

**OPTIMIZATION AND MACHINE
LEARNING IN MRI: APPLICATIONS IN
RAPID MR IMAGE RECONSTRUCTION
AND ENCODING MODELS OF CORTICAL
REPRESENTATIONS**

A DISSERTATION SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

By
Mohammad Shahdloo
February 2020

Optimization and Machine Learning in MRI: Applications in Rapid
MR Image Reconstruction and Encoding Models of Cortical Representations

By Mohammad Shahdloo

February 2020

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Tolga ukur (Advisor)

Ergin Atalar

Mehmet Erkut Erdem

Emine lk Sarıtař ukur

lkay Ulusoy

Approved for the Graduate School of Engineering and Science:

Ezhan Karařan
Director of the Graduate School

ABSTRACT

OPTIMIZATION AND MACHINE LEARNING IN MRI: APPLICATIONS IN RAPID MR IMAGE RECONSTRUCTION AND ENCODING MODELS OF CORTICAL REPRESENTATIONS

Mohammad Shahdloo

Ph.D. in Electrical and Electronics Engineering

Advisor: Tolga Çukur

February 2020

Magnetic Resonance Imaging (MRI) is a non-invasive medical imaging modality that is widely used by clinicians and researchers to picture body anatomy and neuronal function. However, long scan time remains a major problem. Recently, multiple techniques have emerged that reduce the acquired MRI signal samples, hence dramatically accelerating the acquisition. These techniques involve sophisticated signal reconstruction procedures that in essence require solving regularized optimization problems, and clinical adoption of accelerated MRI critically relies on self-tuning solutions for these problems. Further to this, recent experimental approaches in cognitive neuroscience favor employing naturalistic audio-visual stimuli that closely resemble humans' daily-life experience. Yet, these modern paradigms inevitably lead to huge functional MRI (fMRI) datasets that require advanced statistical and computational techniques to uncover the large amount of embedded information. Here, we propose a novel efficient data-driven self-tuning reconstruction method for accelerated MRI. We demonstrate superior performance of the proposed method across various simulated and in vivo datasets and under various scan configurations. Furthermore, we develop statistical analysis tools to investigate the neural representation of hundreds of action categories in natural movies in the brain via fMRI, and study their attentional modulations. Finally, we develop a model-based framework to estimate temporal extent of semantic information integration in the brain, and investigate its attentional modulations using fMRI data recorded during natural story listening. In short, the methodological and analytical approaches introduced in this thesis greatly benefit clinical utility of accelerated MRI, and enhance our understanding of brain function in daily life.

Keywords: magnetic resonance imaging (MRI), self-tuning reconstruction, encoding models, action perception, deep learning, language model.

ÖZET

MRG'DE OPTİMİZASYON VE MAKİNE ÖĞRENİMİ: HIZLI MR GÖRÜNTÜ REKONSTRÜKSİYONU VE BEYİNDEKİ TEMSİLLERİN KODLAMA MODELLERİNE UYGULANIŞI

Mohammad Shahdloo

Elektrik Elektronik Mühendisliği, Doktora

Tez Danışmanı: Tolga Çukur

Şubat 2020

Manyetik rezonans görüntüleme (MRG) klinisyenler ve araştırmacılar tarafından vücut anatomisi ve sinirsel işlevi resmetmede yaygın olarak kullanılan müdahalesiz bir tıbbi görüntüleme yöntemidir. Fakat, görüntülemenin uzun sürüşü büyük bir problem olmaya devam etmektedir. Yakın zamanda, toplanan MRG sinyallerinin miktarını azaltarak ciddi bir hız artışına sebep olan birkaç teknik geliştirildi. Bu teknikler temelde birtakım düzenlenmiş optimizasyon problemleri çözümü gerektiren gelişmiş sinyal rekonstrüksiyon prosedürleri içermekte ve hızlandırılmış MRG'nin klinik uygulamalarda benimsenmesi bu problemlere getirilecek öz-ayar çözümlerine ciddi derecede dayanmaktadır. Öte yandan, bilişsel nörobilimdeki yeni deneysel yaklaşımlar insanların gündelik hayatta başlarından geçenele oldukça benzeyen doğal işitsel-görsel uyarıların kullanımını teşvik etmektedir. Ancak, bu modern paradigmlar kaçınılmaz olarak büyük miktarda gömülü bilginin ortaya çıkarılması için gelişmiş istatistiksel ve sayısal tekniklerin kullanımını gerektiren devasa fonksiyonel MRG (fMRG) veri kümelerinin ortaya çıkmasına yol açmaktadır. Bu çalışmada hızlandırılmış MRG için yeni ve etkili bir veri-güdümlü öz-ayar rekonstrüksiyon metodu ileri sürüyoruz. İleri sürülen metodun çeşitli temsili ve canlı üzerinde elde edilmiş (in vivo) veri kümeleri ve çeşitli görüntüleme konfigürasyonları altında üstün performans ortaya koyduğunu gösteriyoruz. Ayrıca, doğal filmlerdeki yüzlerce eylem kategorisinin beyindeki sinirsel temsili ve bunların dikkate dayalı değişimlerini fMRG ile incelemek için istatistiksel analiz teknikleri geliştirdik. Son olarak beyindeki anlamsal bilgi entegrasyonunun zamansal kapsamını ölçmek ve dikkate dayalı değişimlerini doğal hikaye dinleme deneyinde toplanmış fMRG verisi kullanarak incelemek için modele dayalı bir altyapı geliştirdik. Özetle, bu tezde ortaya konan metodolojik ve analitik yaklaşımlar hızlandırılmış MRG'nin

linik kullanımına ve gündelik hayatta beyin işleyişini anlamamıza büyük katkı sağlamaktadır.

Anahtar sözcükler: manyetik rezonans görüntüleme (MRG), öz-ayar rekonstrüksiyon, eylem algı, derin öğrenme, dil modeli.

Acknowledgement

Contrary to popular belief, doing a Ph.D does not require the highest level of intelligence. It's not the difficulty of the study that makes it hard, it's the process of going through that long road full of various unthinkable obstacles that makes it hard. Exactly for this reason, I am deeply thankful and blessed for the amazing people who supported me going on this road.

I could not have hoped for a better advisor than Prof. Tolga Çukur. He is the most gentle, supportive, and knowledgeable person I know. He provided me the tools and guidance to freely pursue my research directions. He always had the patience to answer my questions. He taught me how to be not a student, but a researcher. For all these, I am always thankful to him.

My research during these years took place mostly at UMRAM. I would like to thank Prof. Ergin Atalar, the director of UMRAM, and all other members of UMRAM family for providing such a professional and motivating research environment.

I am indebted to my thesis committee members, Prof. Ergin Atalar and Prof. Erkut Erdem, who provided me with their precious suggestions during the development of this thesis. I also would like to thank Prof. Emine Ülkü Sarıtaş and Prof. İlkay Ulusoy, jury members, for their time and effort examining this thesis.

I would not have been in this point in my life and career if I did not have Saba Darbani and Sadra Khalkhali beside me. I got the most precious inspirations from them that have been, and would be, extremely invaluable for me forever.

I could not go through these years without having those joyful moments with Sina Rezaei and Sepideh Yekani; the kindest friends I could ever imagine to have. Thank you for all those long discussions that would have started from a cake –usually *horizontal* banana cake– recipe but ended in analyzing a –nonsense– hot political debate up until 4 in the morning! Thank you for coloring my days.

I would also like to express my thanks to all lab members, especially Salman Dar, Efe Ilıcak, Emin Çelik, and Ibrahim Kiremitci, and also my dearest friends Mahkame, Sina, Salar, Mona, Parisa, Melisa, Kerim, Soheil, Akbar, and many others who made these past years at Bilkent a memorable episode of my life.

I would like to thank to my family for their everlasting love and support. My

achievements today are rooted in the calm, kind, and supportive environment that they had provided, from my childhood to adulthood.

Finally and above all, I would like to thank Hasti, the love my life and my best friend. We went through all these years together, shoulder to shoulder. She has always been the greatest source of motivation and encouragement for me. I dedicate this thesis to her. Thank you Hasti!

Mo Shahdloo

24 - Dec - 2019

Contents

1	Introduction	1
1.1	Contributions	3
1.2	Outline	4
2	Fundamentals of MR Image Reconstruction and Analysis	5
2.1	MR Image Reconstruction	5
2.1.1	Magnetic excitation and signal acquisition	5
2.1.2	Accelerated MRI	6
2.2	Functional MRI (fMRI)	7
2.2.1	Encoding Models in fMRI	8
3	Rapid Self-Tuning Compressed Sensing MRI	11
3.1	Introduction	12
3.2	Theory	14
3.2.1	Reconstruction by calibration over tensors	14
3.2.2	PESCaT	15
3.2.3	Parameter tuning by projection onto epigraph sets	17
3.3	Methods	23
3.3.1	Alternative reconstructions	23
3.3.2	Simulations	25
3.3.3	In vivo experiments	27
3.4	Results	30
3.4.1	Simulations	30
3.4.2	In vivo experiments	36
3.5	Discussion	47

3.6	Publications	51
4	Attentional Modulations of Action-Category Representation in the Brain	53
4.1	Introduction	54
4.2	Methods	58
4.2.1	Subjects	58
4.2.2	fMRI data collection	58
4.2.3	Stimuli and experimental design	58
4.2.4	fMRI data preprocessing.	60
4.2.5	Definition of regions of interest	60
4.2.6	Head motion, eye movement, and physiological noise	61
4.2.7	Category features	62
4.2.8	Motion-energy features	62
4.2.9	Space-Time Interest Points (STIP) features	63
4.2.10	Model estimation and testing	63
4.2.11	Action category responses	65
4.2.12	Variance partitioning analysis	65
4.2.13	Semantic representation of actions	66
4.2.14	Characterizing tuning shifts	68
4.3	Results	69
4.3.1	Attention alters category tuning profiles	69
4.3.2	Attention warps semantic representation of actions	72
4.3.3	Distribution of tuning shifts across cortex	73
4.3.4	Tuning shifts for unattended categories	80
4.3.5	Tuning shifts are influenced by intrinsic selectivity for action categories	81
4.4	Discussion	85
4.5	Publications	88
5	Temporal Receptive Windows in the Brain Mapped via Deep Language Models	90
5.1	Introduction	91
5.2	Methods	93

5.2.1	Subjects	93
5.2.2	fMRI data collection	94
5.2.3	Stimuli and experimental design	94
5.2.4	fMRI data preprocessing	95
5.2.5	Definition of regions of interest	95
5.2.6	Semantic embedding of stories	96
5.2.7	Language model features	96
5.2.8	Model estimation and validation	97
5.2.9	Physiological noise controls	98
5.2.10	Temporal receptive windows	98
5.2.11	Alternative approach to measure TRW	100
5.2.12	Sensitivity and bias of TRWs	100
5.3	Results	101
5.3.1	Language model well predicts responses to narrated stories	101
5.3.2	Distribution of TRWs across cortex	104
5.3.3	Attention modulates TRWs across cortex	106
5.3.4	Auditory search for object categories selectively modulates TRWs	108
5.4	Discussion	109
5.5	Publications	113
6	Concluding Remarks and Future Directions	114

