or standard deviation, which in all cases is applied when generating the noise series; after generation and any resampling, the standard deviation of the series is calculated and adjusted to match σ, if necessary.

The right-hand side of each equation has a pseudorandom component and may may include additional structure. The pseudorandom component is denoted χ, and represents a number generated from the standard normal distribution ($\mathcal{N}(0, 1)$) at

each voxel, at each time point. The most common stochastic component, $\sigma\chi$, is equivalent to sampling from $\mathcal{N}(0, \sigma^2)$.

In this section, Δt is used instead of TR to indicate the time step for noise generation, while TR is reserved for the repetition time of an acquisition paradigm. This distinction reflects a distinction to be drawn in this particular simulation, in which data is generated at 2.5s intervals and then subsampled to a 5s TR.

White noise `neuRosim::systemnoise` represents measurement noise, fluctuations inherent to MRI acquisition. Noise in MRI is magnitude-dependent, following a Rician distribution, but for signal-to-noise ratios > 2, a Gaussian approximation suffices (Gudbjartsson and Patz, 1995).

$$\epsilon(t, x, y|\sigma) = \sigma\chi \tag{4.1}$$

Physiological noise `neuRosim::physnoise` models respiratory and cardiac signals, which are the dominant sources of noise generated by subjects' bodies; given the sampling rate (1/TR) of fMRI ($<$ 1Hz), aliasing can result (Biswal et al., 1996). These signals are modeled as sine-waves, with default frequencies of $f_{Resp} = 0.2$Hz and $f_{HR} = 1.17$Hz ($\sim$70 bpm). In this case, the `sigma` parameter applies to the standard deviation of the physiological signal, and the stochastic component has variance of 1.

$$\phi(t, x, y|f_{HR}, f_{Resp}, \sigma) \approx \chi + \sigma\left(\sin(2\pi f_{HR}t) + \cos(2\pi f_{Resp}t)\right) \tag{4.2}$$

$\phi(t, x, y)$ is scaled so that the standard deviation of the sum of sines equals σ, sampled at intervals of Δt. The above approximation reflects that, as $\Delta t \to 0$, the

sum of two independent sine waves has standard deviation of 1.

Note that the random variate contributes $1/(\sigma^2 + 1)$ of the variance in ϕ, and is negligible for large σ.

Temporal noise `neuRosim::temporalnoise` induces temporal autocorrelations, shown by Purdon and Weisskoff (1998) to result in misestimated false positive rates unless "whitened". Temporal noise reflects the nature of fMRI as a repeated measure, as well as modeling the effects of physiological events of non-interest on the BOLD signal. A simple autoregressive model ($AR(p)$) is used, with parameters $\rho_1, ...\rho_p$.

$$\tau(t, x, y|\vec{\rho}, \sigma) = \sigma\chi + \sum_{i=1}^{p} \rho_i \tau(t - i\Delta t, x, y|\vec{\rho}, \sigma) \tag{4.3}$$

Spatial noise `neuRosim::spatialnoise` simulates spatial dependencies in BOLD responses, which occur due to both the spatial spread of the hemodynamic response resulting from neural connectivity and shared vascular resources (Engel et al., 1997) and the limits of fMRI spatial resolution (Robson et al., 1997). Spatial noise may be generated using an $AR(1)$ autoregressive model with parameter ρ or a Gaussian or Gamma random field model. In this study we will consider only the 2-D autoregressive model:

$$\begin{aligned}\varsigma(t, x, y|\rho) = \sigma\chi + \sqrt{1 - \rho^2}(\varsigma(t, x - \Delta x, y|\rho) + \varsigma(t, x, y - \Delta y|\rho) + \\ \varsigma(t, x + \Delta x, y|\rho) + \varsigma(t, x, y + \Delta y|\rho))\end{aligned} \tag{4.4}$$

Task-related noise `neuRosim::tasknoise` adds additional Gaussian (or Rician) noise only when a voxel is considered active. This has the effect of increasing variance during the task, and the authors suggest that it may be interpreted as residual head

motion-related noise or spontaneous neural activity due to the task.

Mixture `neuRosim::simVOLfmri` permits a weighted combination of noise sources to be generated, given a baseline activation b, a signal-to-noise ratio R and a weight vector $\sum \vec{\omega} = 1$. Letting $\sigma = b/R$, a noise series:

$$N(t, x, y|\vec{\omega}, b, R) = \frac{\vec{\omega} \bullet (\epsilon(t|\sigma), \tau(t|\sigma), \phi(t|\sigma), \varsigma(t|\sigma))}{RSS(\vec{\omega})} \tag{4.5}$$

Task-related noise and low frequency drift (`neuRosim::lowfreqdrift`) are not shown for simplicity, but are accessible via this same interface.

Summary The noise models described above capture a range of sources of variance. Assigning these models realistic relative weights is non-trivial, and spatial and temporal dependencies are unlikely to be easily separated. Nonetheless, these sources of noise represent the major sources of systematic error that must be accounted for in standard preprocessing and deconvolution.

4.1.2 Approach

In this study, we simulated a perception-only version of the syllable repetition task from Chapter 2, and constructed artificial fMRI datasets containing representations of the perceived syllable. We defined a representation as a set of distinct receptive fields, or functions mapping a stimulus onto a response magnitude (for simplicity, ranging in $[0, 1]$), that voxels might be assigned. Using stimulus data from the study Chapter 2, we constructed discrete (all-or-nothing) representations of vowels and syllables, and a continuous representation based on the (F_1, F_2) formant frequency values of each stimulus.

By constructing noisy datasets with known underlying representations, we sought to characterize the performance of searchlight classification algorithms, parametrically by strength and density of information (which are defined precisely in Section 4.2). Classification analyses are performed using both vowel and syllable class labels, to test for a classification performance boost when the label matches the underlying representation (an assumption of Naselaris and Kay, 2015), as well as differential performance on discrete vs. continuous representations.

4.1.3 General hypotheses

We hypothesized that classification detects the presence of information, and not representation, or, put another way, there is no advantage to classification for an input representation that matches the chosen class label *so long as knowledge of the input representation implies knowledge of the class label.* Because the syllable contains the vowel, we hypothesized that vowel classification will be high in datasets constructed from vowel receptive fields and syllable receptive fields, alike. On the other hand, multiple syllables share a vowel, so we predicted that syllable classification will be high only in datasets constructed from syllable receptive fields.

Vowels are well-characterized by the first two formant frequencies (Peterson and Barney, 1952), making them ideal targets for comparing discrete and analog representations. We compared the performance of classification with vowel labels on both the discrete (vowel) and analog (formant) datasets. Although vowels cluster in the $F_1 \times F_2$ frequency space, no discriminant can perfectly divide the stimuli in Figure 2.7. We therefore hypothesized that, although a classification analysis that attempts to classify vowels on datasets constructed from formant-based receptive fields will succeed, such an analysis will be outperformed by classifying vowels on a dataset

constructed from vowel receptive fields. Specifically, we predicted lower classification accuracies on formant-based datasets than on vowel-based datasets, for any given density/amplitude combination.

4.2 Materials and methods

This simulation study sought to generate hypothetical speech sound representations in realistic fMRI noise and, using individual event estimation and searchlight classification analysis, assess the comparative performance of different representations on different classification labels. The simulated task was based on the syllable repetition task used in Chapter 2, with stimuli selected from the same set of recordings as in that study. The task itself was reduced to a single, auditory perception event in each trial, as no behavioral requirements were needed to ensure and verify subject engagement.

Datasets were constructed as a single slice, mimicking the two-dimensional geometry of the cortical sheet, and contained independently generated noise and signal components. Simulated signals were generated according to a set of receptive field models in which a given voxel may respond to phonological or auditory phonetic properties of the speech stimulus.

4.2.1 Modeled paradigm

A simulated session consists of 8 runs of 48 trials of 10s duration, with an auditory stimulus presentation at $t = 0$s in each trial. Two volumes are acquired per trial, with acquisition time (TA) of 2.5s and repetition time (TR) of 5s, starting at $t = 0$s. In contrast to the unevenly timed sparse paradigm of Chapter 2, a constant TR of 2×TA simplifies simulation and eliminates the need for scan-timing correction (see Section 2.2.6.1). Each volume is a single $64 \times 64 \times 1$ voxel slice.

Stimuli mimicked those used in Chapter 2: 18 CVC syllables were constructed from the consonants /m/, /t/ and /l/, and the vowels /ɪ/, /ɛ/ and /ʌ/ (Table 2.1). Two male and two female native English speakers recorded the stimuli, and five recordings of each syllable per speaker were collected, to allow for acoustic variation

among recordings of each syllable. Mean formant frequencies (see Figure 2.7) were extracted from approximate midpoints of vowels in the stimuli using custom PRAAT (Boersma and Weenink) scripts (You et al., 2015).

Each run consisted of 43 task trials, with syllable presentations, and 5 control trials, which were treated as containing no stimulus.

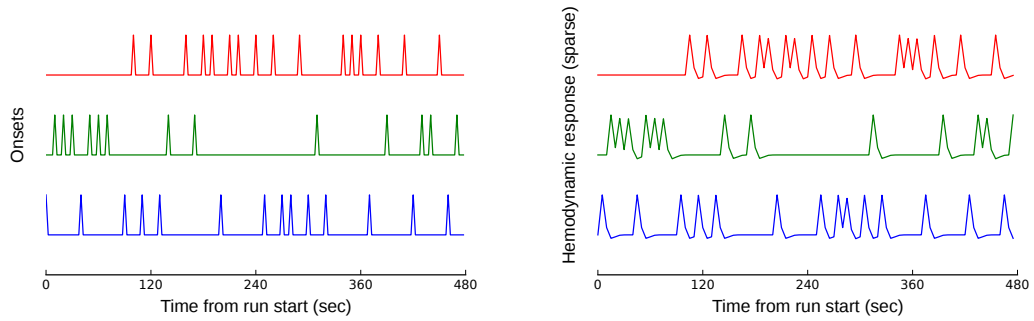
4.2.2 Receptive field types

Three classes of representations were considered: vowel, syllable, and formant, corresponding to possible underlying neural representations that subjects may use or activate during the task. The term receptive field (RF) is used here to refer to the strength of response of a voxel to each stimulus; each representation can be characterized by a set of RFs, and a responsive voxel is assigned one such RF.

Discrete: Vowel The simplest assumed representation of vowels is for a responsive unit to respond to the presentation of a specific vowel, and not to respond in its absence.

Figure 4.1 shows receptive field time series for a single run of 48 trials, constructed from the first run of subject S1 in Chapter 2. Figure 4.1a shows impulse responses corresponding to stimulus onset times for stimuli within each receptive field (blue: /ɛ/; green: /ɪ/; red: /ʌ/). Figure 4.1b shows the expected hemodynamic response to each stimulus series in a responsive voxel for each class, modeled at TR of 5s and TA of 2.5s.

Discrete: Syllable A syllable-responsive receptive field is one in which a responsive voxel responds to the presentation of an entire, specific syllable, and not at all to any other syllable, regardless of phonetic or phonological similarity. In the current case,



(a) Onset time series for three vowels (b) Sparse-sampled HRF series

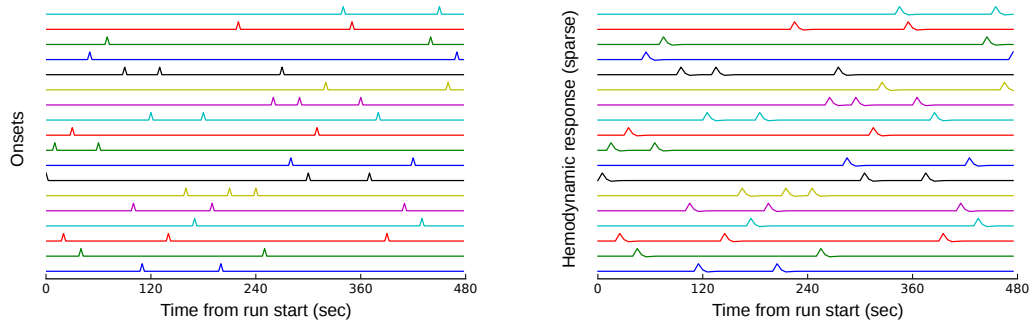
Figure 4.1: **Three category discrete vowel signal.** Time series for a single run of 48 trials, with voxels responsive to vowel identity. (a) Impulse responses of each receptive field to onset of a stimulus containing the preferred vowel identity. (b) Estimated hemodynamic response of corresponding responsive voxels. Each curve is an impulse response train, convolved with a canonical HRF (TA=2.5s), and down-sampled (TR=5s) to model a sparse acquisition paradigm.

we consider 18 syllables, 6 of which contain each vowel.

Figure 4.2 shows receptive field time series for a single run of 48 trials, constructed from the first run of subject S1 in Chapter 2. Figure 4.2a shows impulse responses corresponding to stimulus onset times for stimuli within each receptive field (*i.e.*, each syllable identity). Figure 4.2b shows the expected hemodynamic response to each stimulus series in a responsive voxel for each class, modeled at TR of 5s and TA of 2.5s.

Figure 4.2a shows stimulus onset times for each syllable identity, and Figure 4.2b shows the expected hemodynamic response to each stimulus series, modeled at TR of 5s and TA of 2.5s.

Acoustic: Formant Frequency-Based Receptive Fields A biologically plausible model of acoustic sensitivity is a receptive field with a preferred frequency band. To normalize for frequency discrimination thresholds, frequencies were defined in the



(a) Onset time series for eighteen syllables (b) Sparse-sampled HRF series

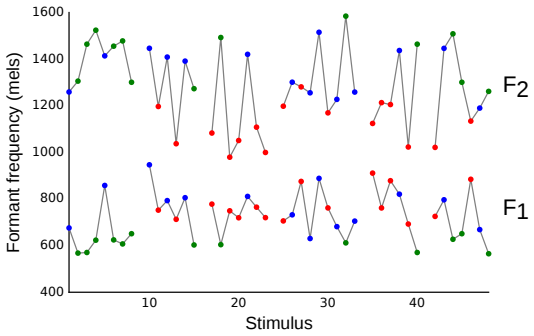
Figure 4.2: **Eighteen category discrete syllable signal.** Time series for a single run of 48 trials, with voxels responsive to syllable identity. (a) Impulse responses of each receptive field to onset of a stimulus containing the preferred syllable identity. (b) Estimated hemodynamic response of corresponding responsive voxels. Each curve is an impulse response train, convolved with a canonical HRF (TA=2.5s), and down-sampled (TR=5s) to model a sparse acquisition paradigm.

mel scale (Stevens et al., 1937). Each receptive field was constructed with a Gaussian fall-off to the distance between its preferred frequency and those of both formants present. No other spectral energy is assumed to activate these units; hence, this model assumes access to formant frequencies and suppression of less salient acoustic markers.