List of Figures

2.1	Illustration of the encoding models.	9
3.1	Flowchart of the PESCiT reconstruction.	17
3.2	The projection onto epigraph sets (PES) approach illustrated in \mathbb{R}^3	19
3.3	The numerical phantom designed for assessment of reliability against variation in level of detail and spatial resolution.	31
3.4	Evolution of the PESCiT cost terms.	32
3.5	Reconstruction performance as a function of the ℓ_1 -epigraph scaling constant.	33
3.6	Reconstruction performance as a function of the TV-epigraph scaling constant.	34
3.7	Optimal sparsity and TV regularization parameters as a function of SNR.	35
3.8	Reconstructions of phase-cycled bSSFP acquisitions of the simulated brain phantom.	36
3.9	Reconstruction performance as a function of acquisition SNR.	37
3.10	Reconstruction performance for varying β_{l_1} as a function of level of detail and resolution of the phantom.	38
3.11	Reconstruction performance for varying β_{TV} as a function of level of detail and resolution of the phantom.	39
3.12	Progression of PESCiT across iterations demonstrated on an in vivo bSSFP dataset.	40
3.13	Progression of PESCiT across iterations demonstrated on an in vivo T1-weighted dataset.	41

3.14	Representative cross-sections from the ToF and the phase-cycled bSSFP datasets.	42
3.15	Reconstructions of in vivo bSSFP acquisitions of the brain at R=6.	43
3.16	Reconstructions of in vivo T1-weighted acquisitions of the brain at R=4.	44
3.17	Reconstructions of in vivo ToF angiography acquisitions of the brain at R=4.	45
3.18	Convergence behavior of self-tuning reconstructions.	47
3.19	Reconstructions of 32-channel in vivo bSSFP acquisitions of the brain at R=10.	48
4.1	Hypothesized changes in semantic representation of action categories.	57
4.2	Model fitting and validation procedure.	64
4.3	Prediction performance of the category model.	71
4.4	Semantic space underlying action category responses.	74
4.5	Cortical flat maps of projections onto the semantic space for subject S1.	76
4.6	Cortical flat maps of projections onto the semantic space for subject S2.	77
4.7	Cortical flat maps of projections onto the semantic space for subject S3.	77
4.8	Cortical flat maps of projections onto the semantic space for subject S4.	78
4.9	Cortical flat maps of projections onto the semantic space for subject S5.	78
4.10	Cortical distribution of tuning shifts.	79
4.11	Tuning shifts for unattended categories.	82
4.12	Tuning shifts during search for individual targets.	84
5.1	Model fitting, validation, and estimation of TRWs.	99
5.2	Hypothesized changes in process memory timescales.	102
5.3	TRW values estimated during passive listening.	103
5.4	TRW values estimated via the alternative approach.	105
5.5	Attentional modulation of TRWs.	107

5.6 Attentional bias of TRWs.	110
5.7 TRWs and their attentional modulations in anatomical ROIs.	111

List of Tables

3.1	Peak signal-to-noise ratio (PSNR) for simulated phantom	26
3.2	PSNR and NRMSE for in vivo bSSFP dataset	28
3.3	PSNR and NRMSE for in vivo T1-weighted dataset	28
3.4	PSNR and NRMSE for in vivo ToF dataset	28
3.5	Reconstruction times for in vivo datasets.	46
3.6	Numerical assessment of bSSFP reconstructions at high acceleration rates.	48
4.1	Clusters of action categories.	67

Chapter 1

Introduction

Magnetic resonance imaging (MRI) is a noninvasive medical imaging technique which is used to visualize body anatomy and acquire functional brain data. Unlike other medical imaging modalities, such as Computed Tomography (CT), MRI does not use ionizing radiation but rather uses static and dynamic magnetic fields, and radio waves to generate images. MRI can be used to produce different tissues contrasts, making it a more efficient choice than CT. Moreover, MRI can be sensitized to measure brain blood flow and concentration, among other metabolites, as indicators of the neural brain activity with high spatio-temporal resolution and without the need to use ionizing radiation, distinguishing it from positron-emission tomography (PET) imaging. These properties have placed anatomical and functional MRI among the most popular imaging modalities for clinical and research applications.

MRI has improved dramatically in functionality, imaging quality and imaging speed since its invention more than 40 years ago. These revolutionary improvements have benefited clinical and research imaging applications dramatically. Apart from improvements in MRI hardware, development of fast pulse sequences and efficient scanning trajectories, this revolution has been realized by emergence of novel signal processing frameworks, such as compressed sensing (CS), that relax MRI signal sampling constraints and enable acquiring far less

samples, leading to dramatically shorter scan times. Recent decades have also seen functional MRI (fMRI) emerging as an increasingly popular tool to study the neural basis of cognitive processes. In its early days, fMRI was employed in cognitive neuroscience to measure regional brain activity in subjects while they performed simplified cognitive tasks, such as looking at gratings or checkerboard patterns, leading to detection of localized activation regions associated with different perceptual tasks. Yet, recent approaches to cognitive neuroscience favor computational frameworks that leverage functional responses to naturalistic stimuli (i.e. movies or stories) acquired from the whole brain during hour-long periods of time. These recent experimental approaches lead to large amounts of functional data and leverage data-driven approaches to gain novel insight into brain's cognitive functions. These developments in anatomical and functional MRI should inevitably be complemented by modern methodological and analytical tools.

Reconstructing accelerated MRI acquisitions using compressed sensing framework involves sophisticated techniques that can in essence be posed as regularized optimization problems. These problems typically weigh fidelity of the reconstruction to the measured data relative to one or more regularization terms. However, optimal weighing of these several terms is needed to produce high quality images from sparse measurements. Thus, optimizing the regularization weights is a critical component of a successful CS reconstruction. However, typical MRI acquisitions contain data from multiple coils and, possibly, multiple contrasts with different signal characteristics. As a result, reconstructing modern multi-coil multi-contrast datasets naturally needs optimizing a multitude of regularization parameters, that is infeasible to perform manually. Moreover, because of the variability in scan, anatomy, and subject parameters that influence acquired MR signal characteristics, hard-coding these regularization terms would lead to sub-optimal reconstructions. Thus, optimal data-driven parameter estimation is a crucial step toward successful implementation of CS in clinical applications.

Modern experimental frameworks in computational neuroscience tend to favor naturalistic stimuli, that lead to huge amounts of fMRI data. Uncovering the vast amount of information within large fMRI datasets acquired during naturalistic experiments requires developing sophisticated techniques that have not been

common in the cognitive neuroscience literature until recently. Many past reports have studied interaction of handfuls of parameters with simplistic perceptual inputs. Yet, analyzing fMRI responses to natural stimuli involves investigating interactions among thousands of model parameters and responses via statistical methods. In a typical such dataset one has to fit nearly 50,000 models, that each map nearly 5000 data samples to the time-course of nearly 1000 features. Further to this, many recent studies suggest employing computational models based on deep-learning. These methods either develop deep models that can mimic naturalistic brain behaviors, or they leverage latent representation of the naturalistic stimuli to model fMRI responses. These drives motivate the increasing need for optimization and machine-learning-based approaches for reconstruction and analysis of anatomical and functional MRI datasets.

1.1 Contributions

The contributions of this thesis are an optimization method for data-driven self-tuning CS-MRI reconstruction, and two machine-learning based analyses of large-scale fMRI experiments. First, we developed PESCaT (i.e. projection onto epigraph sets for reconstruction by calibration over tensors) as a self-tuning CS framework for reconstruction of generic multi-coil multi-acquisition accelerated MRI datasets [176, 177, 180]. Second, we investigated semantic representation of hundreds of action categories in natural scenes across cortex, and their modulations during natural visual search for action categories [181]. Third, we developed a model-based approach to estimate temporal window of past semantic information integration (i.e. temporal receptive windows, TRWs) in the brain, and investigated attentional modulations of TRWs during category-based natural audition [178].

1.2 Outline

The rest of this thesis is organized as follows. In chapter 2, we introduce the reader to fundamentals of MRI data acquisition and image reconstruction, and encoding models for fMRI datasets. In chapter 3, we develop theoretical principles for PESCiT as a self-tuning CS-MRI reconstruction technique and provide empirical evidence for its superior performance in reconstruction of multiple simulated and in vivo accelerated acquisitions. In chapter 4, we employ statistical analysis tools to investigate attentional modulations of semantic representation of hundreds of action categories in the brain using fMRI data acquired during presentation of natural movies. In chapter 5, we develop a model-based approach to measure temporal receptive windows in the brain using fMRI data collected during natural story listening, and we investigate their attentional modulations during category-based auditory attention. Finally, concluding remarks and directions for future work are provided in Chapter 6.

Chapter 2

Fundamentals of MR Image Reconstruction and Analysis

2.1 MR Image Reconstruction

2.1.1 Magnetic excitation and signal acquisition

Magnetic resonance imaging (MRI) employs a combination of static and dynamic magnetic fields, and radio-frequency waves to produce signals depicting body anatomy and brain function. The imaging process starts with transmitting a radio-frequency (RF) wave into the tissue of interest, that is placed in a strong static magnetic field. The energy in the transmitted RF wave gets absorbed by – mostly– hydrogen atoms in the tissue, tilting the direction of their magnetic spin moment. Subsequently, the RF excitation gets turned off and the tilted spins transmit back the absorbed energy, inducing a voltage signal on the receiver coil. However, the collected signal is the superposition of signals transmitted from *all* parts of the excited tissue. To disentangle the signal from different tissue locations, gradient magnetic fields are used. The gradient coils superimpose a spatially varying gradient field on top of the static field, hence, spatially encoding

the received signal by tilting tissue spins in a spatially-varying manner. This spatially-encoded signal is digitized using an analog to digital converter (ADC) to form the raw MR data, namely the *k-space* data. Finally, the image gets reconstructed by performing an inverse Fourier transform of the k-space data, leveraging the underlying MR signal properties.

Variation of chemical tissue properties is the primary source of MRI image contrast. Each tissue type excites the absorbed RF power at a certain rate, leading to tissue-dependent MR signal across the field of view, reflected in contrast difference across pixels of the reconstructed image. Moreover, by modifying the configuration of MR acquisition procedure the received signal can be sensitized to different tissue properties.

2.1.2 Accelerated MRI

In each cycle of MR excitation/acquisition, some portion of the k-space data is acquired by probing the k-space along user-defined trajectories. These cycles are repeated until the acquired data is sufficient to perform a successful inverse transform, leading to prolonged scan times. Several successful techniques have been proposed to reduce scan time, while maintaining reconstruction quality. A group of these methods employ spatial redundancy in the MR signal to partially collect the data from different portions of the k-space via several parallel receive coils. Various reconstruction techniques are then used to reconstruct the MR image from these partial acquisitions, either in the image domain [159] or in the spatial-frequency domain [65]. A more modern approach, compressed sensing MRI (CS-MRI; [127]), leverages sparsity of MR images in a transform (e.g. wavelet, total variation) domain to acquire k-space samples at a rate far less than what Nyquist sampling theorem devises, hence accelerating the scan. CS-MRI theoretically lays on incoherent undersampling of the k-space that leads to noise-like image artifacts. Then, the reconstruction is performed by solving an optimization problem that weighs consistency of the reconstruction to the acquired data versus sparsity in the transform domain. To elaborate, suppose m be the reconstructed image

of interest, y be the measured signal, $\Psi_1 \dots \Psi_k$ be transformations that map the image into k separate sparse representations (i.e. sparsifying transforms), and F_u be the undersampled Fourier transform. The reconstruction problem then can be formulated as

$$\begin{aligned}
 & \text{minimize } \|F_u m - y\|_2 \\
 & \text{s.t. } \|\Psi_1 m\|_1 < \epsilon_1 \\
 & \quad \dots \\
 & \|\Psi_k m\|_1 < \epsilon_k
 \end{aligned} \tag{2.1}$$

Here, the first term enforces fidelity of the reconstruction to the acquired samples, and the subsequent terms enforce sparsity of the reconstruction in each of the transform domains, where ϵ_i controls the sparsity level in the i th transform domain. The sparsifying transforms can be assumed to be the wavelet transform, total-variation transform, or any other sparsifying transform depending on the reconstruction problem requirements. The problem in Eq. 2.1 can be equivalently expressed in the regularized form as

$$\hat{m} = \underset{m}{\text{arg min}} \|F_u m - y\|_2^2 + \sum_{i=1}^k \lambda_i \|\Psi_i m\|_1 \tag{2.2}$$

where \hat{m} is the reconstructed image, and the regularization terms λ_i control the reconstruction sparsity level in each of the transform domains. Finally, the equation in Eq. 2.2 can be solved via various iterative approaches, such as projection onto convex sets (POCS) or alternating direction method of multipliers (ADMM), after tuning the regularization parameters λ_i . In chapter 3 we develop a reconstruction method that optimizes the regularization terms λ_i using the acquired undersampled data.

2.2 Functional MRI (fMRI)

Anatomical MRI acquisitions are typically sensitive to magnetic properties of the hydrogen content of tissues. However, MRI acquisition sequence can be modified

to be sensitive to changes in regional cerebral blood flow. Blood flow distributes the oxygen through different body organs by carrying it via an iron-containing oxygen-transport metalloprotein, namely hemoglobin. However, magnetic properties of hemoglobin is different when it carries oxygen (i.e. oxyhemoglobin) versus when it does not carry oxygen (i.e. deoxyhemoglobin); oxyhemoglobin is diamagnetic, while deoxyhemoglobin is paramagnetic. Moreover, as the brain uses oxygen to perform computational tasks, the local ratio of oxyhemoglobin to deoxyhemoglobin in the brain gradually varies during performing cognitive tasks. Thus, comparing multiple consecutive MR acquisitions that are sensitized to detect deoxyhemoglobin, namely the blood oxygen-dependent (BOLD) signal, can be used as a proxy to image local neural activations in the brain.

2.2.1 Encoding Models in fMRI

A common aim in cognitive neuroscience is to understand what sensory, cognitive or motor information is represented in a specific region of the brain. In this scheme, one attempts to fit an *encoding model* to the fMRI data that reflects how neural activity varies as a function of variation in the perceptual input. To formalize this modeling framework, suppose that inputs lay in an *input space*, and are mapped –possibly nonlinearly– to a latent feature representation. This nonlinear mapping would thus map each point from the input space to a point in the *feature space*. Finally, the encoding model is a linear mapping from points in the feature space to points in the *activity space*, where BOLD responses lay (Fig. 2.1, [136]).

Generalized linear models (GLM) used in the statistical parameter mapping (SPM) framework [57] are a classical example of encoding models. In this framework, the experimenter independently manipulates several dimension of the stimulus and finds model weights that directly relate the response in each voxel to the variation in the manipulated stimulus dimensions. Then, statistical significance of the model weights is assessed in each voxel and aggregated across a given region of interest. Note that encoding models naturally have the potential

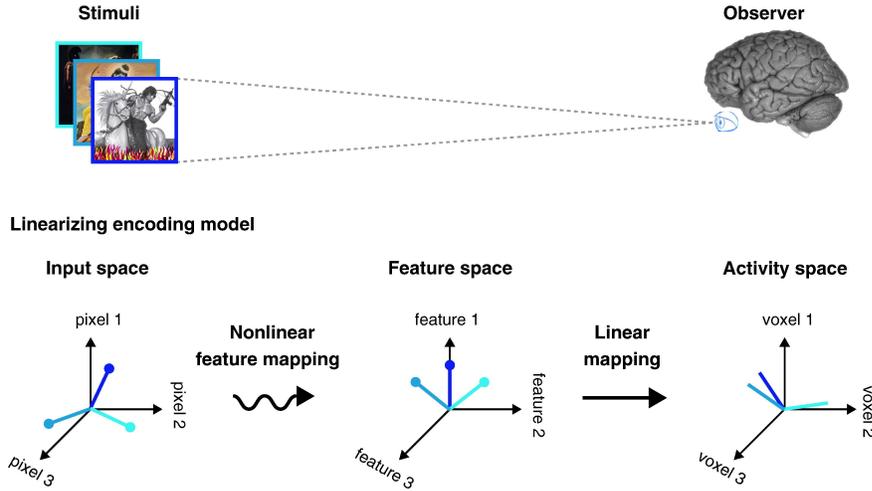


Figure 2.1. Illustration of the encoding models. During cognitive perception, brain can be considered as a nonlinear system mapping the stimulus to neural activity. Each stimulus sample is a point in the input space. A latent representation of the stimulus can be assumed to lay as a point in the feature space, and a nonlinear mapping relates each point in the input space to each point in the feature space. Then, a linear mapping maps each point in the feature space to a response laying in the activity space. Figure reprinted from Neuroimage, 56(2), T. Naselaris, K.N. Kay, S. Nishimoto, J.L. Gallant, “Encoding and decoding in fMRI”, Page 11, Copyright (2011), with permission from Elsevier.

of exploring the capability of multiple different feature representations of a single given stimuli to explain the response variance. This can simply be achieved by projecting inputs from the input space to different feature spaces, each capturing a different dimension of variability in the stimulus. Yet, most approaches of fMRI data analysis undermine this capability by using simplistic stimuli that represent discrete classes, or they probe discrete dimensions of cognitive processing like attending to several specific locations. While, ecologically valid stimuli that contain high variance across multiple cognitive dimensions, for instance natural movies and natural stories, can be used to test different hypotheses having a single set of stimulus and response in hand. However, the drawback of using natural stimuli in the context of linearized encoding models is the need to use sophisticated statistical analyses that can investigate the relationship between the rich information embedded in the stimulus and the responses.

Prediction performance of encoding models is limited by noise. From a modeling perspective, noise is every activity that is not reliably associated with the stimuli or task. This includes noise due to hardware imperfections such as magnetic field instability and heating, physiological effects such as respiratory and cardiac activity, in addition to many other cognitive factors such as expectation, arousal, attention to nuisance stimulus features and so on. An statistical approach can be taken to summarize the relationship between stimuli, response, and noise in encoding models [136]. Assume an stimulus sample s laying in the input space, that has the feature representation $f(s)$ in the feature space, evokes a response r in the activity space. The uncertainty in the encoding model can be formulated as an *encoding distribution* $p(r|f(s))$, where each sample from this distribution indicates the probability of the response being evoked by the latent representation of the stimulus, given the noise. To elaborate, consider several repetitions of presenting a given image and recording the responses. Due to noise, each of these repetitions would result in a separate point in the activity space and the collection of all responses form a cloud of points in the activity space. The encoding distribution would determine the size and shape of this cloud and the best that the encoding model can do is to determine the *most probable* point in this distribution of responses. Thus, this optimal encoding model can be stated as

$$M = W^T f(s) = \underset{r}{\operatorname{arg\,max}} \quad p(r|f(s)) \quad (2.3)$$

where W is the set of model weights that linearly maps the feature representation $f(s)$ to the responses. The encoding distribution in fMRI can be safely considered a multi-dimensional Gaussian distribution

$$p(r|f(s)) \sim \exp\left(-\frac{1}{2}(r - W^T f(s))^T \Sigma^{-1}(r - W^T f(s))\right) \quad (2.4)$$

Under this assumption, the model weights can be assessed via solving a regularized regression problem. In chapters 4, 5 we employ this framework to model large-scale brain activity evoked by natural movies and stories.

Chapter 3

Rapid Self-Tuning Compressed Sensing MRI

Summary

The compressed sensing (CS) framework leverages the sparsity of MR images to reconstruct from undersampled acquisitions. CS reconstructions involve one or more regularization parameters that weigh sparsity in transform domains against fidelity to acquired data. While parameter selection is critical for reconstruction quality, the optimal parameters are subject and dataset specific. Thus, commonly practiced heuristic parameter selection generalizes poorly to independent datasets. Recent studies have proposed to tune parameters by estimating the risk of removing significant image coefficients. Line searches are performed across the parameter space to identify the parameter value that minimizes this risk. Although effective, these line searches yield prolonged reconstruction times. Here, we propose a new self-tuning CS method that uses computationally efficient projections onto epigraph sets of the ℓ_1 and total-variation norms to simultaneously achieve parameter selection and regularization. In vivo demonstrations are provided for balanced steady-state free precession, time-of-flight, and T1-weighted imaging. The proposed method achieves an order of magnitude improvement in computational efficiency over line-search methods while maintaining near-optimal parameter selection.

3.1 Introduction

The compressed sensing (CS) framework was recently proposed for accelerated MRI, where compressibility of MR images are employed to reconstruct from undersampled acquisitions [127, 18, 36]. To do this, CS reconstructions are typically cast as regularized optimization problems that weigh data consistency against sparsity in some transform domain (e.g., wavelet domain, total variation (TV)) [127]. The weighing between data consistency and sparsity is governed by regularization parameters. High parameter values overemphasize sparsity at the expense of introducing inconsistency to acquired data samples, potentially leading to feature losses. Meanwhile, low parameter values render the reconstructions ineffective in suppressing residual aliasing and noise in undersampled acquisitions. Since the optimal regularization parameters are subject and dataset specific, time-consuming and potentially erroneous heuristic selection is performed in many studies, limiting the clinical utility of CS-MRI.

Several unsupervised methods have been proposed to address parameter selection in CS-MRI. Empirical methods including the L-curve criterion (LCC) follow the notion that the optimal parameter should be selected to attain a favorable trade-off between data consistency and regularization objectives [71, 72, 166]. Assuming this trade-off is characterized by an L-shaped curve, LCC selects the parameter on the point of maximum curvature [223]. LCC has been successfully demonstrated for parameter selection in several applications including parallel imaging [123, 221], quantitative susceptibility mapping [13], and diffusion spectrum imaging [12]. However, curvature assessment is computationally inefficient and typically sensitive to numerical perturbation and nonlinearities in the reconstruction problem [210, 110, 61].

Alternatively, parameters can be selected based on analytical estimates of the reconstruction error to optimize the regularization parameters. These methods include generalized cross-validation (GCV) [62], and methods based on Stein’s unbiased risk estimator (SURE) [191, 223]. In GCV, an analytical measure for

reconstruction error is estimated that asymptotically converges to the true error [62]. The GCV measure is derived as a function of the sampling pattern, regularization function, and regularization parameter. Parameter estimation via minimization of the GCV measure has been used in a variety of applications such as functional MRI [26], perfusion imaging [188], and dynamic MRI [189]. However, the GCV measure can be expensive to compute and yields biased estimates of the true error with limited number of data samples [161].