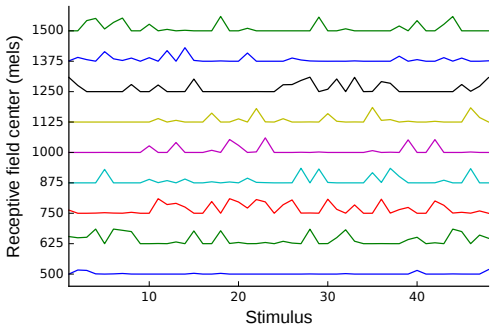
$$response = \exp\left(-\left(\frac{f_{pref} - F_1}{\sigma}\right)^2\right) + \exp\left(-\left(\frac{f_{pref} - F_2}{\sigma}\right)^2\right) \tag{4.6}$$

We considered a set of 9 receptive fields, evenly spaced between 500 and 1500 mels.

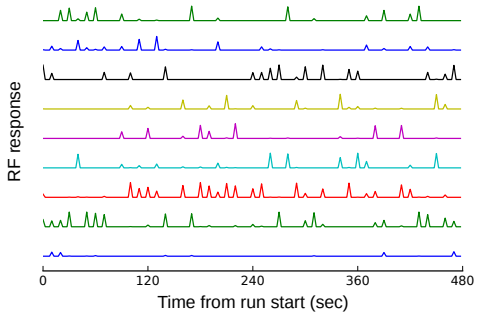
Figure 4.3(a,b) shows formant values and receptive field responses to each trial in a single run. Unlike in Figures 4.1 and 4.2, multiple receptive fields may be active in a given trial, and their response is proportional to the proximity of the two formants to their preferred frequency. Figure 4.3(c,d) shows the time series of these responses at stimulus onset (c) and convolved with a canonical HRF (d), sampled at TR=5s,



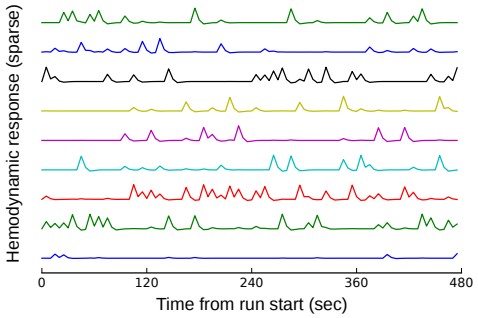
(a) Formant values (in mels) for each trial



(b) Receptive field response magnitudes



(c) Receptive field impulse responses



(d) Sparse-sampled HRF series

Figure 4.3: **Formant receptive fields.** (a) Estimated F_1 and F_2 frequencies are plotted for each trial (1-48) in a sample run. Missing values are for control trials, with no stimulus. (b) Response of each receptive field (f_{pref} plotted on y-axis; $\sigma =$ 60mel) at each trial. The height of each curve takes values in $[0, 1]$, according to Equation 4.6. (c) Impulse responses of each receptive field to the onset of each stimulus; the magnitude of responses are shown in (b). (d) Estimated hemodynamic response of corresponding responsive voxels. Each curve is an impulse response train, convolved with a canonical HRF (TA=2.5s), and down-sampled (TR=5s) to model a sparse acquisition paradigm.

TA=2.5s.

4.2.3 Implementation

Datasets were constructed as a single slice of 64×64 voxels, mimicking the two-dimensional geometry of the cortical sheet. Each voxel was assigned one (or no) receptive field. Sparse acquistion was simulated by generating data with a 2.5s TR and discarding the second and fourth volumes of each trial. Hence, a 48 trial run of 10s trials requires the generation of 4 volumes per trial, or 192 volumes per run, but only 96 volumes are used.

Dataset construction was performed using the R package `neuRosim` (Welvaert et al., 2011) version 0.2-12[1]. See Section 4.1.1 for a brief description of the relevant components.

4.2.3.1 Noise

A noise session was constructed from 8 independently generated runs, each using the following model:

$$\begin{aligned}\texttt{simVOLfmri}(&\texttt{base} = b,\ \texttt{SNR} = R,\ \texttt{weights} = \vec{\omega},\\ &\texttt{dim} = \texttt{c}(64, 64, 1),\ \texttt{nscan} = 192,\ \texttt{TR} = 2.5,\\ &\texttt{noise} = \texttt{"mixture"})\end{aligned}$$

10 sessions of noise were generated.

In addition to the free parameters, b, R and $\vec{\omega}$, the default and derived parameters used in the noise model are summarized in Table 4.1. The combined standard

[1]Minor bugfixes were submitted to the authors, and the fixed version used in this study. All described data may be generated in the unaltered version, however, if with slightly greater inconvenience.

deviation of all noise components $\sigma = b/R$. $\vec{\omega}$ is constrained to sum to 1.

Component	Parameter	Value/formula
`systemnoise`	`sigma`	$\omega_1\sigma$
`systemnoise`	`type`	`"gaussian"`
`temporalnoise`	`sigma`	$\omega_2\sigma$
`temporalnoise`	`rho`	(0.2)
`physnoise`	`sigma`	$\omega_4\sigma$
`physnoise`	`freq.heart`	1.17
`physnoise`	`freq.resp`	0.2
`spatial`	`sigma`	$\omega_6\sigma$
`spatial`	`method`	`"corr"`
`spatial`	`type`	`"gaussian"`
`spatial`	`rho`	0.75

Table 4.1: Parameters for generated noise datasets

Estimating free parameters Free parameters were estimated using the motion-corrected functional volumes for subjects in Project 1. Estimates were based on all time points at all voxels on the graymid surface (see Section 2.2.7). For a given subject, baseline activation was estimated as the mean value of graymid time points, and SNR was estimated as the mean divided by the standard deviation.

In the absence of clear data on relative strengths of noise components, we let $\vec{\omega} = \left(\frac{1}{4}, \frac{1}{4}, 0, \frac{1}{4}, 0, \frac{1}{4}\right)$.

4.2.3.2 Signals

Signal datasets were constructed from session metadata from the task in Chapter 2, with the proportion of simulated voxels responsive to *some* stimulus controlled by a density parameter $\delta \in [0, 1]$. Density captures a notion of redundancy, and may be thought of as a measure of how many voxels' responses can be distinguished across stimuli.

Voxels were chosen to respond to one of K receptive fields – $K = 3$ for vowels, 18 for syllables, 9 for formant receptive fields – with probability δ/K; that is, the probability that a voxel was responsive to some receptive field was δ, and the probability that a responsive voxel was responsive to a given receptive field was $1/K$.

To achieve these properties at each location (i, j), a variate X_{ij} was sampled from a uniform distribution and thresholded as follows to assign a response class k to voxel v_{ij}:

$$X_{ij} \sim U\left(1, 1 + \frac{K}{\delta}\right)$$

$$v_{ij} = \begin{cases} \lfloor X_{ij} \rfloor & \text{if } \lfloor X_{ij} \rfloor \leq K \\ 0 & \text{otherwise} \end{cases}$$

$U(a, b)$ indicates a uniform distribution with support $[a, b)$. Noting that $P(\lfloor X_{ij} \rfloor = k) = \frac{\delta}{K}$ for $k \in \{1..\lfloor K/\delta \rfloor\}$, the desired probabilities may be trivially derived:

$$\begin{aligned} P(v_{ij} > 0) &= \delta \\ P(v_{ij} = k | v_{ij} > 0) &= \frac{1}{K} \qquad\qquad \forall k \in \{1..K\} \end{aligned}$$

The procedures for simulating a single run of signals for discrete and continuous cases are given in Algorithm 4.1 and Algorithm 4.2, respectively.

Discrete signals In the discrete case, the number of (non-control) stimulus classes K was derived from a metadata file describing the stimulus sequence from one subject (S1) from Project 1. In all simulations, a discrete signal tracked either a vowel ($K = 3$) or a syllable ($K = 18$). A voxel is assumed to respond specifically to a single stimulus class, with an `effectsize` of 1% of baseline activation (of the noise sessions), and `SNR`

of 5, using `task-related` noise to provide spatial and temporal variation to voxels responding to the same stimulus.

Algorithm 4.1 Discrete signal generator (single run)

Require: `voxels` - Volume of voxel class sensitivities
Require: `stims` - Stimulus class sequence in $\{0..K\}$
Ensure: `effectsize` - 1% of noise baseline

```
vol ← 0
totaltime ← NTRIALS * TRIAL_LENGTH
for k ← {1..K} do
  onsets ← (which(stims == k) − 1) * TRIAL_LENGTH
  design ← simprepTemporal(totaltime, onsets = onsets, durations = 0.5,
                           effectsize = effectsize, TR = TA)
  image ← simprepSpatial(sum(voxels == k), listcoords(voxels == k),
                         form = "manual")
  vol ← vol + simVOLfmri(design = design, image = image,
                         nscan = ⌈NTRIALS * TRIAL_LENGTH/TA⌉, TR = TA,
                         noise = "task-related", SNR = 5,
                         dim = c(64, 64, 1))
end for
return vol
```

Continuous signals In the continuous case, all responsive voxels may respond to all stimuli, with the strength of response to each stimulus (`effectsize`) defining a preference, or receptive field. Here, K refers to the number of distinct receptive fields to be represented. Because a responsive voxel responded to every stimulus, additional `task-related` noise would simply be white noise, and was thus omitted.

4.2.3.3 Datasets

Datasets were constructed by selecting a noise dataset as a baseline and adding a signal dataset with an amplification factor indicating the peak response amplitude as a percentage of baseline activation.

Algorithm 4.2 Continuous signal generator (single run)

Require: `voxels` - Volume of voxel RF sensitivities
Require: `stims` - Stimulus class sequence
Require: `effectsizes` - RF responses to stimulus sequence
Ensure: K = Number of receptive fields (RFs)
Ensure: max(`effectsizes`) = 1% of noise baseline

```
vol ← 0
totaltime ← NTRIALS * TRIAL_LENGTH
onsets ← (which(stims ≠ 0) − 1) * TRIAL_LENGTH
for k ← {1..K} do
  design ← simprepTemporal(totaltime, onsets = onsets, durations = 0.5,
                           effectsize = effectsizes[k], TR = TA)
  image ← simprepSpatial(sum(voxels == k), listcoords(voxels == k),
                         form = "manual")
  vol ← vol + simVOLfmri(design = design, image = image,
                         nscan = ⌈NTRIALS * TRIAL_LENGTH/TA⌉, TR = TA,
                         dim = c(64, 64, 1))
end for
return vol
```

4.2.4 Analyses

After a session was simulated, it was treated as a series of motion-corrected runs, ready for event response estimation. Individual events were estimated using Nipype (Gorgolewski et al., 2016) and FSL (Smith et al., 2004; Jenkinson et al., 2012), in the same configuration as in Section 3.2.6, modified to fit a single event (stimulus onset) per 10s trial. Classification was performed using linear C-SVMs, as in Chapters 2 and 3, in a 3-voxel radius, volumetric searchlight configuration using a leave-one-run-out cross-validation scheme (Halchenko et al., 2015).

All datasets were classified with both vowel and syllable identities as labels. To allow for searchlight-induced edge effects, all statistics were taken from voxels at least three voxels from the edge of the image.

In order to establish null distributions of classification accuracy, searchlight analyses were first performed on noise datasets, with no signal added. Results from all

ten noise datasets were combined to establish global percentile thresholds.

For a given analysis, in addition to mean accuracy across voxels, we report the proportion of voxels exceeding the 99 percentile accuracy threshold. For analyses with a defined signal density, we also report the proportion of voxels exceeding the threshold, divided by the proportion of voxels with signal added.

A CVC syllable may be characterized by the identity of the whole syllable or those of its constituent phonemes, such as the vowel, both in its neural representation and in selecting class labels for searchlight analysis. We therefore assessed classification performance for both vowel and syllable identity on datasets constructed from both vowel and syllable information, independently, as well as datasets constructed from formant-based representations, in order to assess relative classification on discrete and continuous representations of vowel sounds.

4.3 Results

4.3.1 Parameter estimates

The `baseline` and `SNR` parameters were estimated from the data collected in Project 1, scan-timing corrected and motion corrected, to ensure estimates were taken from gray-matter voxels. To estimate `baseline`, the mean of middle-gray-matter voxels across all trials was calculated for each subject, and the `SNR` parameter by dividing each subject's mean by their standard deviation (see Section 4.2.3.1). Table 4.2 shows the mean and standard deviation across subjects, and the chosen values for the simulations presented here.

Parameter	Value	$\hat{\mu}$	$\hat{\sigma}$
`baseline`	700	694.15	83.48
`SNR`	2.9	2.90	0.17
Std. Dev.	241.38	240.23	31.39

Table 4.2: **Simulation parameters estimated from data.** The `baseline` and `SNR` parameters were estimated from collected subject data, and their distribution across subjects are described in $\hat{\mu}$ and $\hat{\sigma}$ columns. The "Value" column contains the values selected for simulation. The distribution of standard deviations is also included for reference. Its value value is simply `baseline`/`SNR`.

4.3.2 Chance classification accuracy

Performing classifications on noise sessions, with no added signal, permits the characterization of chance classification accuracy. Figure 4.4 shows cumulative distribution functions (CDFs) for classification accuracy rates across 10 separate noise distributions, classified on syllables (green) or vowels (blue). To account for edge effects induced by 3-voxel radius searchlights, only voxels at least three voxels from the edge of the image are counted. Median classification rates are at theoretical chance accuracies: $1/18 = 0.0\bar{5}$ for syllables and $1/3 = 0.\bar{3}$ for vowels. The 99th percentiles are

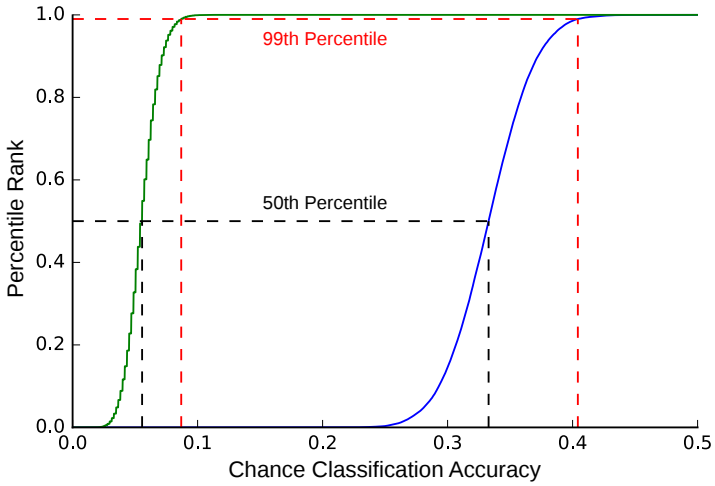
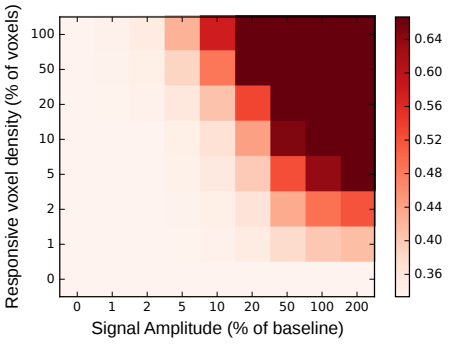


Figure 4.4: **Chance classification accuracy distributions.** Cumulative distribution functions for syllable and vowel classifications are shown in green and blue, respectively. Syllable classifications (1 in 18) have a median of 5.56% and a 99 percentile of 8.69%. Vowel classifications (1 in 3) have a median of 33.3% and a 99 percentile of 40.4%.