A recent approach instead uses the SURE criterion to estimate the expected value of the mean-square error (MSE) of the reconstruction. Given a specific parameter value and an estimate of the noise variance, Stein’s lemma [191] is used to compute online estimates of MSE. Subsequently, a line search over potential parameter values is performed for selecting the optimal parameter at each iteration. SURE-based parameter selection has produced promising results in several sparse recovery applications including CS-MRI [125, 19, 203, 107, 45, 68]. Unfortunately, parameter searches that need to be performed in each iteration cause substantial computational burden.

Here we introduce a computationally efficient self-tuning reconstruction method, named PESCaT (Projection onto Epigraph Sets for reconstruction by Calibration over Tensors), that can handle both single-acquisition and multi-acquisition datasets. To jointly reconstruct undersampled acquisitions, PESCaT performs tensor-based interpolation across acquired data, complemented by sparsity regularization of wavelet coefficients and TV regularization of image coefficients.

Since wavelet coefficients show varying sparsity across subbands and decomposition levels, PESCaT uses different ℓ_1 regularization parameters for each subband and level. Similarly, multi-coil multi-acquisition image coefficients may show varying spatial gradients, so different TV regularization parameters are used for each coil and acquisition. Parameters are efficiently tuned via simple geometric projections onto the boundary of the convex epigraph sets for the ℓ_1 - and TV-norm functions. This formulation transforms the selection of many

different regularization parameters for multiple subbands, levels, coils, and acquisitions into the selection of two scaling factors for the ℓ_1 -norm and TV-norm epigraphs. These factors can be reliably tuned on training data, yielding consistent performance across sequences, acceleration factors, and subjects. Comprehensive demonstrations on simulated brain phantoms, and in vivo balanced steady-state free-precession (bSSFP), T1-weighted, and angiographic acquisitions indicate that PESCaT enables nearly an order of magnitude improvement in computational efficiency compared to SURE-based methods, without compromising reconstruction quality.

3.2 Theory

Our main aim is to introduce a fast joint reconstruction method that automatically selects the free parameters for regularization terms based on ℓ_1 - and TV-norms. We consider the application of this self-tuning reconstruction to single-coil multi-acquisition, multi-coil single-acquisition, and multi-coil multi-acquisition MRI datasets. In the following sections, we introduce the regularized reconstruction problem, and its solution via projection onto epigraph sets for unsupervised parameter selection.

3.2.1 Reconstruction by calibration over tensors

Compressive sensing (CS) techniques proposed for static MRI acquisitions typically leverage encoding information provided either by multiple coils [126, 18, 121] or by multiple acquisitions [97, 11, 87] to enable recovery of unacquired data samples. Yet, simultaneous use of information across coils and acquisitions can benefit phase-cycled bSSFP [16, 82], multi-contrast [63, 14] or parametric imaging [207, 224, 222]. Here we consider a joint reconstruction framework for multi-coil, multi-acquisition datasets, based on a recently proposed method named ReCaT (Reconstruction by Calibration over Tensors) [16]. ReCaT rests on the following spatial encoding model for the signal measured in acquisition $n \in [1, \dots, N]$ and

coil $d \in [1, \dots, D]$:

$$S_{nd}(r) = P_n(r)C_d(r)S_0(r) \quad (3.1)$$

where r is the spatial location, P_n is the acquisition spatial profile, C_d is the coil sensitivity profile, and S_0 is the signal devoid of coil sensitivity and acquisition profile modulations. ReCaT seeks to linearly synthesize missing k-space samples from neighboring acquired samples across all coils and acquisitions. A tensor interpolation kernel is used for this purpose:

$$x_{nd} = \sum_{i=1}^N \sum_{j=1}^D t_{ij,nd}(k_r) \otimes x_{ij}(k_r) \quad (3.2)$$

where x_{nd} is the k-space data from n^{th} acquisition and d^{th} coil, k_r is the k-space location, and \otimes is the convolution operation. Here $t_{ij,nd}(k_r)$ accounts for the contribution of samples from acquisition i and coil j to x_{nd} . Equation 3.2 can be compactly expressed as:

$$x = \mathcal{T}x \quad (3.3)$$

3.2.2 PESCaT

ReCaT considers a basic implementation that does not include any regularization terms to enforce sparsity [16]. Here we introduce PESCaT, that incorporates sparsity and TV penalties:

$$\begin{aligned} \min_{x_{nd}} \left\{ \sum_{n=1}^N \sum_{d=1}^D \|\mathcal{T}x_{nd}\|_2^2 \right. \\ + \sum_{l=1}^L \sum_{s=1}^3 \lambda_{\ell_1,ls} \sum_{n=1}^N \sum_{d=1}^D \|\Psi_{ls}\{\mathcal{F}^{-1}\{x_{nd}\}\}\|_1 \\ \left. + \sum_{n=1}^N \sum_{d=1}^D \lambda_{TV,nd} \|\mathcal{F}^{-1}\{x_{nd}\}\|_{TV} \right\} \quad (3.4) \end{aligned}$$

where Ψ_{ls} is the wavelet operator for subband s and level l , \mathcal{I} is the identity operator, and \mathcal{F}^{-1} is the inverse Fourier operator. A separate ℓ_1 -regularization

parameter, $\lambda_{\ell_1,ls}$, is prescribed for each subband and level of the wavelet coefficients. Sparsity regularization is performed on the three high-pass subbands while the low-pass subband is kept intact to avoid over-smoothing. Meanwhile, a separate TV regularization parameter, $\lambda_{TV,nd}$, is used for each acquisition and coil. Because wavelet coefficients are aggregated across the coil and acquisition dimensions, $\lambda_{\ell_1,ls}$ varies across wavelet levels and subbands but it is fixed across coils or acquisitions.

Here, we implemented PESCaT in a constrained optimization formulation equivalent to the Lagrangian formulation in Eq. 3.4:

$$\begin{aligned}
& \min_{x_{nd}} && \sum_{n=1}^N \sum_{d=1}^D \|(\mathcal{T} - \mathcal{I})x_{nd}\|_2^2 \\
& \text{subject to} && \sum_{n=1}^N \sum_{d=1}^D \|\Psi_{ls}\{\mathcal{F}^{-1}\{x_{nd}\}\}\|_1 \leq \epsilon_{\ell_1,ls} \\
& && s = 1, 2, 3 \\
& && l = 1, \dots, L; \\
& && \|\mathcal{F}^{-1}\{x_{nd}\}\|_{TV} \leq \epsilon_{TV,nd} \\
& && n = 1, \dots, N \\
& && d = 1, \dots, D
\end{aligned} \tag{3.5}$$

where $\epsilon_{\ell_1,ls}$ are the constraints on the sparsity of the reconstruction, and $\epsilon_{TV,nd}$ are the constraints on the TV of the reconstruction. The optimization problem in Eq. 3.5 was solved via an alternating projections onto sets algorithm. As outlined in Fig. 3.1, this algorithm involves three consecutive projections, namely data-consistent calibration, sparsity, and TV projections. The calibration projection linearly synthesized unacquired k-space samples via the tensor interpolating kernel. To perform this projection while enforcing strict consistency to acquired data, an iterative least-squares algorithm was employed [126]. The sparsity projection jointly projected wavelet coefficients of images onto the epigraph set of the ℓ_1 -norm function.

The TV projection projected image coefficients onto the epigraph set of the

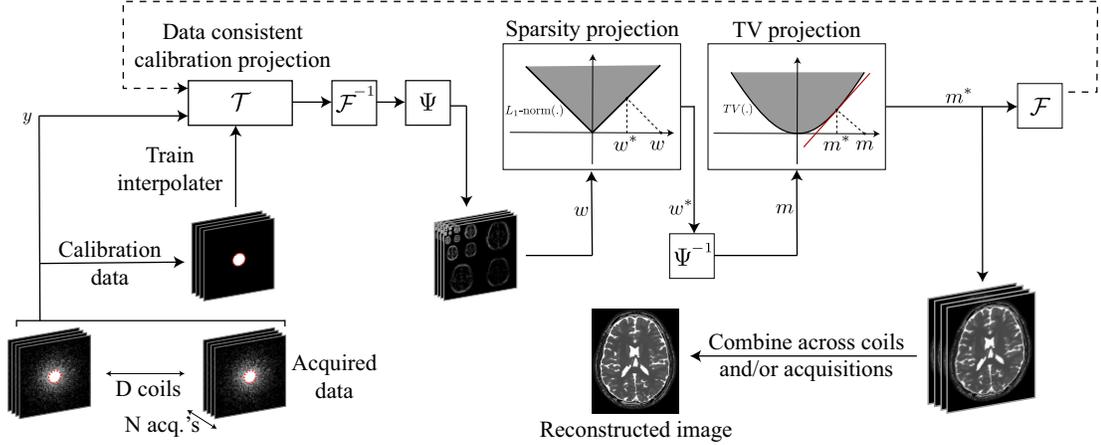


Figure 3.1. Flowchart of the PESCαT reconstruction. PESCαT employs an alternating projections onto sets approach with three subprojections: data-consistent calibration projection, sparsity projection, and TV projection. The calibration projection linearly synthesizes unacquired k-space samples via a tensor interpolating kernel. The sparsity projection jointly projects wavelet coefficients of the multi-coil, multi-acquisition images onto the epigraph set of the ℓ_1 -norm function. The TV projection projects image coefficients onto the epigraph set of the TV-norm function. These projections are performed iteratively until convergence. Lastly, reconstructed images are combined across multiple coils and/or acquisitions.

TV-norm function. These projections were performed iteratively until convergence. At each iteration, MSE between the reconstructed image in the current iterate and the previous iterate was first measured, and the percentage change in MSE across consecutive iterations was then calculated. Convergence was taken to be the iteration at which the percentage change in MSE fell below 20%. Lastly, reconstructed images were combined across multiple coils and/or acquisitions. Note that because PESCαT is structured modularly regarding the calibration, sparsity, and TV projections, it is trivial to implement variants that only employ sparsity or TV regularization.

3.2.3 Parameter tuning by projection onto epigraph sets

Careful tuning of constraint parameters in Eq. 3.5 is critical for a successful reconstruction. Selecting too tight constraints can lead to loss of important image features, whereas selecting too loose constraints will yield substantial residual

noise and aliasing. When only a few parameters are to be tuned, an exhaustive search over a relevant range of values followed by visual inspection is typically exercised. However, even in a modest dataset with $D = 4$ coils and $N = 4$ acquisitions, and assuming $L = 4$ wavelet decomposition levels there are 28 distinct parameters involved in Eq. 3.5. Thus, the exhaustive search approach is impractical.

Here, we perform self-tuning of the constraint parameters in Eq. 3.5 via projections onto epigraph sets of the respective regularization terms. Let $\mathcal{U} \in \mathbb{R}^k$ be a closed convex set, $\Phi : \mathbb{R}^k \rightarrow \mathbb{R}$ be a convex function (e.g., ℓ_1 -norm and TV-norm functions), and $\hat{u} \in \mathbb{R}^k$ be an input vector (e.g., wavelet coefficients for ℓ_1 -norm or image coefficients for TV-norm). The proximal operator of Φ^2 is:

$$\text{prox}_{\Phi^2}(\hat{u}) = \arg \min_{u \in \mathcal{U}} \|\hat{u} - u\|_2^2 + \Phi^2(u) \quad (3.6)$$

where u is the auxiliary variable. We prefer to use Φ^2 here since it allows us to express the solution as a simple geometric projection. Specifically, the problem in Eq. 3.6 can be stated in vector form by mapping onto \mathbb{R}^{k+1} :

$$\min_{u \in \mathcal{U}} \left\| \begin{bmatrix} \hat{u} \\ 0 \end{bmatrix} - \begin{bmatrix} u \\ \Phi(u) \end{bmatrix} \right\|_2^2 \quad (3.7)$$

Here we propose to implement the proximal operator in Eq. 3.6 by identifying the closest vector $[u^* \ \Phi(u^*)]^T \in \mathbb{R}^{k+1}$ to $[\hat{u} \ 0]^T$. This solution can be shown to be equivalent to the orthogonal projection of the vector $[\hat{u} \ 0]^T$ onto the epigraph set of Φ (epi_{Φ}) defined as:

$$\text{epi}_{\Phi} = \left\{ \begin{bmatrix} u \\ z \end{bmatrix}^T : z \geq \Phi(u) \right\} \quad (3.8)$$

where z denotes an upper bound for the function $\Phi(u)$. The projection onto epi_{Φ} is the closest solution to \hat{u} that lies on the boundary of the epigraph set. Since the epigraph set of a convex function is also convex, this projection will yield the global optimum solution. Note that projections onto the epigraph set will yield the solution of the proximal operator only if the search space of the proximal operator is a convex set $\mathcal{U} \in \mathbb{R}^k$ [29]. In practice, a family of solutions can be obtained by introducing a scaling parameter to alter the size of the epigraph set:

$$\text{epi}_{\Phi'} = \left\{ \begin{bmatrix} u \\ z \end{bmatrix}^T : z \geq \beta_{\Phi} \Phi(u) \right\} \quad (3.9)$$

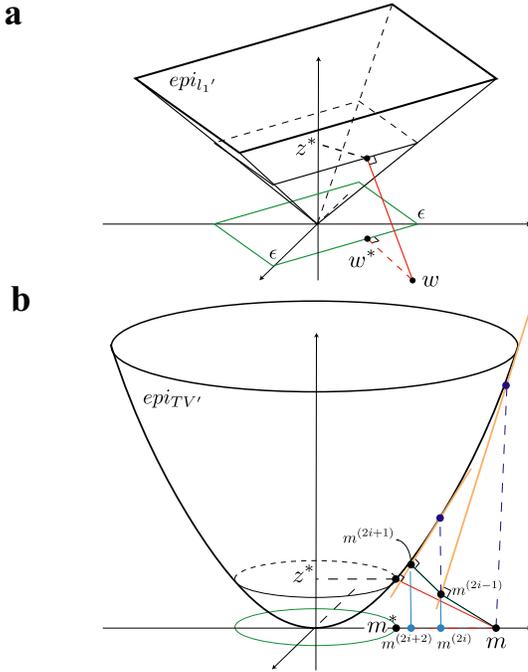


Figure 3.2. The projection onto epigraph sets (PES) approach illustrated in \mathbb{R}^3 .

(a) PES for ℓ_1 -regularization. An input vector w (e.g., vector of wavelet coefficients of image m) is projected onto the epigraph set of the ℓ_1 -norm function (epi_{ℓ_1}). This projection results in the output $[w^* z^*]^T$, thereby inherently calculating the projection of w onto the ℓ_1 -ball in \mathbb{R}^2 (w^*). The size of the ℓ_1 -ball (ϵ) depicted in green color depends on z^* . (b) PES for TV regularization. Unlike PES- ℓ_1 , PES-TV has no closed-form solution, and is instead implemented via an iterative epigraphical splitting procedure. At the i^{th} iteration, the input vector $m^{(2i-1)}$ is projected onto the supporting hyperplane (orange line), resulting in $m^{(2i+1)}$. This intermediate vector is then projected on the level set to compute $m^{(2i+2)}$. Through successive iterations the output gradually converges to the desired projection point on the epigraph set $[m^* z^*]^T$, thereby inherently calculating the projection of m onto the TV-ball in \mathbb{R}^2 (m^*). The size of the TV-ball depends on z^* .

Here, β_Φ serves to control the allowed degree of deviation of u^* from \hat{u} . Note that both z^* and u^* are computed via an orthogonal projection of the input onto epi'_Φ . Since the scales of z^* and u^* vary proportionately to the scale of \hat{u} , β_Φ can be described in absolute terms. $\beta_\Phi > 1$ scales down the epigraph set, resulting in a solution u^* that deviates further from \hat{u} . Meanwhile, $0 < \beta_\Phi < 1$ scales up epi_Φ , resulting in a solution u^* that is closer to \hat{u} , where $u^* = \hat{u}$ as $\beta_\Phi \rightarrow 0$. To obtain more conservative solutions, here we used $0 < \beta_\Phi < 1$ for both sparsity and TV projections. The resulting projection point determines both the size of the Φ -ball in \mathbb{R}^k (i.e. ℓ_1 -ball or TV-ball, see Fig. 3.2) and the actual projection onto the ball. Hence, the proximal operator in Eq. 3.6 enables assessing the optimal constraint parameters in Eq. 3.5 using the input vector \hat{u} as explained below.

3.2.3.1 Self-tuning sparsity projection

The sparsity projections were implemented using projections onto the epigraph set of the ℓ_1 -norm function, applied on wavelet coefficients. The image coefficients $m_{nd} = \mathcal{F}^{-1}\{x_{nd}\}$ are obtained by inverse Fourier transformation of k-space data, x_{nd} , for acquisition n and coil d . The wavelet coefficients for m_{nd} are then given by $w_{ls,nd} = \Psi_{ls}\{m_{nd}\}$ at subband s and level l , and w_{ls} denotes the aggregate vector pooling $w_{ls,nd}$ across coils and acquisitions. Assuming $\hat{w} = w_{ls}$ is the input vector, the proximal formulation in Eq. 3.6 becomes:

$$\text{prox}_{\ell_1^2}(\hat{w}) = \arg \min_u \|\hat{w} - u\|_2^2 + \|u\|_1^2 \quad (3.10)$$

The solution to Eq. 3.10 is then obtained by projecting $\begin{bmatrix} w_{ls} & 0 \end{bmatrix}^T$ onto the scaled epigraph set (see Fig. 3.2a):

$$\text{epi}_{\ell_1} = \left\{ \begin{bmatrix} u & z \end{bmatrix}^T \in \mathbb{R}^{k+1} : z \geq \beta_{\ell_1} \|u\|_1 \right\} \quad (3.11)$$

where β_{ℓ_1} denotes the epigraph scaling factor for the ℓ_1 -norm. As demonstrated in Fig. 3.2a, the closest orthogonal projection of $\begin{bmatrix} w_{ls} & 0 \end{bmatrix}^T$ onto the epigraph set lies on the boundary of epi_{ℓ_1} . For the simple case of \mathbb{R}^2 ($k = 1$), $\begin{bmatrix} w_{ls} & 0 \end{bmatrix}^T$ is projected onto the $z = \beta_{\ell_1} |u|$ line, yielding $z_{ls}^* = \frac{\beta_{\ell_1} |w_{ls}|}{\beta_{\ell_1}^2 + 1}$. It can be shown that for arbitrary k , the z -intercept is:

$$z_{ls}^* = \frac{\beta_{\ell_1} \|w_{ls}\|_1}{\beta_{\ell_1}^2 k + 1} \quad (3.12)$$

The value of the z -intercept also determines the size of the respective ℓ_1 -ball, $B_{\ell_1,ls} = \{u \in \mathbb{R}^k : \|u\|_1 \leq \epsilon_{\ell_1,ls}\}$, as:

$$\epsilon_{\ell_1,ls} = \frac{z_{ls}^*}{\beta_{\ell_1}} \quad (3.13)$$

Therefore, w_{ls}^* can be computed by finding the projection of w_{ls} onto the ℓ_1 -ball of size $\epsilon_{\ell_1,ls}$. To efficiently implement this projection, we used a soft-thresholding operation [106]:

$$w_{ls}^* = e^{i\angle w_{ls}} \max(|w_{ls}| - \theta_{ls}, 0) \quad (3.14)$$

where magnitudes of wavelet coefficients are subjected to a threshold of θ_{ls} , and phases of coefficients are individually restored via $e^{i\angle w_{ls}}$. We propose to determine the value of θ_{ls} given $\epsilon_{\ell_1,ls}$ using an efficient ranking algorithm [48]. The proposed algorithm first sorts the absolute values of the wavelet coefficients $w_{ls,nd}$ to attain a rank-ordered sequence $\{\mu_i\}_{i=1}^k$ where $\mu_1 > \mu_2 > \dots > \mu_k$. This sequence is then analyzed to find the threshold that approximately yields a resultant ℓ_1 -norm of value $\epsilon_{\ell_1,ls}$ in the thresholded coefficients:

$$\begin{aligned} \rho_{ls} &= \max\{j \in \{1, 2, \dots, k\} : \mu_j - \frac{1}{j} \left(\sum_{r=1}^j \mu_r - \epsilon_{\ell_1,ls} \right) > 0\}, \\ \theta_{ls} &= \frac{1}{\rho_{ls}} \left(\sum_{n=1}^{\rho_{ls}} \mu_n - \epsilon_{\ell_1,ls} \right). \end{aligned} \quad (3.15)$$

Note that the determined threshold directly translates to $\lambda_{\ell_1,ls}$ in Eq. 3.4 by [153]:

$$\lambda_{\ell_1,ls} = 2\theta_{ls} \quad (3.16)$$

Projections were separately performed for each subband s at each wavelet decomposition level l to determine the respective w_{ls}^* , and $\epsilon_{\ell_1,ls}$. Since wavelet coefficients were pooled across coils and acquisitions, parameter selection is performed jointly across coils and acquisitions. Since the only free parameter in the proposed method is the epigraph scaling constant β_{ℓ_1} , the selection of $3 \times L$ parameters in Eq. 3.5 are transformed into the selection of a single parameter. Here, the optimal β_{ℓ_1} was empirically determined in a group of training subjects and then used to obtain reconstructions in held-out test subjects.