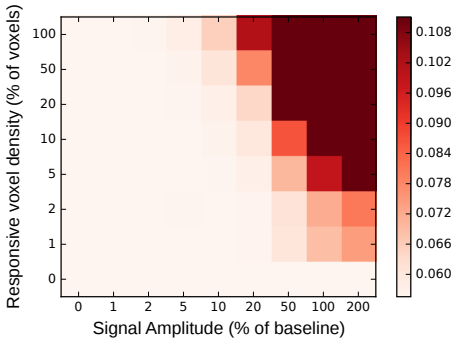
0.0869 and 0.404, respectively. From these values we can characterize an analysis by the percentage of voxels exceeding the chance 99th percentile.

Consider simulations for which vowel or syllable signals are introduced. Let the signal amplitude, the peak HRF value in generated signals, vary from 0% to 200% of the `baseline` activation (700; unitless), and let the density, the proportion of voxels responsive to some stimulus, vary from 0% to 100% of all voxels. Together, amplitude and density represent the availability of information for a classifier.

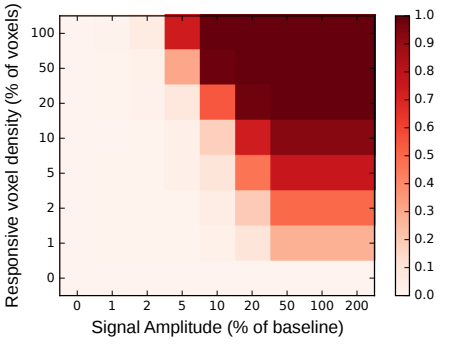
Figure 4.5 examines the behavior of classifiers constructed on congruent trial labels (*i.e.*, vowel labels for datasets with vowel signals added; syllable labels for datasets with syllable signals added). Sub-figures (a) and (b) show mean classification accuracies, scaled from chance (1/3 for vowels, 1/18 for syllables) to 2× chance. For vowels, the mean accuracy exceeds the 99th percentile chance accuracy (0.404) as density approaches 100% for low amplitudes (5% of baseline) and at any density $\geq$ 2% for



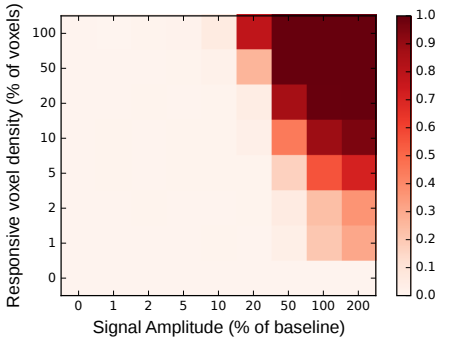
(a) Accuracy (vowel)



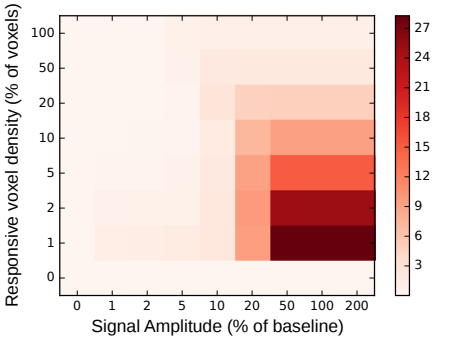
(b) Accuracy (syllable)



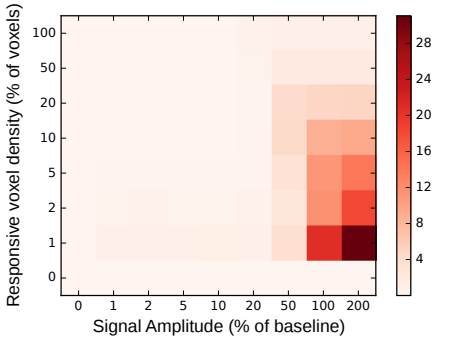
(c) Proportion voxels $>$ 99%-ile (vowel)



(d) Proportion voxels $>$ 99%-ile (syllable)



(e) Density-normalized (vowel)



(f) Density-normalized (syllable)

Figure 4.5: **Classification accuracy relative to chance.** Performance of searchlight classification of vowels and syllables on datasets containing vowel (left) and syllable-derived (right) signals. (a) and (b) show mean accuracy across all voxels, scaled between chance and 2$\times$ chance. (c) and (d) show the proportion of voxels exceeding the empirical 99-percentile chance classification accuracy. (e) and (f) show the proportion exceeding that value, normalized by the density, or proportion of voxels *containing signal*.

high amplitudes (50% of baseline). For syllables, the mean accuracy exceeds the 99th percentile chance accuracy (0.0869) as density approaches 100% for an amplitude of 20% of baseline, as density exceeds 10% for an amplitude of 50% of baseline, and for densities $\geq 5\%$ for amplitudes $\geq 100\%$.

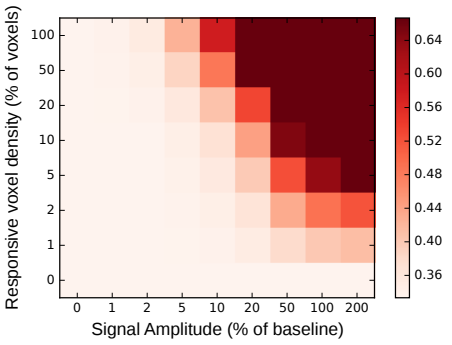
Sub-figures (c) and (d) show the proportion of voxels exceeding the 99th percentile of chance classification accuracies, following roughly similar patterns for 50% of voxels exceeding this strict threshold. Finally, sub-figures (e) and (f) normalize these proportions by density, giving a measure of what proportion of voxels exceed a chance threshold, relative to the proportion of voxels containing a signal. For vowels, this shows that, for signal amplitudes $\leq 10\%$, the proportion of super-threshold voxels roughly tracks the number of voxels containing signal. For syllables, this is true for signal amplitudes $\leq 20\%$.

4.3.3 Label / Sublabel Interchangeability

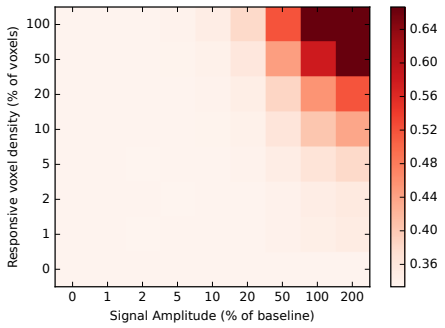
Next we consider the question of whether searchlight MVPA performance can help distinguish between different levels of categorical representation.

Figures 4.6 and 4.7 show the results of classifying datasets created using signals encoding vowel (left) and syllable (right) categories, with vowel (4.6) and syllable (4.7) class labels.

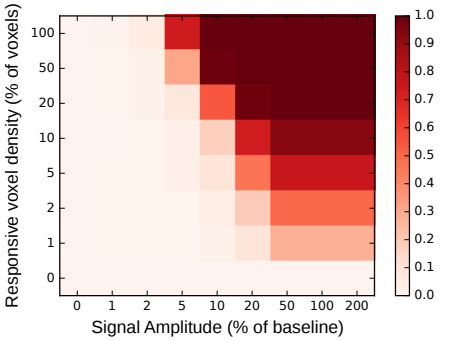
Figure 4.6 shows a strong effect of underlying signal, with classification gains in the syllable simulation requiring half of an order of magnitude higher amplitude, for a given density, or an order of magnitude higher density, for a given amplitude, than for simulations using a signal tracking vowel identity. This disparity in vowel classifications is most likely due to a difference in the number of voxels active on each trial, for a given density. In contrast, Figure 4.7 shows both simulations produce



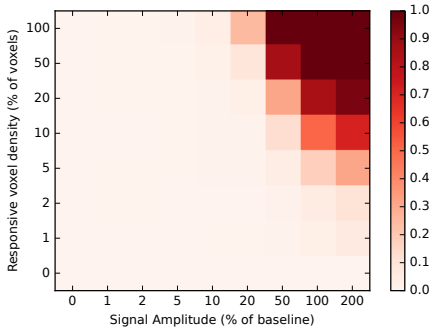
(a) Accuracy (vowel)



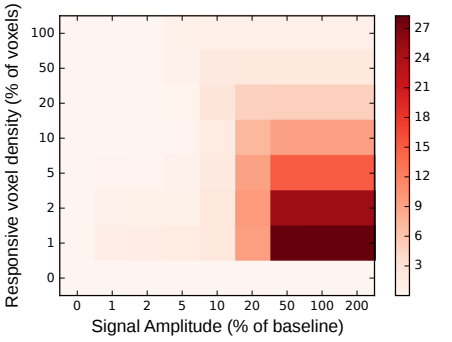
(b) Accuracy (syllable)



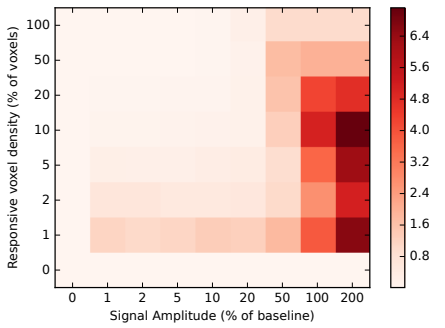
(c) Proportion voxels $>$ 99%-ile (vowel)



(d) Proportion voxels $>$ 99%-ile (syllable)

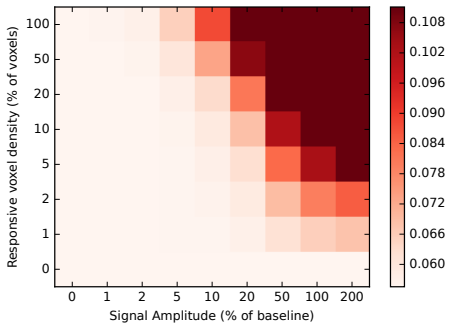


(e) Density-normalized (vowel)

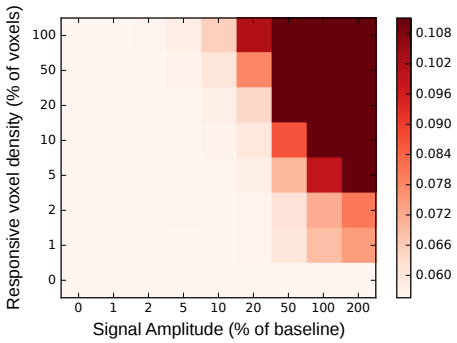


(f) Density-normalized (syllable)

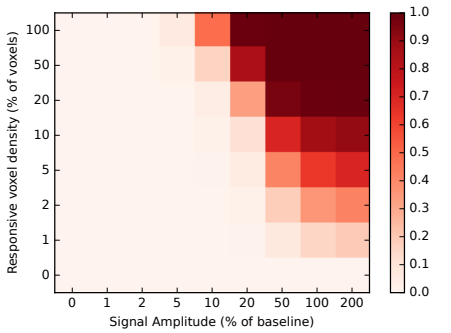
Figure 4.6: **Vowel classification accuracy on datasets with vowel- and syllable-related signals added.** Performance of searchlight-classification of vowels based on datasets with vowel (left) and syllable-derived (right) signals.



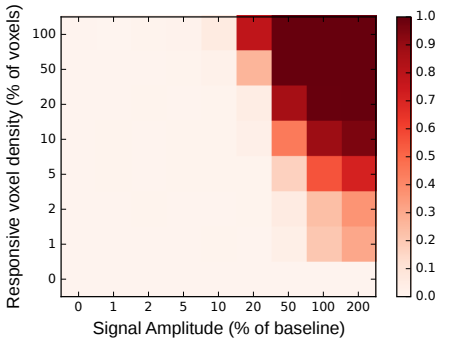
(a) Accuracy (vowel)



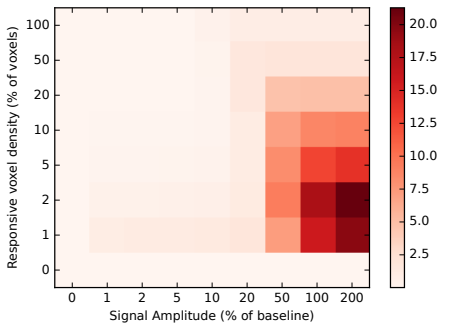
(b) Accuracy (syllable)



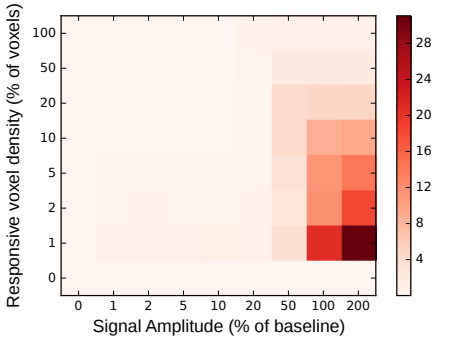
(c) Proportion voxels $>$ 99%-ile (vowel)



(d) Proportion voxels $>$ 99%-ile (syllable)



(e) Density-normalized (vowel)



(f) Density-normalized (syllable)

Figure 4.7: **Syllable classification accuracy on datasets with vowel- and syllable-related signals added.** Performance of searchlight-classification of syllables based on datasets with vowel (left) and syllable-derived (right) signals.

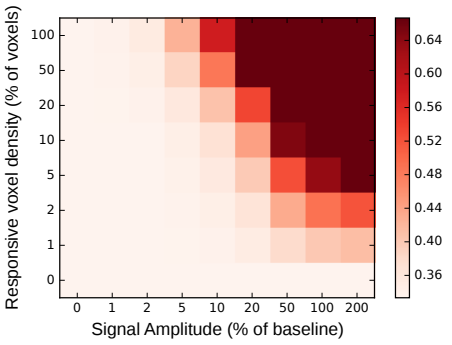
similar classification accuracy profiles when classifying on syllable identity, and both are, in turn, similar to classification of a vowel-based dataset on vowel identity. Thus, if classifying the same dataset on vowels and syllables produces similar maps, when normalizing by chance classification measures, then the underlying representation would appear to be vowel-level; if syllable classification is relatively improved over vowel classification, a syllable-level representation is more likely.

4.3.4 Discrete and Analog Representations

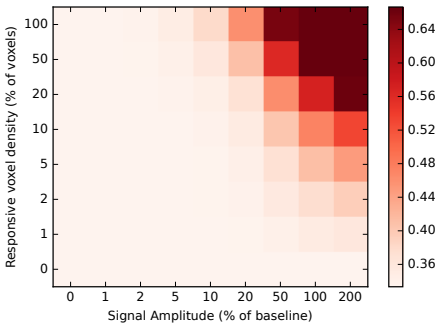
Finally, we consider relative classification performance of discrete and analog representations of vowels. As described above, the chosen analog representation is a series of receptive fields in perceptual frequency space, *i.e.*, the mel scale. A sensitive voxel responds proportional to the proximity of either formant to the preferred frequency of its receptive field.