3.2.3.2 Self-tuning TV projection

The TV projections were implemented using projections onto the epigraph set of the TV-norm function, applied on image coefficients. Letting $\hat{m} = m_{nd}$ be the input vector, the proximal formulation in Eq. 3.6 becomes:

$$\text{prox}_{TV^2}(\hat{m}) = \arg \min_u \|\hat{m} - u\|_2^2 + \|u\|_{TV}^2 \quad (3.17)$$

The solution to Eq. 3.17 is then obtained by projecting $\begin{bmatrix} m_{nd} & 0 \end{bmatrix}^T$ onto the scaled epigraph set (see Fig. 3.2b):

$$epi_{TV'} = \left\{ \begin{bmatrix} u & z \end{bmatrix}^T \in \mathbb{R}^{k+1} : z \geq \beta_{TV} \|u\|_{TV} \right\} \quad (3.18)$$

where β_{TV} denotes the epigraph scaling factor for the TV-norm. Unlike the projection onto the ℓ_1 -norm epigraph, projection onto generic epigraph sets (including TV-norm epigraph) does not have a closed-form solution. As demonstrated in Fig. 3.2b, PESCaT uses an iterative epigraphical splitting method to perform the projection efficiently [198]. In the initial step of this approach, complex-valued $m^{(0)} = m_{nd}$ is projected onto the supporting hyperplane of $epi_{TV'}$ at $\begin{bmatrix} m_{nd} & \beta_{TV} \|m_{nd}\|_{TV} \end{bmatrix}^T$ resulting in $m^{(1)}$. The supporting hyperplane is determined by evaluating the gradient of the epigraph surface. In the following step, $m^{(1)}$ is projected onto the level set, $L_{TV} = \left\{ \begin{bmatrix} u & z \end{bmatrix}^T : z \leq 0 \right\}$, by forcing the last element of $m^{(1)}$ to zero. This projection yields the next estimate $m^{(2)}$. These two projections are iterated. Note that all steps of the splitting procedure are performed in complex domain, thereby, regularizing magnitude and phase channels simultaneously. Previous studies have shown that the second derivative of distance between the input vector and the projections on the supporting hyperplanes ($\|m_{nd} - m^{(2i+1)}\|_2$) is negative as the projections approach to the true projection solution and is positive as the projections deviate from it [29]. Thus, in case of a sign change in the second derivative a refinement step is performed, where $m^{(2i)}$ is projected onto the supporting hyperplane at $\frac{m^{(2i+1)} + m^{(2i-1)}}{2}$. This heuristic approach has been shown to converge to the global solution for TV projections [198]. Note that the projection uniquely specifies the z-intercept, z_{nd}^* . Hence, the size of the corresponding TV-ball, $B_{TV,nd} = \{u \in \mathbb{R}^k : \|u\|_{TV} \leq \epsilon_{TV,nd}\}$, can be calculated as:

$$\epsilon_{TV,nd} = \frac{z_{nd}^*}{\beta_{TV}} \quad (3.19)$$

Note that it is nontrivial to explicitly express $\lambda_{TV,nd}$ in Eq. 3.4 in terms of $\epsilon_{TV,nd}$ in Eq. 3.5. Yet, constraining $\epsilon_{TV,nd}$ implicitly enforces a set of regularization parameters $\lambda_{TV,nd}$.

Projections were separately performed for each acquisition n and coil d to determine the respective m_{nd}^* and $\epsilon_{TV,nd}$. Since the only free parameter is the

epigraph scaling constant β_{TV} , the selection of $N \times D$ parameters in Eq. 3.5 is transformed into the selection of a single parameter. Here, the optimal β_{TV} was empirically determined in a group of training subjects and then used to obtain reconstructions in held-out test subjects.

All reconstruction algorithms were executed in MATLAB (MathWorks, MA). The implementations used libraries from the SPIRiT toolbox [126]. The PESCiT algorithm is available for general use at <http://github.com/icon-lab/mrirecon>.

3.3 Methods

3.3.1 Alternative reconstructions

To demonstrate the performance of PESCiT, we compared it against several alternative reconstructions that aim to select regularization parameters.

3.3.1.1 Self-tuning regularized ReCaT (ReCaT_{SURE})

A recently previously proposed multi-coil multi-acquisition method, ReCaT, did not include any regularization parameters [16]. We have implemented a variant of ReCaT incorporating sparsity and TV regularization terms where the regularization parameters are automatically selected using the data. This reconstruction method iteratively synthesizes unacquired data as weighted combinations of collected data across coils and/or acquisitions. It uses sparsity and TV projections to enforce image sparsity. At each iteration, the regularization parameter for the sparsity term is selected based on the SURE criterion. The regularization parameter for the TV term is selected based on the local standard deviation of the reconstructed image from the previous iteration.

An alternating projections onto sets algorithm was used in ReCaT_{SURE} to solve

the reconstruction problem cast in Eq. 3.4. ReCaT_{SURE} used a single sparsity regularization parameter for all subbands and levels of wavelet coefficients. The sparsity regularization parameter was determined via a line search over the range $[2 \times 10^{-5}, 2 \times 10^{-1}]$. The TV regularization parameter was taken as one-third of the median local standard deviation [96]. All remaining reconstruction parameters were kept identical to PESCiT.

3.3.1.2 ReCaT_{SURE} with early stop

The projections performed in each iteration of PESCiT do not involve any line searches, and therefore they are more efficient compared to ReCaT_{SURE} . To enable a fair comparison, we implemented a variant of ReCaT_{SURE} that was stopped once the total reconstruction time reached that of PESCiT. All reconstruction parameters except the total number of iterations were kept identical to ReCaT_{SURE} .

3.3.1.3 ReCaT with empirically-tuned parameters (ReCaT_{fixed})

To demonstrate the effects of prescribing separate regularization parameters for different subbands/levels or coils/acquisitions in PESCiT, we implemented a variant of ReCaT with a single sparsity parameter across all subbands/levels and a single TV parameter across all coils/acquisitions. Similar to PESCiT, this reconstruction method iteratively synthesizes unacquired data as weighted combinations of collected data across coils and/or acquisitions. ReCaT_{fixed} was tuned using held-out data. The sparsity and TV parameters were independently varied across a broad range $[10^{-5}, 0.5]$. Separate reconstructions were obtained for each parameter set, and reconstruction quality was taken as peak signal-to-noise ratio (PSNR) between the reconstructed image and the fully-sampled reference image. The parameter set that yielded the maximum PSNR was selected. Sparsity and TV parameters were fixed across iterations. All remaining reconstruction parameters were kept identical to PESCiT.

3.3.1.4 Brute-force reconstruction

To evaluate the success of PESCiT in selecting the optimal parameters, a brute-force reconstruction was implemented to solve the problem in Eq. 3.4. The brute-force method used a constant set of regularization parameters across iterations. The sparsity and TV parameters were independently varied across the range $[10^{-5}, 0.5]$. Separate reconstructions were obtained for each parameter set, and reconstruction quality was taken as PSNR between the reconstructed image and the fully-sampled reference image. The parameter set that yielded the maximum PSNR was selected. All remaining reconstruction parameters were kept identical to PESCiT.

3.3.1.5 ESPIRiT with PES parameter tuning (PESSPIRiT)

To compare the performance of PESCiT against conventional parallel imaging, we implemented a variant of ESPIRiT [200] that included sparsity and TV regularization terms. Similar to ESPIRiT, this method iteratively reconstructs images based on coil sensitivities estimated from central calibration data. In each iteration, the sparsity and TV regularization parameters were tuned using PES. Two other variants, PESSPIRiT with only the sparsity regularization (PESSPIRiT $_{\ell_1}$) and PESSPIRiT with only the TV regularization (PESSPIRiT $_{TV}$) were also implemented. In all variants, the stopping criterion was identical to PESCiT to enable a fair comparison.

3.3.2 Simulations

Simulations were performed using a realistic brain phantom at 0.5 mm isotropic resolution (<http://www.bic.mni.mcgill.ca/brainweb>). Phase-cycled bSSFP signals were assumed with T1/T2: 3000/1000 ms for cerebrospinal fluid, 1200/250 ms for blood, 1000/80 ms for white matter, 1300/110 ms for gray matter, 1400/30 ms for muscle, and 370/130 ms for fat [97]. Single-coil three-dimensional (3D)

Table 3.1. Peak signal-to-noise ratio (PSNR) for simulated phantom

	R=2	R=4	R=6
Brute-force	<i>30.29±0.24</i>	<i>28.16±0.27</i>	<i>27.47±0.22</i>
PESCaT	29.56±0.34	26.68±0.24	26.25±0.19
ReCaT_{SURE}	29.44±0.21	25.52±0.23	24.72±0.16
Early stop	27.67±0.26	24.96±0.23	23.66±0.21

PSNR was measured between the reconstructed image and a fully-sampled reference image. Measurements were obtained for brute-force, PESCaT, ReCaT_{SURE} and ReCaT_{SURE} with early stop methods. Results are reported as mean±std across five cross-sections.

acquisitions were assumed with TR/TE=5.0/2.5 ms, flip angle=45°, and phase-cycling increments $\Delta\phi=2\pi\frac{[0:1:N-1]}{N}$. We used a simulated field inhomogeneity distribution corresponding to an off-resonance shift with zero mean and 62 Hz standard deviation. A bivariate Gaussian noise was added to simulated acquisitions to attain signal-to-noise ratio (SNR)=20, where SNR was taken as the ratio of the mean power in the phantom image to the noise variance. Data were undersampled by a factor (R) of 2, 4, and 6 in the two phase-encode directions using disjoint, variable density random undersampling [36] and normalized so that zero-filled density compensated k-space data had unity norm [127]. Reconstruction quality was taken as PSNR between reconstructions and a fully-sampled reference. To prevent bias, the 98th percentile of image intensities were adjusted to [0, 1]. PSNR values were then averaged across five central axial cross-sections.

To examine the effect of noise on optimal regularization parameters, we performed experiments on the simulated brain phantom where the noise level was systematically varied. The simulations output single-coil single-acquisition brain images with SNR varying in the range [5, 25]. Data were undersampled by R=2, 4, and 6 in the two phase-encode directions using disjoint, variable density random undersampling. Multiple separate reconstructions were obtained for each undersampled dataset via ReCaT_{fixed}, while ℓ_1 and TV regularization parameters were independently varied in the range [0.001, 0.1]. At each SNR level, fully-sampled data were used as reference. PSNR was measured between the reconstructions and the reference. The optimal regularization parameters were selected according to PSNR.

To examine the reliability of the epigraph scaling parameters against noise, reconstructions of the brain phantom were obtained at three separate levels of SNR = 10, 18, 25. Meanwhile, β_{ℓ_1} was varied in the range [0.05, 0.6] and β_{TV} was varied in the range [0.1, 1]. To examine the reliability of the epigraph scaling parameters against variations in the level of detail and spatial resolution, we performed experiments on a simulated numerical phantom. A circular phantom of radius 125 voxels (for a 256×256 field of view) was designed with the background resembling muscle tissue and vertical bright bars of width 12 and height [190, 220, 238, 238, 220, 190] voxels resembling blood vessels (Fig. 3.3). Phase-cycled bSSFP signals were assumed with T1/T2: 870/47 ms for muscle, and 1273/259 ms for blood. Three dimensional acquisitions were assumed with TR/TE=4.6/2.3 ms, flip angle=60°, and phase-cycling increments $\Delta\phi=2\pi \frac{[0:1:N-1]}{N}$. A simulated field inhomogeneity distribution corresponding to an off-resonance shift with zero mean and 62 Hz standard deviation was used. Level of detail was varied from low to high by incrementally placing [1, 3, 6] vertical bars in the phantom. Spatial resolution was varied from low to high by low-pass filtering k-space data to select circular regions of radius [20, 55, 125] voxels. Reconstructions were obtained while β_{ℓ_1} and β_{TV} were varied in the range [0.05, 0.5].

3.3.3 In vivo experiments

Experiments were performed to acquire 3D multi-coil multi-acquisition phase-cycled bSSFP, and multi-coil single-acquisition T1-weighted and time-of-flight (ToF) angiography data in the brain. Data were collected on a 3T Siemens Magnetom scanner (maximum gradient strength of 45 mT/m and slew rate of 200 T/m/s). bSSFP and ToF data were collected using a 12-channel receive-only head coil that was hardware compressed to 4 channels.