Figure 4.8 compares the performance of searchlight classification of vowel identity in datasets constructed with a discrete vowel representation (left) and analog, formant-based representation (right). By all measures, the analog representation requires higher density or higher signal amplitude to achieve the same classification performance as the discrete representation.

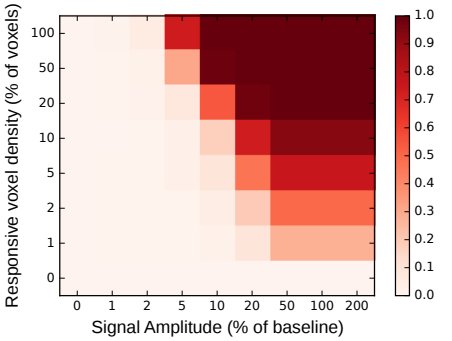
For completeness, Figure 4.9 compares the performance of searchlight classification of syllable identity in datasets constructed with a discrete syllable representation (left) and analog, formant-based representation (right). By all measures, the analog representation requires higher density or higher signal amplitude to achieve the same classification performance as the discrete representation.



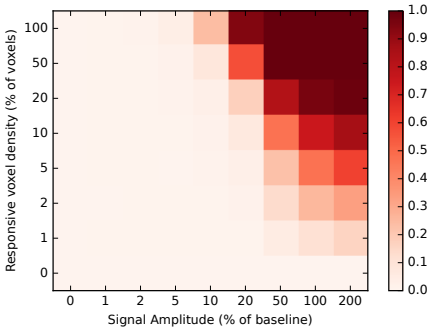
(a) Accuracy (vowel)



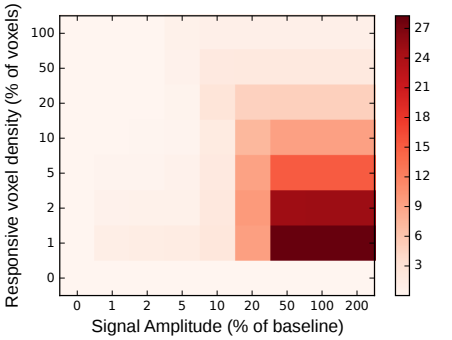
(b) Accuracy (formant)



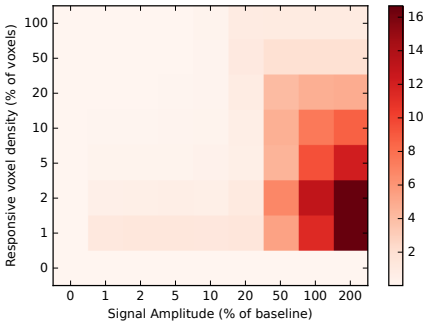
(c) Proportion voxels $>$ 99%-ile (vowel)



(d) Proportion voxels $>$ 99%-ile (formant)

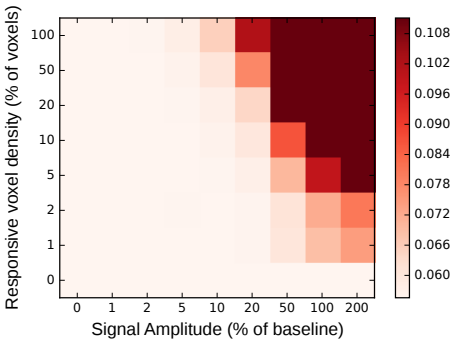


(e) Density-normalized (vowel)

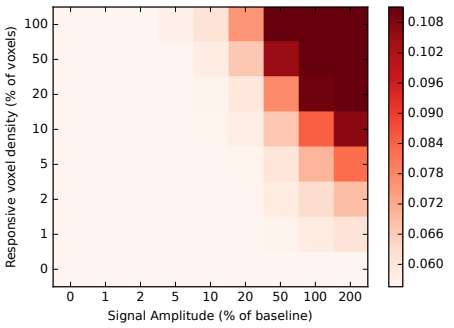


(f) Density-normalized (formant)

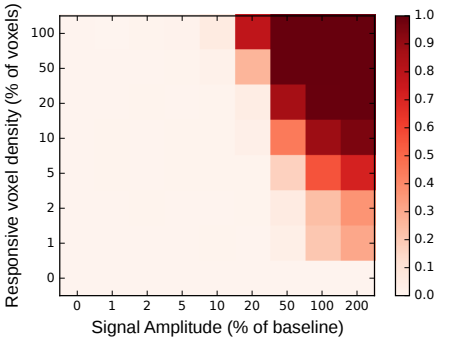
Figure 4.8: **Vowel classification accuracy on datasets with vowel- and formant-related signals added.** Performance of searchlight-classification of vowels based on datasets with vowel (left) and formant-derived (right) signals.



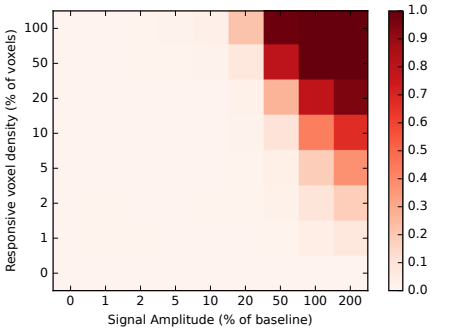
(a) Accuracy (syllable)



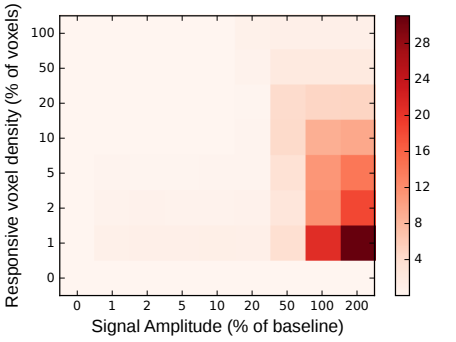
(b) Accuracy (formant)



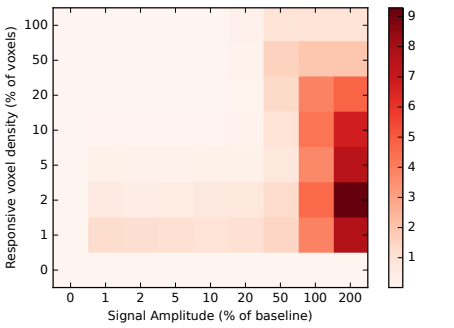
(c) Proportion voxels $>$ 99%-ile (syllable)



(d) Proportion voxels $>$ 99%-ile (formant)



(e) Density-normalized (syllable)



(f) Density-normalized (formant)

Figure 4.9: **Syllable classification accuracy on datasets with syllable- and formant-related signals added.** Performance of searchlight-classification of syllables based on datasets with syllables (left) and formant-derived (right) signals.

4.4 Discussion

This simulation study begins to address issues of representation and interpretation raised by the speech fMRI experiments described in Chapters 2 and 3. The simulated task was a perception-only version of the syllable-repetition task from Chapter 2, with trial length modified to easily simulate a constant TR sparse paradigm, and the simulated signals (*i.e.*, underlying neural codes for the stimuli) were distributed uniformly and randomly across a synthetic time series for a single image slice in order to characterize results by density (proportion of responsive voxels) and magnitude (relative strength of response compared to baseline, or mean gray-matter activation).

Three receptive field models were considered: a *vowel* receptive field had an all-or-nothing impulse response to the presentation of a syllable containing the preferred vowel; a *syllable* receptive field had an all-or-nothing impulse response to the presentation of the preferred syllable; and a *formant* receptive field had an impulse response whose magnitude was modulated by the proximity of the stimulus vowel formants (F_1 and F_2 only) to a preferred frequency. Following construction of artificial datasets parameterized by density and signal amplitude, responses to individual stimulus presentations were modeled using techniques that mirrored those used with actual data in Chapters 2 and 3, to be classified according to the identity either of the vowel or the syllable presented.

In all cases, we found that high classification accuracies ($\geq 2\times$ chance) could be achieved, regardless of the underlying receptive field representation, but the necessary densities and signal amplitudes required to achieve a constant level of accuracy varied across analyses. For datasets containing vowel or syllable receptive fields, classifying the vowel identity required markedly lower density and amplitude in the vowel-based dataset (Figure 4.6). On the other hand, classifying the syllable identity had com-

parable performance between the two representations, with somewhat improved performance on vowels (Figure 4.7). For datasets containing vowel or formant-based receptive fields, classifying the vowel identity again required lower density and amplitude in the vowel based dataset (Figure 4.8). Similarly, classifying the syllable identity on datasets containing syllable and formant-based RFs showed improved performance on the discrete, syllable representation over the analog, formant-based representation (Figure 4.9).

4.4.1 Summary of main hypotheses

Two specific hypotheses were proposed. The first was stated *classification detects the presence of information, and not representation*, and produced the expectation that vowel classification would be insensitive to whether the underlying receptive fields were tuned to vowel or syllable identity, while syllable classification would perform well only on datasets constructed from syllable receptive fields. In fact we observed the opposite trend, with syllable classification proving relatively insensitive to representational bases. Thus, if one assumes that, in a region of interest, a subset of voxels may be responding categorically to either a vowel or syllable identity, a significant result for both vowel and syllable classifications could be indicative of a vowel representation, while a syllable-only classification more likely reflects an underlying syllable representation. Rather than outright disconfirming the hypothesis, though, this pattern prompts a more nuanced notion of information, in terms of density and conditional probability. Density was defined as a parameter (δ) for constructing datasets (Section 4.2.3.2) and refers to the proportion of voxels that respond to some stimulus; for discrete all-or-nothing RFs considered here, the proportion of voxels active on any given trial was δ/K, where K was the number of distinct RFs. Normal-

izing with respect to this proportion, instead of δ, would likely result in the originally proposed performance patterns, though whether this is a more realistic density criterion is non-obvious. Additionally, in the formulation of the hypothesis, "information" was implicitly used in the sense of a Boolean conditional: knowledge of the syllable identity implies knowledge of the vowel identity. However, by Bayes' rule it is trivially shown that likelihood of correctly guessing the syllable is 1/6 the likelihood of correctly guessing the vowel. It was thus incorrect to pose one condition (in which the underlying representation was vowel-based) as lacking information about syllable identity. In particular, classification performance supported by partial knowledge must be considered in the terms of "relative accuracy" that MVPA classification results are cast in. We have shown our results from chance to 2× chance because, in our experimental studies, results below chance are not of interest, and positive results typically do not approach the maximum possible value 1. The low chance accuracy for a high K classification problem means that even partial information is sufficient to produce a significant positive result. Thus, by revising the notion of "information", this hypothesis stands with the following result: a syllable may be classified above chance on the basis of its constituent phonemes alone. From the perspective of natural language, the number of possible syllables grows combinatorially with the number of phonemes in a language. The disparity in density of discrete representations (in the sense of the all-or-nothing receptive fields proposed here) of syllables and phonemes in a human brain would be correspondingly more stark; given that phoneme-level representations can support syllable classification, it appears more likely that syllable classification reflects phoneme-level representations.

The second hypothesis was that vowel classification would be better supported by datasets constructed from discrete vowel receptive fields than by datasets constructed

from continuous, formant-based receptive fields. In the case of this specific analog representation, the hypothesis was clearly confirmed: high classification rates for vowel identity are achievable based on a formant-based receptive field representation, but require substantially higher density or signal amplitude, in comparison with a phoneme-based representation, to achieve. In general, we expect this pattern to hold for "phonological" (abstract /categorical) vs. phonetic (continuous) representations of phonemes.

These preliminary results demonstrate the utility of this model-based approach in assessing the effects of assumed discrete and analog representations on the behavior of searchlight classification.

4.4.2 Future directions

Beyond exploring the interpretive questions posed by the analyses in this dissertation, the aim of this work is to begin to address outstanding questions in the field of speech neuroscience with regard to the representational units of speech sounds. Contemporary models of speech production and processing suggest separate representations of speech in different brain areas, at various linguistic levels. For example, Hickok (2012) suggests different areas may represent speech sounds at syllabic (BA44 for motor syllable programs, STg/STs for auditory syllable targets) and somato-phonemic (*i.e.*, vocal tract constriction; BA6 for motor phoneme programs, aSMg for somatic phoneme targets) levels. Bohland et al. (2010) suggest a phonemic representation for planning a forthcoming speech sequence in left IFs and a syllabic representation in left ventral premotor cortex. Similar proposals are found in models of speech perception, where there is evidence, for example, that the lexicon can be accessed via an abstract, phoneme-based representation or via a more acoustic, less categorical representation

(Hickok and Poeppel, 2007). Recent electrocorticography studies have demonstrated evidence for distributed, feature-based representation of phonemes in STg (Mesgarani et al., 2014) and articulatory-motor representations in ventral motor/somatosensory cortex (Bouchard et al., 2013; Conant et al., 2014; Arsenault and Buchsbaum, 2015). In this analysis, we considered a formant frequency-based representation of vowels. Because formant frequencies covary with vowel identity, a hypothetical formant-based representation would represent an abstracted, phonetic version of the acoustics for a vowel, but not a fully categorical model. Such an intermediate representation is often assumed in neurocomputational models of aspects of speech perception (*e.g.*, Guenther and Gjaja, 1996) and production (Guenther et al., 2006; Kröger et al., 2009). Functional MRI, although a coarse measure of neural activity, is an irreplaceable tool for studying the neural processing underlying speech in healthy subjects. In order to understand how fMRI and MVPA can best make contact with theoretical models, it is essential to characterize the behavior of these analysis techniques, given "ground truth" simulated neural signals that match hypothetical representations.

The comparison of computational models and neuroimaging data was also an explicit design goal in representational similarity analysis (RSA Kriegeskorte et al., 2008; Kriegeskorte and Kievit, 2013; Nili et al., 2014). The central idea in RSA is that representations are reflected in a notion of "distance", or dissimilarity between objects to be represented. The more similar a model's pattern of dissimilarity (encoded in a representational dissimilarity matrix; RDM) is to the response pattern observed for a set of voxels, the more those voxels are deemed likely to have a representational basis that mirrors that found in the model. This approach was used by Evans and Davis (2015) to identify regions that respond more or less consistently to phonological or acoustic properties of CV stimuli. While this is a valuable approach, classification-

based MVPA remains a popular analysis technique, and it is necessary to enhance our ability to interpret such opaque measures as classification accuracy. Further, RSA does not use a simulated BOLD signal, instead assuming the dissimilarity between conditions should directly reflect differences observed in BOLD patterns. Beyond the MVPA techniques analyzed here, artificial datasets can be subjected to new analyses – including RSA – to gauge their appropriateness for detecting or distinguishing different representations.

4.4.3 Limitations and improvements

A full analysis of the analytic methods used in this dissertation is beyond the scope of this initial study. This simulation considers only a single subject, under multiple noise conditions, with voxel activation patterns that are unchanged across analyses. To more directly address the summary statistics presented in Chapters 2 and 3, multiple subjects need to be simulated, with some degree of functional variation, and the analysis pipeline will need to extend to group-level analyses, including voxel- and cluster-level thresholding. In practice, this will require spatially-varying representations (*i.e.*, regions of greater or lesser density of responsive voxels), which will allow the spatial variation of functional representations across subjects, and the construction of data-driven, voxel-wise thresholds.

In addition to a more complete exploration of the analytical techniques used in the analysis of actual fMRI data, the simulation model leaves considerable room for more realistic dynamics and alternative signal implementations.

Two distinct, but related, representational considerations omitted from this analysis are the effects of multiple sensitivity and signal spread. It is unrealistic that single voxels, containing many thousands of neurons, would respond collectively to

one stimulus attribute or another, or that a neural sub-population that does respond to a single stimulus would exhibit hemodynamic effects confined to a single voxel. A neural population that responds to speech sounds is much more likely to appear at the voxel level as a graded response to different speech sounds than as an all-or-nothing response to subsets of speech sounds. Such a graded response may be induced by signal spread alone, at a sufficiently high density of responsive voxels; while this may be a useful technique for generating sensitivity to multiple stimulus classes, it may also be worth considering separately signal spread due to tissue connectivity and the vascular/measurement spread (which has here been modeled as spatially-correlated noise).

The representations modeled here are extremely simple, but have very limited connection to existing proposals of neural representations of speech sounds. A number of theoretical representations were discussed in Section 4.4.2, the implementation and analysis of which would be a valuable contribution in itself. However, simpler acoustic models are also worth considering, because lower-level representations necessarily contain the basis for deriving higher-level representations, and our analysis of a formant-based representation showed this is sufficient to drive classification of phonemes, if less strongly than a pure categorical phoneme-based representation. Tonotopic maps have been observed in primary auditory cortex (*e.g.*, Talavage et al., 2003); extending this technique to generate and analyze tonotopic or somatotopic maps to determine whether or not they would provide a basis for classifying speech at different levels of analysis would be a worthwhile exploration of the acoustic and motoric end-points of the speech processing systems.

In both the dataset construction and analysis presented here, a central metric is the density of voxels containing some simulated receptive field. Although a useful

parameter in the creation of datasets, it does not have a clear interpretation in neural data. As noted when discussing signal spread, if a voxel is responsive to speech sounds, it is more likely to have a graded response to different sounds, and *all* voxels in a responsive region are likely to have some response, even if not preferential to any given stimulus. As this framework is expanded to incorporate signal spread, a more relevant parameter may be the number of contributory signals to a given voxel, reflecting a *sub-voxel* density of responsive neural populations. Even in this limited, preliminary framework, density is not the only, nor perhaps the best, axis for making comparisons between receptive field models. To illustrate the issue, consider as an alternative the density of voxels responsive to a given stimulus: By this measure, for parity (*i.e.*, so that the same number of units respond to a given stimulus) with vowel RFs – in which 1 in 3 RFs responds to each stimulus – syllable RFs ought to be six times more dense than currently constructed. For an analog model such as the formant RFs proposed here, if one assumes two receptive fields respond to each formant, then four of nine receptive fields are active in any trial and thus should be made 75% as dense to achieve parity with vowel RFs. Other metrics may be devised, and care must be taken to understand the assumptions being imposed.

As a final point, it is worth considering the limitations imposed by `neuRosim`. The mutual independence of spatial and temporal noise is difficult to justify; a noise signal arising from neurophysiological processes should have both a spatial and temporal component, whereas neuRosim adds two independently generated signals. Spatial convolution of temporal noise with a point-spread function, or temporal convolution of spatial noise with a hemodynamic response function may prove to be more realistic noise models. Additionally, the signal generation is decidedly coarse, intended to simulate studies that will use traditional univariate analyses. These facilities could

be expanded to include the generation of local regions with stimulus-correlated spatial means or patterns of variation (see Coutanche, 2013).

In summary, the work presented here constitutes the first steps in a larger exploration of multivariate pattern analysis. We have shown that simple models of vowel, syllable, and formant-based receptive fields can support successful classification in a searchlight analysis, and that these models result in differential performance, permitting comparisons between hypothetical representations. At present, the model makes no provision for spatial inhomogeneity, a necessary precondition for topographic maps. Multi-subject, group level analyses are also not fully developed, preventing an analysis of "clusters of predictive information", the standard reporting unit of MVPA studies. Additionally, it is important to carefully consider the evaluation criteria used to compare the different models and analyses; in this preliminary effort, we focused on the density and amplitude of responsive voxels, but these metrics assume that density and amplitude are independent considerations from model specification. As work progresses, these assumptions may need to be revisited. Further work will permit more detailed and realistic models of speech sounds to be assessed and compared, contributing to our ability to interpret MVPA results in light of theoretical models of speech perception and production.

Chapter 5

Conclusion

The repetition of a novel word requires a person to translate an acoustic pattern into a series of articulatory gestures, a process that relies on a series of neural representations. The speaker must abstract linguistically-relevant features from their specific acoustic context, maintain a working memory representation, and generate a speech motor plan, before fluently producing the target output sequence. This dissertation describes efforts to systematically map the neural correlates of perception and production during speech repetition.

The fMRI study discussed in Chapter 2 sought to capture BOLD responses to speech items at distinct input and output stages of a simple syllable repetition task, and to use multivariate analysis to discover informational correlates of the repeated speech sounds. Our approach utilized a fast, event-related design, which allowed us to separately estimate input- and output-related BOLD responses. These individual trial response estimates allowed for the localization of clusters of cortical vertices that, across subjects, predicted the sounds heard or produced at above chance levels. Multivariate pattern analyses revealed informational correlates of speech items predominantly in areas traditionally associated with the speech network (superior temporal, inferior frontal, pre-SMA, motor and somatosensory cortices). Of particular interest, the perceived vowel (the most acoustically salient sound available on a trial) could be predicted in responses linked to the input portion of the task at above chance levels in the left inferior frontal sulcus, an area suggested previously to code for individual planned phonemes (Bohland and Guenther, 2006; Bohland et al.,

2010). Likewise, the produced vowel could be detected from output-related responses in bilateral posterior STs, which indicate a target sound representation for production that is activated following the stimulus input, or auditory input of the speaker's own voice used for speech monitoring. Efforts to disentangle phonological (categorical) representations from phonetic (continuous) representations were limited in their utility, failing to reveal any significant dependence of classifiers on responses linearly related to formant frequencies.

Building on these results, Chapter 3 describes a further fMRI experiment designed to dissociate the representations of speech sounds involved in auditory syllable perception from those involved in preparing and overtly producing a syllable, as well as to systematically compare the responses to and representations of words and nonwords in a novel speech repetition task. In this repetition task, subjects were informed of the syllable to be spoken after they listened to an aurally presented syllable, isolating the processes of auditory syllable perception, preparation to produce a syllable, and overt production of a selected syllable. In half of trials, the auditory stimulus was to be repeated, while, in the remaining trials, subjects instead were instructed to speak a (previously) visually-presented syllable. In addition to confirming a greater engagement of so-called dorsal stream areas in perception and motor planning for nonwords than words (Saur et al., 2008), our results suggested that subjects automatically generated motor plans upon listening to the auditory stimulus, regardless of which syllable was to be spoken, relying on a posterior temporal and inferior frontal network (*i.e.*, the dorsal stream) to refresh the speech motor plan, when cued to produce the unheard syllable. Multivariate analysis showed early representations of the auditory stimulus in inferior frontal sulcus, superior temporal sulcus and even motor cortex, while the vocal target (once known to the subject) was consistently represented near

the junction of IFs, the precentral sulcus, and middle frontal gyrus. Interestingly, although trials with words and nonwords did not show differences in overall activation levels during overt production, multivariate analyses indicate differences in representations in bilateral Spt (word preference), posterior superior temporal gyrus (nonword preference), and left insula, which revealed an inferior/superior division in preference for words and nonwords, respectively.

These two studies yielded a pattern of results and interpretation that are consistent in broad terms, but with many specific discrepancies. In Chapter 2, the large IFs cluster found at the *input* phase of the task was taken to indicate that subjects construct a speech motor plan immediately on perception, and this conclusion was corroborated and elaborated in Chapter 3, demonstrating that this process is difficult (potentially impossible, in a strict interpretation based on the Motor Theory of speech perception) to preempt and generates a decodable representation in left ventral motor cortex. This finding in vMC is one apparent discrepancy: while consonants and syllables could be detected in ventral motor and somatosensory cortices in the first study, vowel information was not found near these regions. Similarly, the IFs cluster itself became drastically smaller in the second study, while the large, bilateral pSTs clusters observed at output entirely failed to reproduce in the second study. Instead, we saw relatively strong evidence that the pSTs represented the vowel in the syllable that was heard in both the input and cue datasets. Based on the evidence that the vocalized syllable was planned for production at the start of the delay period, the output portion of the task can be expected to engage similar processes, particularly in trials in which the auditory stimulus was repeated. Thus, this difference in pSTs might instead be an effect of stimulus choice, which differed, for example, in the balance of words and nonwords across vowels (Okada and Hickok, 2006a). Alternatively, the

absence of a delay in the second task between the previous scan and the production cue (compared to 0.5-1.5s delay in the first) may altered the response to the auditory processing of self-produced speech, resulting in reduced ability to decode the vowel in pSTs at output.

Together, through multivariate pattern analysis of multiple events within individual trials, these studies have begun to clarify the within-trial dynamics of speech repetition, a seemingly simple task that appears to rely on multiple distinct representations linked through a series of complex neural pathways. They have also demonstrated the utility of classification of phonological content in tracking representations of syllables in the different stages of the task. As these techniques mature, a natural extension of this task is the repetition and/or construction of syllable sequences to further clarify the preparation phase of the task and expose ordering information to multivariate analysis. Consider, for example, a stimulus sequence of *visual syllable – auditory syllable – ordering cue – production cue*, where the subject produces two syllables, and the ordering cue indicates whether one of the syllables is to be repeated twice, or one before the other. Building on the multivariate contrast technique of the second study, it should be possible to distinguish between representations that reflect motor plan contents and ordering, which have been proposed to have anatomically distinct representations by Majerus (2013).

Additionally, it is clear that the phoneme itself is not (always) the neural representation being tracked, and room remains for further exploration of the effect of stimulus choices and their interactions with theoretical underlying neural representations on identified patterns. Chapter 4 approached this problem from the other direction, developing a framework for comparing the behavior of MVPA techniques on hypothetical representations of stimuli. Here, we considered searchlight classifi-

cation over simulated datasets containing discrete and continuous representations of the stimuli from Chapter 2. These datasets were constructed based on a simplified variation (listening) of the task from Chapter 2, and the simulated representations were based on acoustic parameters of stimulus sequences from the original task. Representations were defined in terms of simple, theoretically motivated receptive fields for a voxel, that were sensitive either to vowel identity, syllable identity, or formant frequency. Following construction of artificial datasets, parameterized by density and strength (amplitude) of receptive fields (responsive voxels), responses to individual trials were modeled using standard fMRI processing tools, to be classified according to the identity either of the vowel or the syllable presented. It was shown that accurate classification of vowels was possible on any of the given representations, but required a higher density and/or amplitude in order to achieve similar levels of performance on syllable- or formant-based representations than on vowel-based representations. Surprisingly, classification of syllables showed little difference in performance over datasets constructed with vowel- or syllable-based representations, demonstrating that incomplete information (*i.e.*, the knowledge of the syllable implies knowledge of the vowel, but not vice versa) is sufficient to boost classification rates significantly above chance.

Chapter 4 demonstrated the viability of the method using simple, idealized representations, and confirms that the detection of representation (or information) is not equivalent to the identification of the underlying neural representation. One of the most important extensions to this research will be to work on constructing and analyzing representations hypothesized in the literature, including acoustic and articulatory feature-based representations of phonemes (Bouchard et al., 2013; Mesgarani et al., 2014), tono- and somatotopic maps (Talavage et al., 2003; Takai et al., 2010; Conant

et al., 2014), and phoneme-position maps (Bohland et al., 2010). Self-organizing maps are popular models of vowel categorization effects (*e.g. Guenther and Gjaja, 1996; Kröger et al., 2009)*; translations of these models into voxel response patterns, along with the specific formant-based input representations they presuppose, would allow us to assess the plausibility of these representations, or identify the analysis technique most likely to find such representations, if they exist.

In conclusion, this work has built on a theoretical framework that understands speech perception and production as processes that are deeply intertwined. Using variations on delayed speech repetition tasks to separate these processes temporally, and multivariate pattern analysis to identify neural correlates of stimulus phonology, we have identified patterns of responses that are consistent with a largely dorsal-stream mediated process of speech plan preparation. Finally, we introduced a simulation framework for assessing performance of MVPA techniques on simulated fMRI datasets containing hypothetical representations, which will aid in the design and interpretation of MVPA-based studies to evaluate theoretical models of speech processes.

Appendix A

Experiment 2 Protocol

Following is the protocol followed prior to the fMRI session.

A.1 Protocol checklist

Consent and safety

1. Provide subject with:
    - 2239E fMRI consent form
    - BU MRI Safety form
    - Edinburgh Handedness Inventory
    - If subject is female, provide a pregnancy test strip
2. Quickly summarize the consent form, noting that the subject can withdraw from the study (*i.e.*, stop the scanning session) at any time though we would like them to try to complete 8 functional runs if possible. Note that our contact information is provided, and that they can have a copy of the consent form. Ask them if they have any questions.
3. If there are any MRI safety concerns, talk with Andy
4. Ask the subject to remove any metal, and offer a locker for belongings (inside the restroom)

5. While one researcher is going over the protocol with the (next) subject, another should work in the control room

Subject training

1. Read over the resting state instructions (see Appendix A.2), and ask if the subject has any questions.

2. Then, tell the subject they will get a demo of the task, and read out the instructions, making sure the subject seems to be following along.

3. Provide the demo, and "guide the subject" through the first couple trials, especially noting that he/she should speak immediately after the cross turns orange. Also note that the orange cross will immediately appear at the end of one scan, so it can be anticipated somewhat. Be sure to also include an example of and discussion of the control trials in the demo.

Equipment / protocol setup

- Generate stimulus files (.npy) for subject(s) and add to Git repository
- Plug VGA cable to projector into presentation laptop
- Plug USB from scanner (for apostrophe triggers) into presentation laptop
- Make sure presentation laptop has power
- Turn on earphone amplifier
- Attach earphones and canal tips

- Set up to record optical microphone signal using Audacity (device should default to USB microphone, but this can be changed); input volume should be set to max.
- Make sure fiber optic mic box is switched on
- Check audio levels for earphones (investigator should do first check of the day, then check each subject before starting scan)
- Check microphone levels in Audacity

A.2 Subject instructions

Resting-State Instructions

During this scan, which will last a little over 9 minutes, we would like you to relax with your eyes open and comfortably focused on the cross in the middle of the screen. It is important that you do not fall asleep, but you will not have to perform any tasks, and we ask that you just let thoughts wander.