T1-weighted data were collected using a 12-channel receive-only head coil. Separate bSSFP datasets were also collected using a 32-channel head coil. Balanced SSFP data were acquired using a bSSFP sequence with the following parameters: flip angle=30°, TR/TE=8.08/4.04 ms, field-of-view (FOV)=218 mm \times 218

Table 3.2. PSNR and NRMSE for in vivo bSSFP dataset

	R=2		R=4		R=6	
	PSNR	NRMSE $\times 10^3$	PSNR	NRMSE $\times 10^3$	PSNR	NRMSE $\times 10^3$
Brute-force	<i>44.31\pm0.72</i>	<i>7.31\pm0.23</i>	<i>40.21\pm0.78</i>	<i>11.18\pm0.45</i>	<i>37.62\pm0.61</i>	<i>16.79\pm0.56</i>
PESCaT	43.93 \pm 0.65	8.16 \pm 0.27	39.64\pm0.61	11.93\pm0.43	36.72\pm0.29	18.36\pm0.34
ReCaT_{fixed}	44.11\pm0.62	7.57\pm0.33	39.08 \pm 0.47	12.59 \pm 0.48	36.51 \pm 0.48	18.45 \pm 0.95
ReCaT_{SURE}	42.20 \pm 0.78	10.09 \pm 0.59	37.89 \pm 0.76	16.19 \pm 0.79	35.15 \pm 0.59	20.44 \pm 0.82
Early stop	41.83 \pm 0.66	10.37 \pm 0.52	36.82 \pm 0.63	18.18 \pm 0.77	34.01 \pm 0.32	22.08 \pm 0.31
PESSPIRiT	43.37 \pm 0.44	8.80 \pm 0.35	38.39 \pm 0.59	14.37 \pm 0.84	35.06 \pm 0.35	20.56 \pm 0.71
PESSPIRiT_{ℓ_1}	41.11 \pm 0.62	10.84 \pm 0.62	35.96 \pm 0.47	19.36 \pm 0.86	33.15 \pm 0.55	23.05 \pm 1.40
PESSPIRiT_{TV}	42.41 \pm 0.67	9.61 \pm 0.59	35.57 \pm 0.68	19.88 \pm 1.35	32.24 \pm 0.61	24.28 \pm 1.69

PSNR and NRMSE were measured between the reconstructed image and a fully-sampled reference image. Measurements were obtained for brute-force, PESCaT, ReCaT_{SURE}, ReCaT_{SURE} with early stop, ReCaT_{fixed}, and variants of PESSPIRiT methods. Results are averaged across three subjects, and reported as mean \pm std across five cross-sections.

Table 3.3. PSNR and NRMSE for in vivo T1-weighted dataset

	R=2		R=4	
	PSNR	NRMSE $\times 10^3$	PSNR	NRMSE $\times 10^3$
Brute-force	<i>36.75\pm0.55</i>	<i>18.25\pm0.89</i>	<i>32.15\pm0.41</i>	<i>23.75\pm0.85</i>
PESCaT	35.62\pm0.95	19.86\pm0.99	31.44\pm1.09	27.75\pm1.61
ReCaT_{fixed}	35.27 \pm 0.67	20.02 \pm 0.85	30.89 \pm 0.63	29.59 \pm 1.31
ReCaT_{SURE}	35.02 \pm 0.93	20.64 \pm 0.94	30.67 \pm 0.93	30.10 \pm 1.57
Early stop	34.64 \pm 1.03	21.47 \pm 0.95	30.03 \pm 1.06	30.81 \pm 1.65
PESSPIRiT	35.21 \pm 0.74	20.14 \pm 1.10	29.92 \pm 0.63	31.66 \pm 1.60
PESSPIRiT_{ℓ_1}	31.75 \pm 0.66	27.37 \pm 1.42	27.39 \pm 0.80	34.37 \pm 2.72
PESSPIRiT_{TV}	34.93 \pm 0.69	21.04 \pm 1.03	29.48 \pm 0.60	31.83 \pm 1.61

PSNR and NRMSE were measured between the reconstructed image and a fully-sampled reference image. Measurements were obtained for brute-force, PESCaT, ReCaT_{SURE}, ReCaT_{SURE} with early stop, ReCaT_{fixed}, and variants of PESSPIRiT methods. Results are averaged across three subjects, and reported as mean \pm std across five cross-sections.

Table 3.4. PSNR and NRMSE for in vivo ToF dataset

	R=2		R=4	
	PSNR	NRMSE $\times 10^3$	PSNR	NRMSE $\times 10^3$
Brute-force	<i>36.57\pm1.61</i>	<i>18.61\pm0.71</i>	<i>33.08\pm1.55</i>	<i>22.75\pm0.97</i>
PESCaT	36.32\pm1.14	18.79\pm0.57	31.86\pm1.21	27.55\pm1.03
ReCaT_{fixed}	35.86 \pm 0.60	19.41 \pm 0.86	30.51 \pm 0.60	30.17 \pm 1.60
ReCaT_{SURE}	35.55 \pm 1.25	19.97 \pm 0.94	31.34 \pm 1.13	28.40 \pm 1.57
Early stop	35.46 \pm 1.12	20.14 \pm 0.95	27.69 \pm 1.19	28.71 \pm 1.65
PESSPIRiT	35.86 \pm 0.57	19.42 \pm 0.90	30.83 \pm 0.47	29.72 \pm 1.41
PESSPIRiT_{ℓ_1}	32.00 \pm 0.55	24.89 \pm 1.46	27.22 \pm 0.58	37.84 \pm 2.71
PESSPIRiT_{TV}	35.66 \pm 0.53	19.72 \pm 0.84	30.41 \pm 0.64	30.42 \pm 1.89

PSNR and NRMSE were measured between the reconstructed image and a fully-sampled reference image. Measurements were obtained for brute-force, PESCaT, ReCaT_{SURE}, ReCaT_{SURE} with early stop, ReCaT_{fixed}, and variants of PESSPIRiT methods. Results are averaged across three subjects, and reported as mean \pm std across five cross-sections.

mm, matrix size of $256 \times 256 \times 96$, resolution of $0.9 \text{ mm} \times 0.9 \text{ mm} \times 0.8 \text{ mm}$, right/left readout direction, and $N=8$ separate acquisitions with phase-cycling values in the range $[0, 2\pi)$ in equispaced intervals. Total acquisition time for the bSSFP sequence was 20:56. T1-weighted data were acquired using an MP-RAGE sequence with the parameters: flip angle= 9° , TR/TE= $2300/2.98$ ms, TI= 900 ms, FOV= $256 \text{ mm} \times 240 \text{ mm}$, matrix size of $256 \times 240 \times 160$, resolution of $1.0 \text{ mm} \times 1.0 \text{ mm} \times 1.2 \text{ mm}$, and superior/inferior readout direction. Total acquisition time for the MP-RAGE sequence was 9:14. ToF angiograms were acquired using a multiple overlapping thin-slab acquisition (MOTSA) sequence with parameters: flip angle= 18° , TR/TE= $38/3.19$ ms, FOV= $204 \text{ mm} \times 204 \text{ mm}$, matrix size of $256 \times 256 \times 75$, isotropic resolution of 0.8 mm , and anterior/posterior readout direction. Total acquisition time for the MOTSA sequence was 14:16. The imaging protocols were approved by the local ethics committee, and all six participants gave written informed consent.

Phase-cycled bSSFP acquisitions with 4 channels were retrospectively under-sampled at $R=2, 4$, and 6 . Following phase-cycles were selected: $\Delta\phi = 2\pi \frac{[0:1:N-1]}{N}$ for $N=2$ and 4 , and $[0, \frac{\pi}{2}, \frac{3\pi}{4}, \pi, \frac{5\pi}{4}, \frac{7\pi}{4}]$ for $N=6$. For this bSSFP dataset, $N=R$ was used. T1-weighted and ToF acquisitions were retrospectively undersampled at $R=2$ and 4 (note that in these cases $N=1$). Undersampling was performed across the two phase encode directions: superior/inferior and anterior/posterior for bSSFP, right/left and anterior/posterior for T1-weighted, superior/inferior and right/left for ToF. Data were normalized so that zero-filled density compensated k-space data had unity norm. Entire volumes were reconstructed, five axial cross-sections equispaced across the entire brain were selected for quantitative assessment. PSNR and normalized root mean-squared error (NRMSE) measurements were averaged across cross-sections.

To investigate the convergence behavior of PESCiT, we studied the evolution of the three cost terms in Eq. 3.4 separately (Fig. 3.4). Normalized cost terms associated with calibration consistency, sparsity, and TV terms at the end of each iteration were plotted across iterations. In all datasets, all cost terms diminish smoothly.

To optimize epigraph scaling constants for ℓ_1 - and TV-norm functions, PESCiT was performed on data acquired from three subjects reserved for this purpose. Volumetric reconstructions were performed at R=2, 4, and 6 for bSSFP datasets, and R=2 and 4 for T1-weighted and ToF datasets. Separate reconstructions were obtained while β_{ℓ_1} was varied in the range [0.1, 1], and β_{TV} was varied in the range [0.05, 0.6]. PSNR was measured between the reconstructed and fully-sampled reference images (Figs. 3.5, 3.6). Consistently across subjects and different types of datasets, PSNR values within 95% of the optimum value were maintained in the range $\beta_{\ell_1} = [0.1, 0.3]$, and $\beta_{TV} = [0.2, 0.4]$. Near-optimal PSNR values were attained around $\beta_{\ell_1} = 0.2$ and $\beta_{TV} = 0.3$. Thus, these scaling constants were prescribed for reconstructions thereafter.

To demonstrate the reconstruction performance of PESCiT at high acceleration rates, phase-cycled bSSFP acquisitions with 32 channels were analyzed. This bSSFP dataset was retrospectively undersampled at R= 8, 10 (where N=8). Entire volumes were reconstructed, and PSNR and NRMSE measurements were averaged across five axial cross-sections.

3.4 Results

3.4.1 Simulations

MRI data may show differential noise and structural characteristics for separate coils and acquisitions, or for separate wavelet subbands and levels. In turn, the optimal regularization parameters can vary across each of these dimensions. To test this prediction, we performed experiments on the simulated brain phantom, where the noise level was systematically varied and ReCaT_{fixed} reconstructions were performed. For both ℓ_1 and TV regularization, the optimal regularization parameters show a clear increasing trend as SNR is lowered (Fig. 3.7). These results suggest that prescribing a fixed parameter can cause performance loss when a good compromise cannot be achieved across subbands/levels or coils/acquisitions.