Task Instructions

In this task, you will be asked to read, listen to, and say some syllables out loud. Some of the syllables will be words and some will be non-words. In order to accurately perform the task, you will need to pay close attention to the display screen and to sounds presented over your earphones, so we ask for you to do your best to remain attentive throughout. In the scanner, we will give you short breaks to rest between runs.

On every trial, the task will proceed as follows:

1. You'll see two syllables spelled out on the screen, which will remain visible for only a short time. We would like you to remember both of those syllables.

2. When the syllables disappear, a cross will appear, which you should keep your eyes focused on.

3. You will then hear one of the syllables that you just read over the headphones.

4. After you hear the syllable, a green or yellow rectangle will appear, which will provide instructions for what you will do next. If the green rectangle appears, your task will be to repeat the syllable that you heard, and you can now forget about the other syllable. If the yellow rectangle appears, your task will be to speak the syllable that you read but did not hear, and you can now forget about the syllable that you heard.

5. You are not to speak immediately after the rectangle cue. Instead, you need to remember the syllable you are going to produce until the cross changes color to orange. When this happens, you are to speak out loud the syllable as clearly as possible without moving your head. You should start speaking as soon as you can after the cross changes color.

6. To help minimize movements, you should always return your mouth to a comfortable position with lips closed and tongue relaxed after speaking and remain that way until the next trial.

7. On some trials, the sound you will hear will not be a syllable, but rather a noise that sounds somewhat like static. On trials when you hear the static noise, you are not to speak either syllable, but rather to press a button on the button box with your right index finger when the cross changes to orange.

8. While you are performing this task, you will hear regular, intermittent scanner noise. As you hear this noise, try to keep as still as possible. If the scanner noise starts while you are speaking, you should stop and return to the resting mouth position. If this happens frequently during any run, please notify us between runs.

9. Each of the syllables will contain one of three vowels – "a" as in "hat", "i" as in "hit", or "u" as in "hut." You should do your best to always pronounce them this way when you speak. The consonants contained in the syllables should have obvious pronunciations, with the possible exception of "th," which we would like you to produce as in "thin" rather than "then."

Bibliography

Abrams, D.A., Ryali, S., Chen, T., Balaban, E., Levitin, D.J., Menon, V., 2013. Multivariate activation and connectivity patterns discriminate speech intelligibility in Wernicke's, Broca's, and Geschwind's areas. Cerebral Cortex 23, 1703–1714. doi:10.1093/cercor/bhs165.

Acheson, D.J., MacDonald, M.C., 2009. Twisting tongues and memories: Explorations of the relationship between language production and verbal working memory. Journal of Memory and Language 60, 329–350. doi:10.1016/j.jml.2008.12.002.

Ackermann, H., Riecker, A., 2004. The contribution of the insula to motor aspects of speech production: A review and a hypothesis. Brain and Language 89, 320–328. doi:10.1016/S0093-934X(03)00347-X.

Adank, P., 2012. The neural bases of difficult speech comprehension and speech production: Two Activation Likelihood Estimation (ALE) meta-analyses. Brain and Language 122, 42–54. doi:10.1016/j.bandl.2012.04.014.

Alario, F.X., Chainay, H., Lehericy, S., Cohen, L., 2006. The role of the supplementary motor area (SMA) in word production. Brain Research 1076, 129–143. doi:10.1016/j.brainres.2005.11.104.

Anderson, M.L., Oates, T., 2010. A critique of multi-voxel pattern analysis, in: Ohlsson, S., Catrambone, R. (Eds.), Proceedings of the 32nd Annual Meeting of the Cognitive Science Society, Portland, Oregon, USA. pp. 1511–1516.

Arsenault, J.S., Buchsbaum, B.R., 2015. Distributed Neural Representations of Phonological Features during Speech Perception. The Journal of Neuroscience 35, 634–642. doi:10.1523/JNEUROSCI.2454-14.2015.

Awh, E., Jonides, J., Smith, E.E., Schumacher, E.H., Koeppe, R.A., Katz, S., 1996. Dissociation of Storage and Rehearsal in Verbal Working Memory: Evidence From Positron Emission Tomography. doi:10.1111/j.1467-9280.1996.tb00662.x.

Baayen, R.H., Piepenbrock, R., Rijn, v., 1993. The CELEX lexical data base on CD-ROM. Linguistic Data Consortium, Philadelphia, PA.

Baddeley, A., 1992. Working memory. Science 255, 556–559. doi:10.1126/science.1736359.

Baldo, J.V., Katseff, S., Dronkers, N.F., 2012. Brain Regions Underlying Repetition and Auditory-Verbal Short-term Memory Deficits in Aphasia: Evidence from

Voxel-based Lesion Symptom Mapping. Aphasiology 26, 338–354. doi:10.1080/02687038.2011.602391.

Baldo, J.V., Wilkins, D.P., Ogar, J., Willock, S., Dronkers, N.F., 2011. Role of the precentral gyrus of the insula in complex articulation. Cortex 47, 800–807. doi:10.1016/j.cortex.2010.07.001.

Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. Cerebral Cortex 19, 2767–2796. doi:10.1093/cercor/bhp055.

Binder, J.R., Frost, J.A., Hammeke, T.A., Rao, S.M., Cox, R.W., 1996. Function of the left planum temporale in auditory and linguisitic production. Brain 119, 1239–1247.

Binder, J.R., Medler, D.A., Desai, R.H., Conant, L.L., Liebenthal, E., 2005. Some neurophysiological constraints on models of word naming. NeuroImage 27, 677–693. doi:10.1016/j.neuroimage.2005.04.029.

Birn, R.M., Bandettini, P.A., Cox, R.W., Jesmanowicz, A., Shaker, R., 1998. Magnetic field changes in the human brain due to swallowing or speaking. Magnetic Resonance in Medicine 40, 55–60. doi:10.1002/mrm.1910400108.

Biswal, B., Deyoe, E.A., Hyde, J.S., 1996. Reduction of physiological fluctuations in fMRI using digital filters. Magnetic Resonance in Medicine 35, 107–113. doi:10.1002/mrm.1910350114.

Boersma, P., Weenink, D., . Praat: doing phonetics by computer [Computer program]. Version 5.3.16, retrieved May 12, 2012, from http://www.praat.org/.

Bohland, J.W., Bullock, D., Guenther, F.H., 2010. Neural representations and mechanisms for the performance of simple speech sequences. Journal of Cognitive Neuroscience 22, 1504–1529. doi:10.1162/jocn.2009.21306.

Bohland, J.W., Guenther, F.H., 2006. An fMRI investigation of syllable sequence production. NeuroImage 32, 821–841. doi:10.1016/j.neuroimage.2006.04.173.

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., Formisano, E., 2014. Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. The Journal of Neuroscience 34, 4548–57. doi:10.1523/JNEUROSCI.4339-13.2014.

Bouchard, K.E., Mesgarani, N., Johnson, K., Chang, E.F., 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495, 327–332. doi:10.1038/nature11911, arXiv:15334406.

Boynton, G.M., Engel, S.A., Glover, G.H., Heeger, D.J., 1996. Linear systems analysis of functional magnetic resonance imaging in human V1. The Journal of Neuroscience 16, 4207–21.

Britton, B., Blumstein, S.E., Myers, E.B., Grindrod, C., 2009. The role of spectral and durational properties on hemispheric asymmetries in vowel perception. Neuropsychologia 47, 1096–1106. doi:10.1016/j.neuropsychologia.2008.12.033.

Buchsbaum, B.R., Baldo, J.V., Okada, K., Berman, K.F., Dronkers, N., D'Esposito, M., Hickok, G., 2011. Conduction aphasia, sensory-motor integration, and phonological short-term memory – An aggregate analysis of lesion and fMRI data. Brain and Language 119, 119–128. doi:10.1016/j.bandl.2010.12.001.

Buchsbaum, B.R., D'Esposito, M., 2008. The Search for the Phonological Store: From Loop to Convolution. Journal of Cognitive Neuroscience 20, 762–778. doi:10.1162/jocn.2008.20501.

Buxton, R.B., Wong, E.C., Frank, L.R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. Magnetic Resonance in Medicine 39, 855–864. doi:10.1002/mrm.1910390602.

Carreiras, M., Mechelli, A., Price, C.J., 2006. Effect of word and syllable frequency on activation during lexical decision and reading aloud. Human Brain Mapping 27, 963–972. doi:10.1002/hbm.20236.

Carreiras, M., Riba, J., Vergara, M., Heldmann, M., Münte, T.F., 2009. Syllable congruency and word frequency effects on brain activation. Human Brain Mapping 30, 3079–3088. doi:10.1002/hbm.20730.

Caspers, S., Geyer, S., Schleicher, A., Mohlberg, H., Amunts, K., Zilles, K., 2006. The human inferior parietal cortex: Cytoarchitectonic parcellation and interindividual variability. NeuroImage 33, 430–448. doi:10.1016/j.neuroimage.2006.06.054.

Catani, M., Jones, D.K., Ffytche, D.H., 2005. Perisylvian language networks of the human brain. Annals of Neurology 57, 8–16. doi:10.1002/ana.20319.

Celsis, P., Boulanouar, K., Doyon, B., Ranjeva, J.P., Berry, I., Nespoulous, J.L., Chollet, F., 1999. Differential fMRI responses in the left posterior superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. NeuroImage 9, 135–144. doi:10.1006/nimg.1998.0389.

Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M., Knight, R.T., 2010. Categorical speech representation in human superior temporal gyrus. Nature Neuroscience 13, 1428–1432. doi:10.1038/nn.2641.

Chen, Y., Namburi, P., Elliott, L.T., Heinzle, J., Soon, C.S., Chee, M.W.L., Haynes, J.D., 2011. Cortical surface-based searchlight decoding. NeuroImage 56, 582–592. doi:10.1016/j.neuroimage.2010.07.035.

Cogan, G.B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., Pesaran, B., 2014. Sensory-motor transformations for speech occur bilaterally. Nature 507, 94–98. doi:10.1038/nature12935.

Conant, D., Bouchard, K.E., Chang, E.F., 2014. Speech map in the human ventral sensory-motor cortex. Current Opinion in Neurobiology 24, 63–67. doi:10.1016/j.conb.2013.08.015.

Corfield, D.R., Murphy, K., Josephs, O., Fink, G.R., Frackowiak, R.S.J., Guz, A., Adams, L., Turner, R., 1999. Cortical and subcortical control of tongue movement in humans: a functional neuroimaging study using fMRI. Journal of Applied Physiology 86, 1468–1477.

Correia, J.M., Jansma, B.M.B., Bonte, M., 2015. Decoding Articulatory Features from fMRI Responses in Dorsal Speech Regions. The Journal of Neuroscience 35, 15015–15025. doi:10.1523/JNEUROSCI.0977-15.2015.

Coutanche, M.N., 2013. Distinguishing multi-voxel patterns and mean activation: Why, how, and what does it tell us? Cognitive, Affective, & Behavioral Neuroscience 13, 667–673. doi:10.3758/s13415-013-0186-2.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical Surface-Based Analysis. NeuroImage 194, 179–194.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., Fadiga, L., 2009. The Motor Somatotopy of Speech Perception. Current Biology 19, 381–385. doi:10.1016/j.cub.2009.01.017.

Davis, T., LaRocque, K.F., Mumford, J.A., Norman, K.A., Wagner, A.D., Poldrack, R.A., 2014. What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. NeuroImage 97, 271–283. doi:10.1016/j.neuroimage.2014.04.037.

Davis, T., Poldrack, R.A., 2013. Measuring neural representations with fMRI: Practices and pitfalls. Annals of the New York Academy of Sciences 1296, 108–134. doi:10.1111/nyas.12156.

Dell, G.S., Burger, L.K., Svec, W.R., 1997. Language production and serial order: a functional analysis and a model. Psychological Review 104, 123–47. doi:10.1037/0033-295X.104.1.123.

Deschamps, I., Tremblay, P., 2014. Sequencing at the syllabic and supra-syllabic levels during speech perception: an fMRI study. Frontiers in Human Neuroscience 8, 1–14. doi:10.3389/fnhum.2014.00492.

Desikan R. S. et al., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980. doi:10.1016/j.neuroimage.2006.01.021.

Diehl, R.L., Lotto, A.J., Holt, L.L., 2004. Speech Perception. Annual Review of Psychology 55, 149–179. doi:10.1146/annurev.psych.55.090902.142028.

Dijkstra, E.W., 1959. A note on two problems in connexion with graphs. Numerische Mathematik 1, 269–271. doi:10.1007/BF01386390.

Dronkers, N.F., 1996. A new brain region for coordinating speech articulation. Nature 384, 159–161. doi:10.1038/384159a0.

Du, Y., Buchsbaum, B.R., Grady, C.L., Alain, C., 2014. Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proceedings of the National Academy of Sciences of the United States of America 111, 7126–31. doi:10.1073/pnas.1318738111.

Eden, G.F., Joseph, J.E., Brown, H.E., Brown, C.P., Zeffiro, T.A., 1999. Utilizing hemodynamic delay and dispersion to detect fMRI signal change without auditory interference: The behavior interleaved gradients technique. Magnetic Resonance in Medicine 41, 13–20. doi:10.1002/(SICI)1522-2594(199901)41:1<13::AID-MRM4>3.0.CO;2-T.

Edmister, W.B., Talavage, T.M., Ledden, P.J., Weisskoff, R.M., 1999. Improved auditory cortex imaging using clustered volume acquisitions. Human Brain Mapping 7, 89–97. doi:10.1002/(SICI)1097-0193(1999)7:2<89::AID-HBM2>3.0.CO;2-N.

Engel, S.A., Glover, G.H., Wandell, B.A., 1997. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cerebral Cortex 7, 181–192. doi:10.1093/cercor/7.2.181.

Etzel, J.A., Zacks, J.M., Braver, T.S., 2013. Searchlight analysis: Promise, pitfalls, and potential. NeuroImage 78, 261–269. doi:10.1016/j.neuroimage.2013.03.041.

Evans, S., Davis, M.H., 2015. Hierarchical Organization of Auditory and Motor Representations in Speech Perception: Evidence from Searchlight Similarity Analysis. Cerebral Cortex 25, 4772–88. doi:10.1093/cercor/bhv136.

Fadiga, L., Craighero, L., 2003. New insights on sensorimotor integration: From hand action to speech perception. Brain and Cognition 53, 514–524. doi:10.1016/S0278-2626(03)00212-4.

Fiez, J.A., Raife, E.A., Balota, D.A., Schwarz, J.P., Raichle, M.E., Petersen, S.E., 1996. A positron emission tomography study of the short-term maintenance of verbal information. The Journal of Neuroscience 16, 808–822.

Fischl, B., 2012. FreeSurfer. NeuroImage 62, 774–781. doi:10.1016/j.neuroimage.2012.01.021.

Flinker, A., Chang, E.F., Kirsch, H.E., Barbaro, N.M., Crone, N.E., Knight, R.T., 2010. Single-trial speech suppression of auditory cortex activity in humans. The Journal of Neuroscience 30, 16643–16650. doi:10.1523/JNEUROSCI.1809-10.2010.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322, 970–973. doi:10.1126/science.1164318.

Friederici, A.D., Meyer, M., von Cramon, D.Y., 2000a. Auditory Language Comprehension: An Event-Related fMRI Study on the Processing of Syntactic and Lexical Information. Brain and Language 74, 289–300. doi:10.1006/brln.2000.2313.

Friederici, A.D., Wang, Y., Herrmann, C.S., Maess, B., Oertel, U., 2000b. Localization of early syntactic processes in frontal and temporal cortical areas: A magnetoencephalographic study. Human Brain Mapping 11, 1–11. doi:10.1002/1097-0193(200009)11:1<1::AID-HBM10>3.0.CO;2-B.

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R., 1998. Event-Related fMRI: Characterizing Differential Responses. NeuroImage 7, 30–40. doi:10.1006/nimg.1997.0306.

Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S.J., Turner, R., 1996. Movement-related effects in fMRI time-series. Magnetic Resonance in Medicine 35, 346–355. doi:10.1002/mrm.1910350312.

Gathercole, S.E., 1995. Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. Memory & Cognition 23, 83–94. doi:10.3758/BF03210559.

Ghaziri J. et al., 2015. The Corticocortical Structural Connectivity of the Human Insula. Cerebral Cortex , 1–13doi:10.1093/cercor/bhv308.

Goldrick, M., Rapp, B., 2007. Lexical and post-lexical phonological representations in spoken production. Cognition 102, 219–260. doi:10.1016/j.cognition.2005.12.010.

Gorgolewski K. J. et al., 2016. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. 0.12.0-rc1. doi:10.5281/zenodo.50186.

Gracco, V.L., Tremblay, P., Pike, B., 2005. Imaging speech production using fMRI. NeuroImage 26, 294–301. doi:10.1016/j.neuroimage.2005.01.033.

Graves, W.W., Grabowski, T.J., Mehta, S., Gupta, P., 2008. The left posterior superior temporal gyrus participates specifically in accessing lexical phonology. Journal of Cognitive Neuroscience 20, 1698–1710. doi:10.1162/jocn.2008.20113.

Gudbjartsson, H., Patz, S., 1995. The Rician distribution of noisy MRI data. Magnetic Resonance in Medicine 34, 910–914. doi:10.1002/mrm.1910340618, arXiv:NIHMS150003.

Guenther, F.H., 1994. A neural network model of speech acquisition and motor equivalent speech production. Biological Cybernetics 72, 43–53. doi:10.1007/BF00206237.

Guenther, F.H., Ghosh, S.S., Tourville, J.A., 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. Brain and Language 96, 280–301. doi:10.1016/j.bandl.2005.06.001.

Guenther, F.H., Gjaja, M.N., 1996. The perceptual magnet effect as an emergent property of neural map formation. The Journal of the Acoustical Society of America 100, 1111–1121. doi:10.1121/1.416296.

Halberstam, B., Raphael, L.J., 2004. Vowel normalization: the role of fundamental frequency and upper formants. Journal of Phonetics 32, 423–434. doi:10.1016/j.wocn.2004.03.001.

Halchenko Y. et al., 2015. PyMVPA: 2.4.1. doi:10.5281/zenodo.33988.

Hall, D.A., Haggard, M.P., Akeroyd, M.A., Palmer, A.R., Summerfield, A.Q., Elliott, M.R., Gurney, E.M., Bowtell, R.W., 1999. 'Sparse' temporal sampling in auditory fMRI. Human Brain Mapping 7, 213–223. doi:10.1002/(SICI)1097-0193(1999)7:3<213::AID-HBM5>3.0.CO;2-N.

Hanke, M., Halchenko, Y.O., Sederberg, P.B., Hanson, S.J., Haxby, J.V., Pollmann, S., 2009. PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. Neuroinformatics 7, 37–53. doi:10.1007/s12021-008-9041-y.

Hanley, J.R., Dell, G., 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, 147–158. doi:10.1080/02643290342000339.

Hartley, T., Houghton, G., 1996. A linguistically constrained model of short-term memory for nonwords. Journal of Memory and Language 35, 1–31. doi:10.1006/jmla.1996.0001.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430. doi:10.1126/science.1063736.

Hester, R., D'Esposito, M., Cole, M.W., Garavan, H., 2007. Neural mechanisms for response selection: comparing selection of responses and items from working memory. NeuroImage 34, 446–454. doi:10.1016/j.neuroimage.2006.08.001.

Hickok, G., 2009. The functional neuroanatomy of language. Physics of Life Reviews 6, 121–143. doi:10.1016/j.plrev.2009.06.001.

Hickok, G., 2010. The role of mirror neurons in speech perception and action word semantics. Language and Cognitive Processes 25, 749–776. doi:10.1080/01690961003595572.

Hickok, G., Erhard, P., Kassubek, J., Helms-Tillery, A.K., Naeve-Velguth, S., Strupp, J.P., Strick, P.L., 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. Neuroscience letters 287, 156–160.

Hickok, G., Houde, J., Rong, F., 2011. Sensorimotor integration in speech processing: computational basis and neural organization. Neuron 69, 407–22. doi:10.1016/j.neuron.2011.01.019.

Hickok, G., Poeppel, D., 2000. Towards a functional neuroanatomy of speech perception. Trends in Cognitive Sciences 4, 131–138. doi:10740277.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99. doi:10.1016/j.cognition.2003.10.011.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8, 393–402. doi:10.1038/nrn2113.

Hickok, G.S., 2012. Computational neuroanatomy of speech production. Nature Reviews Neuroscience 13, 135–145. doi:10.1038/nrn3158.

Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American English vowels. The Journal of the Acoustical Society of America 97, 3099–111.

Hinrichs, H., Scholz, M., Tempelmann, C., Woldorff, M.G., Dale, A.M., Heinze, H.J., 2000. Deconvolution of Event-Related fMRI Responses in Fast-Rate Experimental Designs: Tracking Amplitude Variations. Journal of Cognitive Neuroscience 12, 76–89. doi:10.1162/089892900564082.

Houde, J.F., Nagarajan, S.S., Sekihara, K., Merzenich, M.M., 2002. Modulation of the auditory cortex during speech: an MEG study. Journal of Cognitive Neuroscience 14, 1125–1138. doi:10.1162/089892902760807140.

Hynd, G.W., Hynd, C.R., 1984. Dyslexia : Neuroanatomical / neurolinguistic perspectives. Reading Research Quarterly 19, 482–498.

Iacoboni, M., 2008. The role of premotor cortex in speech perception: Evidence from fMRI and rTMS. Journal of Physiology-Paris 102, 31–34. doi:10.1016/j.jphysparis.2008.03.003.

Indefrey, P., Levelt, W.J.M., 2004. The spatial and temporal signatures of word production components. Cognition 92, 101–144. doi:10.1016/j.cognition.2002.06.001.

Jacquemot, C., Dupoux, E., Bachoud-Lévi, A.C., 2007. Breaking the mirror: Asymmetrical disconnection between the phonological input and output codes. doi:10.1080/02643290600683342.

Jacquemot, C., Scott, S.K., 2006. What is the relationship between phonological short-term memory and speech processing? Trends in Cognitive Sciences 10, 480–486. doi:10.1016/j.tics.2006.09.002.

Jefferies, E., Crisp, J., Ralph, M.A.L., 2006. The impact of phonological or semantic impairment on delayed auditory repetition: Evidence from stroke aphasia and semantic dementia. Aphasiology 20, 963–992. doi:10.1080/02687030600739398.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. NeuroImage 62, 782–790. doi:10.1016/j.neuroimage.2011.09.015.

Johnson, K., 1990. The role of perceived speaker identity in F0 normalization of vowels. The Journal of the Acoustical Society of America 88, 642–654. doi:10.1121/1.399767.

Jonas, S., 1981. The supplementary motor region and speech emission. Journal of Communication Disorders 14, 349–373. doi:10.1016/0021-9924(81)90019-8.

Jones, E., Oliphant, T., Peterson, P., et al., 2001–. SciPy: Open source scientific tools for Python. URL: http://www.scipy.org/. [Online; accessed July 23, 2015].

Jonides, J., Schumacher, E.H., Smith, E.E., Koeppe, R.A., Awh, E., Reuter-Lorenz, P.A., Marshuetz, C., Willis, C.R., 1998. The role of parietal cortex in verbal working memory. Journal of Neuroscience 18, 5026–5034.

Josephs, O., Turner, R., Friston, K., 1997. Event-related fMRI. Human Brain Mapping 5, 243–248. doi:10.1002/(SICI)1097-0193(1997)5:4<243::AID-HBM7>3.0.CO;2-3.

Kilian-Hütten, N., Valente, G., Vroomen, J., Formisano, E., 2011. Auditory cortex encodes the perceptual interpretation of ambiguous sound. The Journal of Neuroscience 31, 1715–20. doi:10.1523/JNEUROSCI.4572-10.2011.

Kohavi, R., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence 14, 1137–1143.

Kriegeskorte, N., 2009. Relating population-code representations between man, monkey, and computational models. Frontiers in Neuroscience 3, 363–373. doi:10.3389/neuro.01.035.2009.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 103, 3863–3868. doi:10.1073/pnas.0600244103.

Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends in Cognitive Sciences 17, 401–412. doi:10.1016/j.tics.2013.06.007, arXiv:9809069v1.

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience 2, 4. doi:10.3389/neuro.06.004.2008.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience 12, 535–540. doi:10.1038/nn.2303.

Kröger, B.J., Kannampuzha, J., Neuschaefer-Rube, C., 2009. Towards a neurocomputational model of speech production and perception. Speech Communication 51, 793–809. doi:10.1016/j.specom.2008.08.002.

Ladefoged, P., Broadbent, D.E., 1957. Information Conveyed by Vowels. doi:10.1121/1.1908694.

Lee, Y.S., Turkeltaub, P., Granger, R., Raizada, R.D.S., 2012. Categorical Speech Processing in Broca's Area: An fMRI Study Using Multivariate Pattern-Based

Analysis. Journal of Neuroscience 32, 3942–3948. doi:10.1523/JNEUROSCI.3814-11.2012.

Leonard, M.K., Chang, E.F., 2014. Dynamic speech representations in the human temporal lobe. Trends in Cognitive Sciences 18, 472–479. doi:10.1016/j.tics.2014.05.001.

Levelt, W.J.M., Wheeldon, L., 1994. Do speakers have access to a mental syllabary? Cognition 50, 239–269. doi:10.1016/0010-0277(94)90030-2.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M., 1967. Perception of the speech code. Psychological Review 74, 431–461. doi:10.1037/h0020279.

Lichtheim, L., 1885. On Aphasia. Brain 7, 433–484. doi:10.1093/brain/7.4.433.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150–157. doi:10.1038/35084005.

Long M. A. et al., 2016. Functional Segregation of Cortical Regions Underlying Speech Timing and Articulation. Neuron 89, 1187–1193. doi:10.1016/j.neuron.2016.01.032.

Lotze, M., Seggewies, G., Erb, M., Grodd, W., Birbaumer, N., 2000. The representation of articulation in the primary sensorimotor cortex. Neuroreport 11, 2985–2989. doi:10.1097/00001756-200009110-00032.

MacNeilage, P.F., 1998. The frame/content theory of evolution of speech production. Behavioral and Brain Sciences 21, 499–511. doi:10.1017/S0140525X98001265.

Majerus, S., 2013. Language repetition and short-term memory: an integrative framework. Frontiers in Human Neuroscience 7, 1–16. doi:10.3389/fnhum.2013.00357.

Mandeville, J.B., Marota, J.J.A., Ayata, C., Moskowitz, M.A., Weisskoff, R.M., Rosen, B.R., 1999. MRI measurement of the temporal evolution of relative CMRO(2) during rat forepaw stimulation. Magnetic Resonance in Medicine 42, 944–951. doi:10.1002/(SICI)1522-2594(199911)42:5<944::AID-MRM15>3.0.CO;2-W.

Markiewicz, C.J., 2016. philips-cdas v0.1. doi:10.5281/zenodo.49853.

Markiewicz, C.J., Bohland, J.W., 2016. Mapping the cortical representation of speech sounds in a syllable repetition task. NeuroImage 141, 174–190. doi:10.1016/j.neuroimage.2016.07.023.

Martin, R.C., Breedin, S.D., Damian, M.F., 1999. The relation of phoneme discrimination, lexical access, and short-term memory: A case study and interactive activation account. Brain and Language 70, 437–482. doi:10.1006/brln.1999.2184.

Martuzzi, R., Ramani, R., Qiu, M., Rajeevan, N., Constable, R.T., 2010. Functional connectivity and alterations in baseline brain state in humans. NeuroImage 49, 823–834. doi:10.1016/j.neuroimage.2009.07.028.

Massaro, D.W., Chen, T.H., 2008. The motor theory of speech perception revisited. Psychonomic Bulletin & Review 15, 453–457. doi:10.3758/PBR.15.2.453, arXiv:NIHMS150003.

McGettigan, C., Warren, J.E., Eisner, F., Marshall, C.R., Shanmugalingam, P., Scott, S.K., 2011. Neural correlates of sublexical processing in phonological working memory. Journal of Cognitive Neuroscience 23, 1–17. doi:10.1162/jocn.2010.21491.

Mei L. et al., 2014. Artificial language training reveals the neural substrates underlying addressed and assembled phonologies. PLoS ONE 9, e93548. doi:10.1371/journal.pone.0093548.

Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., Iacoboni, M., 2007. The Essential Role of Premotor Cortex in Speech Perception. Current Biology 17, 1692–1696. doi:10.1016/j.cub.2007.08.064.

Menenti, L., Segaert, K., Hagoort, P., 2012. The neuronal infrastructure of speaking. Brain and Language 122, 71–80. doi:10.1016/j.bandl.2012.04.012.

Merrill, J., Sammler, D., Bangert, M., Goldhahn, D., Lohmann, G., Turner, R., Friederici, A.D., 2012. Perception of words and pitch patterns in song and speech. Frontiers in Psychology 3, 1–13. doi:10.3389/fpsyg.2012.00076.

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

Molnar-Szakacs, I., Iacoboni, M., Koski, L., Mazziotta, J.C., 2005. Functional Segregation within Pars Opercularis of the Inferior Frontal Gyrus: Evidence from fMRI Studies of Imitation and Action Observation. Cerebral Cortex 15, 986–994. doi:10.1093/cercor/bhh199.

Mumford, J.A., Turner, B.O., Ashby, F.G., Poldrack, R.A., 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. NeuroImage 59, 2636–2643. doi:10.1016/j.neuroimage.2011.08.076.

Myers, E.B., 2007. Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: An fMRI investigation. Neuropsychologia 45, 1463–1473. doi:10.1016/j.neuropsychologia.2006.11.005.

Myers, E.B., Blumstein, S.E., Walsh, E., Eliassen, J., 2009. Inferior frontal regions underlie the perception of phonetic category invariance. Psychological Science 20, 895–903. doi:10.1111/j.1467-9280.2009.02380.x.

Naselaris, T., Kay, K.N., 2015. Resolving Ambiguities of MVPA Using Explicit Models of Representation. Trends in cognitive sciences 19, 551–4. doi:10.1016/j.tics.2015.07.005.

Nichols, T., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Statistical Methods in Medical Research 12, 419–446. doi:10.1191/0962280203sm341ra.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A Toolbox for Representational Similarity Analysis. PLoS Computational Biology 10. doi:10.1371/journal.pcbi.1003553.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10, 424–430. doi:10.1016/j.tics.2006.07.005.

Nozari, N., Dell, G.S., 2013. How damaged brains repeat words: A computational approach. Brain and Language 126, 327–337. doi:10.1016/j.bandl.2013.07.005.

Numminen, J., Salmelin, R., Hari, R., 1999. Subject's own speech reduces reactivity of the human auditory cortex. Neuroscience Letters 265, 119–122. doi:10.1016/S0304-3940(99)00218-9.

Ogawa, S., Lee, T.M., Kay, A.R., Tank, D.W., 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proceedings of the National Academy of Sciences of the United States of America 87, 9868–72. doi:10.1073/pnas.87.24.9868.

Okada, K., Hickok, G., 2006a. Identification of lexical-phonological networks in the superior temporal sulcus using functional magnetic resonance imaging. NeuroReport 17, 1293–1296. doi:10.1097/01.wnr.0000233091.82536.b2.

Okada, K., Hickok, G., 2006b. Left posterior auditory-related cortices participate both in speech perception and speech production: Neural overlap revealed by fMRI. Brain and Language 98, 112–117. doi:10.1016/j.bandl.2006.04.006.

Okada, K., Smith, K.R., Humphries, C., Hickok, G., 2003. Word length modulates neural activity in auditory cortex during covert object naming. Neuroreport 14. doi:10.1097/01.wnr.0000094104.16607.

Oosterhof, N.N., Wiestler, T., Downing, P.E., Diedrichsen, J., 2011. A comparison of volume-based and surface-based multi-voxel pattern analysis. NeuroImage 56, 593–600. doi:10.1016/j.neuroimage.2010.04.270.

Papoutsi, M., de Zwart, J.A., Jansma, J.M., Pickering, M.J., Bednar, J.A., Horwitz, B., 2009. From phonemes to articulatory codes: An fMRI study of the role of Broca's area in speech production. Cerebral Cortex 19, 2156–2165. doi:10.1093/cercor/bhn239.

Paulesu, E., Frith, C.D., Frackowiak, R.S., 1993. The neural correlates of the verbal component of working memory. Nature 362, 342–345. doi:10.1038/362342a0.

Paus, T., 1996. Modulation of cerebral blood flow in the human auditory cortex during speech: Role of motor-to-sensory discharges. European Journal of Neuroscience 8, 2236–2246. doi:10.1111/j.1460-9568.1996.tb01187.x.

Peeva, M.G., Guenther, F.H., Tourville, J.A., Nieto-Castanon, A., Anton, J.L., Nazarian, B., Alario, F.X., 2010. Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech production network. NeuroImage 50, 626–638. doi:10.1016/j.neuroimage.2009.12.065.

Perrachione, T.K., Ghosh, S.S., 2013. Optimized design and analysis of sparse-sampling fMRI experiments. Frontiers in Neuroscience 7, 55. doi:10.3389/fnins.2013.00055.

Perrachione, T.K., Ghosh, S.S., Ostrovskaya, I., Gabrieli, J.D.E., Kovelman, I., 2017. Phonological working memory for words and nonwords in cerebral cortex. Journal of Speech, Language and Hearing Research .

Peterson, G.E., Barney, H.H., 1952. Control methods used in a study of the vowels. Journal of the Acoustical Society of America 24, 175–184.

Price, C.J., 2012. A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. NeuroImage 62, 816–847. doi:10.1016/j.neuroimage.2012.04.062.

Pugh, K.R., Mencl, W.E., Jenner, A.R., Katz, L., Frost, S.J., Lee, J.R., Shaywitz, S.E., Shaywitz, B.A., 2001. Neurobiological studies of reading and reading disability. Journal of Communication Disorders 34, 479–492. doi:10.1016/S0021-9924(01)00060-0.

Pulvermüller, F., Fadiga, L., 2010. Active perception: sensorimotor circuits as a cortical basis for language. Nature Reviews Neuroscience 11, 351–360. doi:10.1038/nrn2811.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proceedings of the National Academy of Sciences of the United States of America 103, 7865–7870. doi:10.1073/pnas.0509989103.

Purdon, P.L., Weisskoff, R.M., 1998. Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. Human Brain Mapping 6, 239–249. doi:10.1002/(SICI)1097-0193(1998)6:4<239::AID-HBM4>3.0.CO;2-4.

Raettig, T., Kotz, S.A., 2008. Auditory processing of different types of pseudowords: An event-related fMRI study. NeuroImage 39, 1420–1428. doi:10.1016/j.neuroimage.2007.09.030.

Raizada, R.D.S., Poldrack, R.A., 2007. Selective Amplification of Stimulus Differences during Categorical Processing of Speech. Neuron 56, 726–740. doi:10.1016/j.neuron.2007.11.001.

Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience 12, 718–724. doi:10.1038/nn.2331.

Ravizza, S.M., Delgado, M.R., Chein, J.M., Becker, J.T., Fiez, J.A., 2004. Functional dissociations within the inferior parietal cortex in verbal working memory. NeuroImage 22, 562–73. doi:10.1016/j.neuroimage.2004.01.039.

Riecker, A., Ackermann, H., Wildgruber, D., Dogil, G., Grodd, W., 2000. Opposite hemispheric lateralization effects during speaking and singing at motor cortex, insula and cerebellum. Neuroreport 11, 1997–2000. doi:10.1097/00001756-200006260-00038.

Robson, M.D., Gore, J.C., Constable, R.T., 1997. Measurement of the point spread function in MRI using constant time imaging. Magnetic Resonance in Medicine 38, 733–740. doi:10.1002/mrm.1910380509.

Roelofs, A., 1997. The WEAVER model of word-form encoding in speech production. Cognition 64, 249–284. doi:10.1016/S0010-0277(97)00027-9.

Rogalsky, C., Poppa, T., Chen, K.H., Anderson, S.W., Damasio, H., Love, T., Hickok, G., 2015. Speech repetition as a window on the neurobiology of auditory-motor integration for speech: A voxel-based lesion symptom mapping study. Neuropsychologia 71, 18–27. doi:10.1016/j.neuropsychologia.2015.03.012.

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A.R., Schulz, J.B., Fox, P.T., Eickhoff, S.B., 2012. Modelling neural correlates of working memory: A coordinate-based meta-analysis. NeuroImage 60, 830–846. doi:10.1016/j.neuroimage.2011.11.050.

Saur D. et al., 2008. Ventral and dorsal pathways for language. Proceedings of the National Academy of Sciences 105, 18035–18040. doi:10.1073/pnas.0805234105.

Schmidt, C.F., Zaehle, T., Meyer, M., Geiser, E., Boesiger, P., Jancke, L., 2008. Silent and continuous fMRI scanning differentially modulate activation in an auditory language comprehension task. Human Brain Mapping 29, 46–56. doi:10.1002/hbm.20372.

Segawa, J.A., Tourville, J.A., Beal, D.S., Guenther, F.H., 2015. The Neural Correlates of Speech Motor Sequence Learning. Journal of Cognitive Neuroscience 27, 819–831. doi:10.1162/jocn_a_00737, arXiv:1511.04103.

Siegel, J.S., Power, J.D., Dubis, J.W., Vogel, A.C., Church, J.A., Schlaggar, B.L., Petersen, S.E., 2014. Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. Human Brain Mapping 35, 1981–1996. doi:10.1002/hbm.22307, arXiv:NIHMS150003.

Smith S. M. et al., 2004. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23, S208–S219. doi:10.1016/j.neuroimage.2004.07.051.

Sörös, P., Sokoloff, L.G., Bose, A., McIntosh, A.R., Graham, S.J., Stuss, D.T., 2006. Clustered functional MRI of overt speech production. NeuroImage 32, 376–387. doi:10.1016/j.neuroimage.2006.02.046.

Stasenko, A., Bonn, C., Teghipco, A., Garcea, F.E., Sweet, C., Dombovy, M., McDonough, J., Mahon, B.Z., 2015. A causal test of the motor theory of speech perception: a case of impaired speech production and spared speech perception. Cognitive Neuropsychology 32, 38–57. doi:10.1080/02643294.2015.1035702, arXiv:15334406.

Stasenko, A., Garcea, F.E., Mahon, B.Z., 2013. What happens to the motor theory of perception when the motor system is damaged? Language and Cognition 5, 225–238. doi:10.1515/langcog-2013-0016.

Stelzer, J., Chen, Y., Turner, R., 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. NeuroImage 65, 69–82. doi:10.1016/j.neuroimage.2012.09.063.

Stevens, S.S., Volkmann, J., Newman, E.B., 1937. Scale for the Measurement of the Psychological Magnitude Pitch 19, 14–19. doi:10.1121/1.1915893.

Stone, M., 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society 36, 111–147.

Storkel, H.L., 2013. A corpus of consonant-vowel-consonant real words and nonwords: comparison of phonotactic probability, neighborhood density, and consonant age of acquisition. Behavior research methods 45, 1159–67. doi:10.3758/s13428-012-0309-7, arXiv:NIHMS150003.

Takai, O., Brown, S., Liotti, M., 2010. Representation of the speech effectors in the human motor cortex: Somatotopy or overlap? Brain and Language 113, 39–44. doi:10.1016/j.bandl.2010.01.008.

Talavage, T.M., Sereno, M.I., Melcher, J.R., Ledden, P.J., Rosen, B.R., Dale, A.M., 2003. Tonotopic Organization in Human Auditory Cortex Revealed by Progressions of Frequency Sensitivity. Journal of Neurophysiology 91, 1282–1296. doi:10.1152/jn.01125.2002.

Todd, M.T., Nystrom, L.E., Cohen, J.D., 2013. Confounds in multivariate pattern analysis: Theory and rule representation case study. NeuroImage 77, 157–165. doi:10.1016/j.neuroimage.2013.03.039.

Tomasi, D., Goldstein, R.Z., Telang, F., Maloney, T., Alia-Klein, N., Caparelli, E.C., Volkow, N.D., 2007. Widespread disruption in brain activation patterns to a working memory task during cocaine abstinence. Brain Research 1171, 83–92. doi:10.1016/j.brainres.2007.06.102.

Tourville, J.A., Guenther, F.H., 2012. Automatic cortical labeling system for neuroimaging studies of normal and disordered speech, in: Neuroscience Meeting Planner, Society for Neuroscience, New Orleans, LA. p. 681.06.

Tourville, J.A., Reilly, K.J., Guenther, F.H., 2008. Neural mechanisms underlying auditory feedback control of speech. NeuroImage 39, 1429–1443. doi:10.1016/j.neuroimage.2007.09.054.

Tremblay, P., Baroni, M., Hasson, U., 2013a. Processing of speech and non-speech sounds in the supratemporal plane: Auditory input preference does not predict sensitivity to statistical structure. NeuroImage 66, 318–332. doi:10.1016/j.neuroimage.2012.10.055.

Tremblay, P., Deschamps, I., Gracco, V.L., 2013b. Regional heterogeneity in the processing and the production of speech in the human planum temporale. Cortex 49, 143–157. doi:10.1016/j.cortex.2011.09.004.

Tremblay, P., Gracco, V.L., 2006. Contribution of the frontal lobe to externally and internally specified verbal responses: fMRI evidence. NeuroImage 33, 947–957. doi:10.1016/j.neuroimage.2006.07.041.

Tremblay, P., Gracco, V.L., 2009. Contribution of the pre-SMA to the production of words and non-speech oral motor gestures, as revealed by repetitive transcranial magnetic stimulation (rTMS). Brain Research 1268, 112–124. doi:10.1016/j.brainres.2009.02.076.

Tremblay, P., Gracco, V.L., 2010. On the selection of words and oral motor responses: Evidence of a response-independent fronto-parietal network. Cortex 46, 15–28. doi:10.1016/j.cortex.2009.03.003.

Tremblay, P., Small, S.L., 2011. Motor response selection in overt sentence production: a functional MRI study. Frontiers in Psychology 2, 1–14. doi:10.3389/fpsyg.2011.00253.

Vaden, K.I., Muftuler, L.T., Hickok, G., 2010. Phonological repetition-suppression in bilateral superior temporal sulci. NeuroImage 49, 1018–1023. doi:10.1016/j.neuroimage.2009.07.063.

Vitevitch, M.S., Luce, P.A., 1999. Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. Journal of Memory and Language 40, 374–408. doi:10.1006/jmla.1998.2618.

Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., Rosseel, Y., 2011. neuRosim: An R package for generating fMRI data. Journal of Statistical Software 44, 1–18. doi:10.18637/jss.v044.i10.

Wildgruber, D., Ackermann, H., Klose, U., Kardatzki, B., Grodd, W., 1996. Functional lateralization of speech production at primary motor cortex: a fMRI study. Neuroreport 7, 2791–2795.

Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nature Neuroscience 7, 701–2. doi:10.1038/nn1263.

Wise, R.J.S., Scott, S.K., Blank, S.C., Mummery, C.J., Murphy, K., Warburton, E.A., 2001. Separate neural subsystems within 'Wernicke's area'. Brain 124, 83–95. doi:10.1093/brain/124.1.83.

You, J., Markiewicz, C.J., Bohland, J.W., 2015. Formant detection scripts for "mapping the cortical representation of speech sounds in a syllable repetition task. doi:10.5281/zenodo.51362.

Zaehle, T., Schmidt, C.F., Meyer, M., Baumann, S., Baltes, C., Boesiger, P., Jancke, L., 2007. Comparison of "silent" clustered and sparse temporal fMRI acquisitions in tonal and speech perception tasks. NeuroImage 37, 1195–1204. doi:10.1016/j.neuroimage.2007.04.073.

Zhang, Q., Hu, X., Luo, H., Li, J., Zhang, X., Zhang, B., 2016. Deciphering phonemes from syllables in BOLD signals in human superior temporal gyrus. European Journal of Neuroscience 43, 773–781. doi:10.1111/ejn.13164.

Ziegler, W., Kilian, B., Deger, K., 1997. The role of the left mesial frontal cortex in fluent speech: Evidence from a case of left supplementary motor area hemorrhage. Neuropsychologia 35, 1197–1208. doi:10.1016/S0028-3932(97)00040-7.

[redacted]

Curriculum Vitae

[redacted]