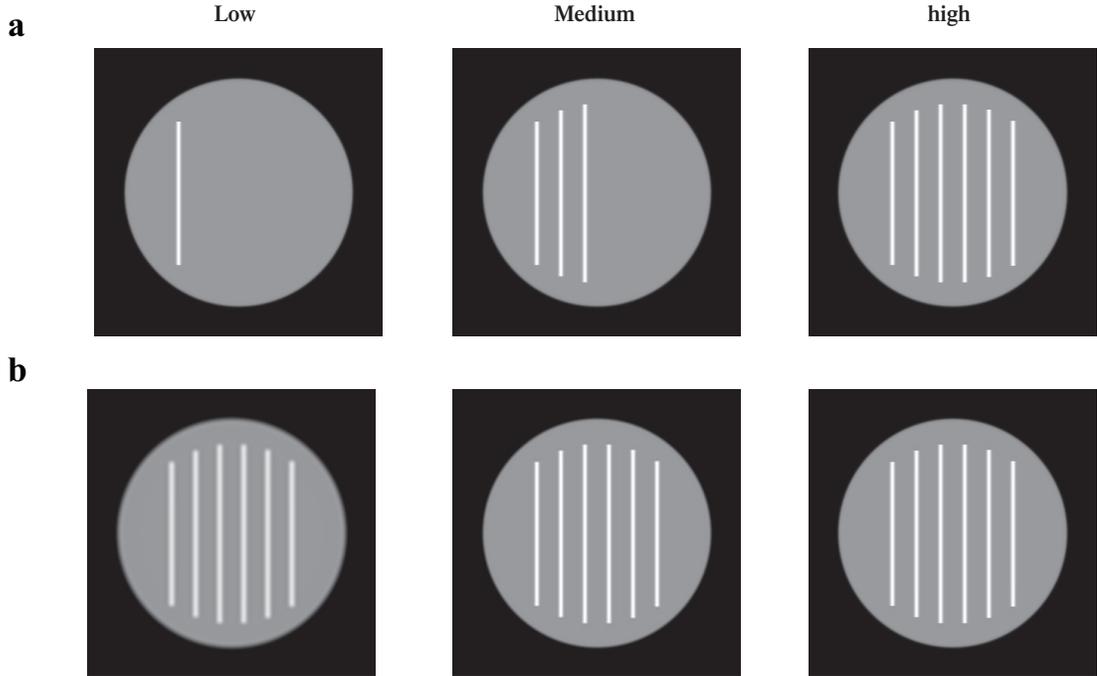


Figure 3.3. The numerical phantom designed for assessment of reliability against variation in level of detail and spatial resolution. A circular phantom of radius 125 voxels (for a 256x256 field of view) was designed with the background resembling muscle tissue and vertical bright bars of width 12 and height [190, 220, 238, 238, 220, 190] voxels resembling blood vessels. Phase-cycled bSSFP signals were assumed with T1/T2: 870/47 ms for muscle, and 1273/259 ms for blood. Level of detail was controlled by incrementally placing the vertical bars in the phantom. Spatial resolution was controlled by low-pass filtering k-space data. Level of detail (**a**), and resolution (**b**) of the phantom were varied from low (**left**) to medium (**middle**) to high (**right column**).

It can also render the reconstruction more susceptible to deviations from the optimal value of the regularization parameter.

In contrast, PESCaT uses only two global parameters to control the overall sparsity of the solutions in wavelet domain (β_{l_1}) and TV domain (β_{TV}). Given these scaling parameters, regularization parameters for individual subbands/levels and coils/acquisitions are determined adaptively in a data-driven manner. To examine the reliability of the scaling parameters against noise, reconstructions were obtained at varying SNR levels. The PSNR curves as a function of β_{l_1} and β_{TV} demonstrate substantial flatness, yielding near-optimal performance across the entire range of values examined (Fig. 3.9). To further examine

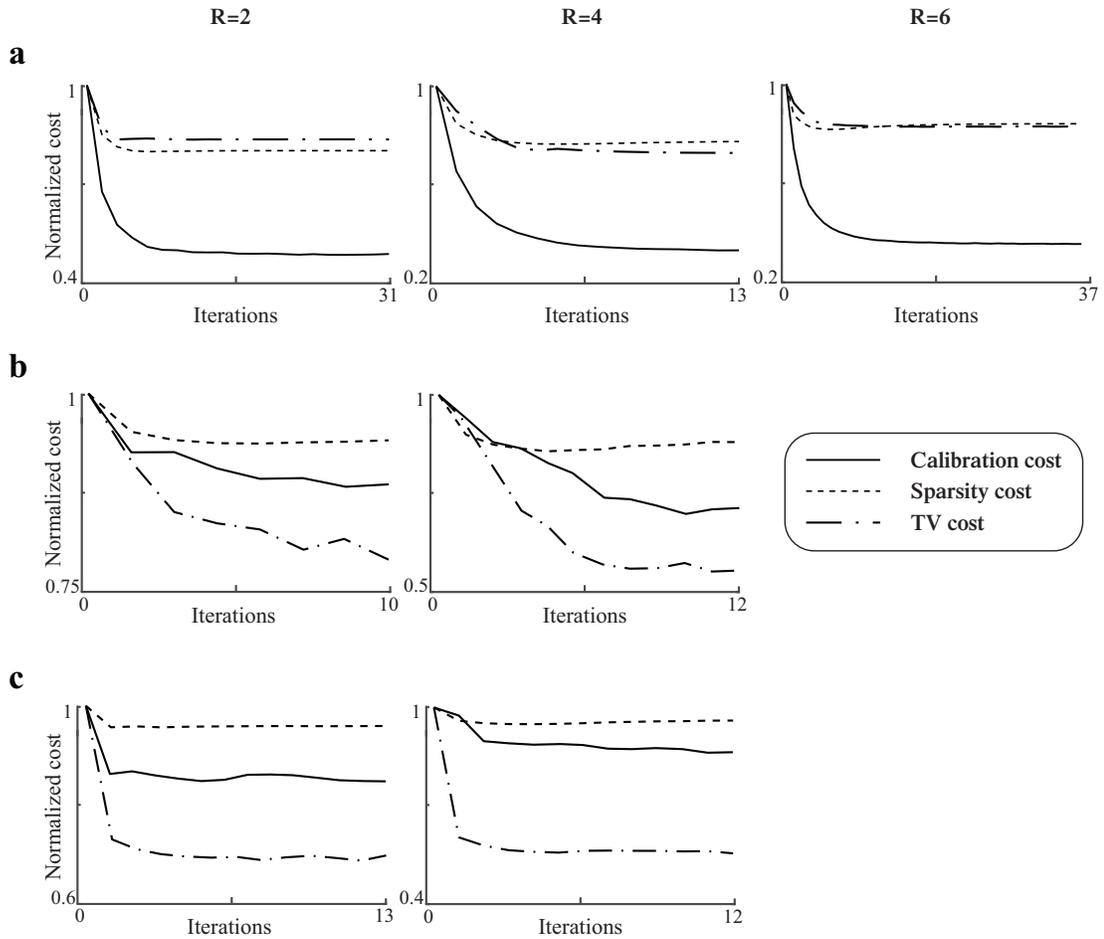


Figure 3.4. Evolution of the PESCiT cost terms. Evolution of the PESCiT cost terms was assessed on in vivo (a) bSSFP, (b) T1-weighted, and (c) ToF acquisitions of the brain. Calibration consistency (solid line), sparsity (dashed line) and TV (dash-dot line) costs were separately calculated at the end of each iteration of the reconstruction. The progression of costs across iterations is shown for a representative cross-section at $R=2$ (left), 4 (middle) and 6 (right column). Reconstructions were stopped once convergence criteria were reached (see Methods). The cost terms diminish smoothly across iterations.

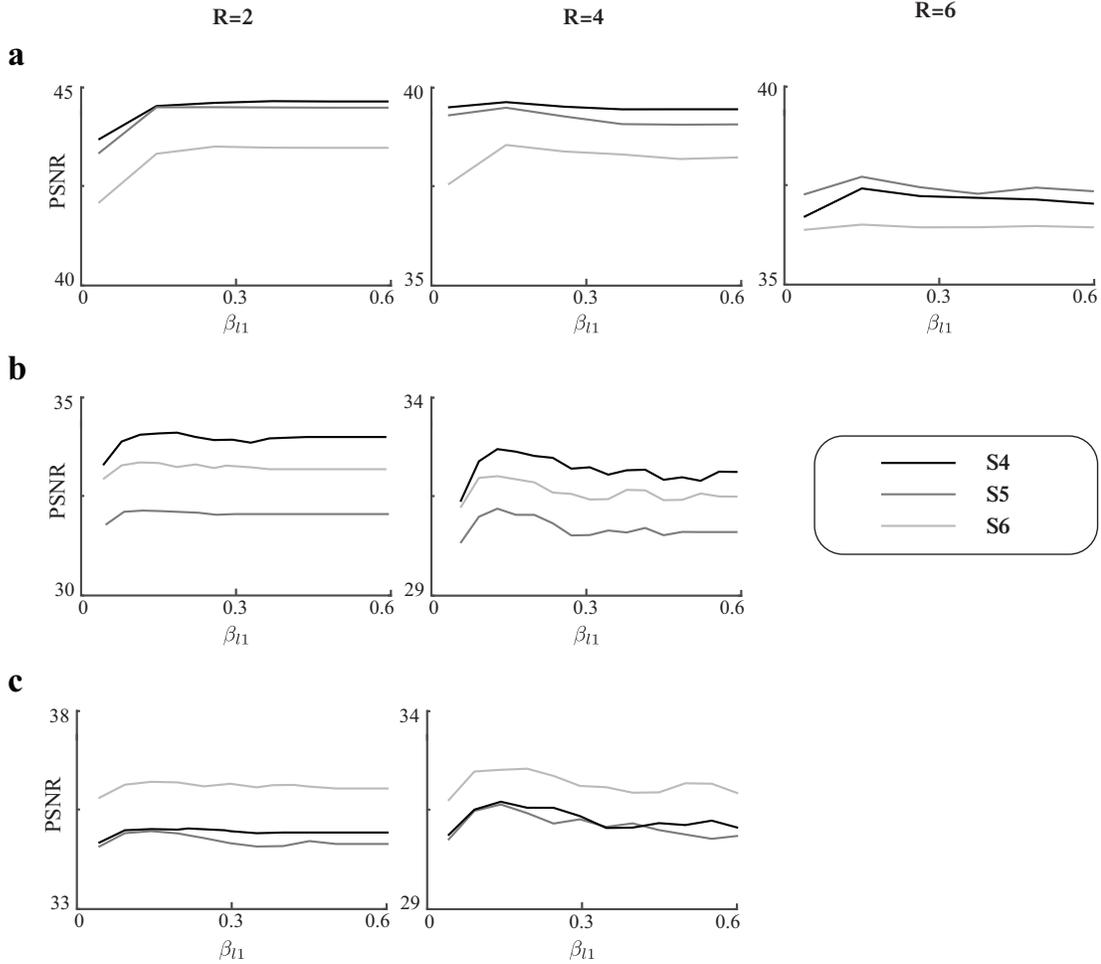


Figure 3.5. Reconstruction performance as a function of the ℓ_1 -epigraph scaling constant. Reconstruction performance was assessed as a function of the epigraph scaling constant for the l_1 -norm function (β_{l_1}). PESCiT reconstructions were performed on in vivo (a) bSSFP, (b) T1-weighted, and (c) ToF acquisitions of the brain. Peak signal-to-noise ratio (PSNR) was measured between reconstructed and fully-sampled reference images. Results averaged across five cross-sections are displayed for the held-out subjects S4, S5 and S6 at $R = 2$ (left), 4 (middle), 6 (right column). Consistently across subjects and across different acquisitions, near-optimal PSNR values are attained for $\beta_{l_1} = 0.2$. Note that PSNR curves demonstrate a high degree of reliability against variations from the optimal β_{l_1} .

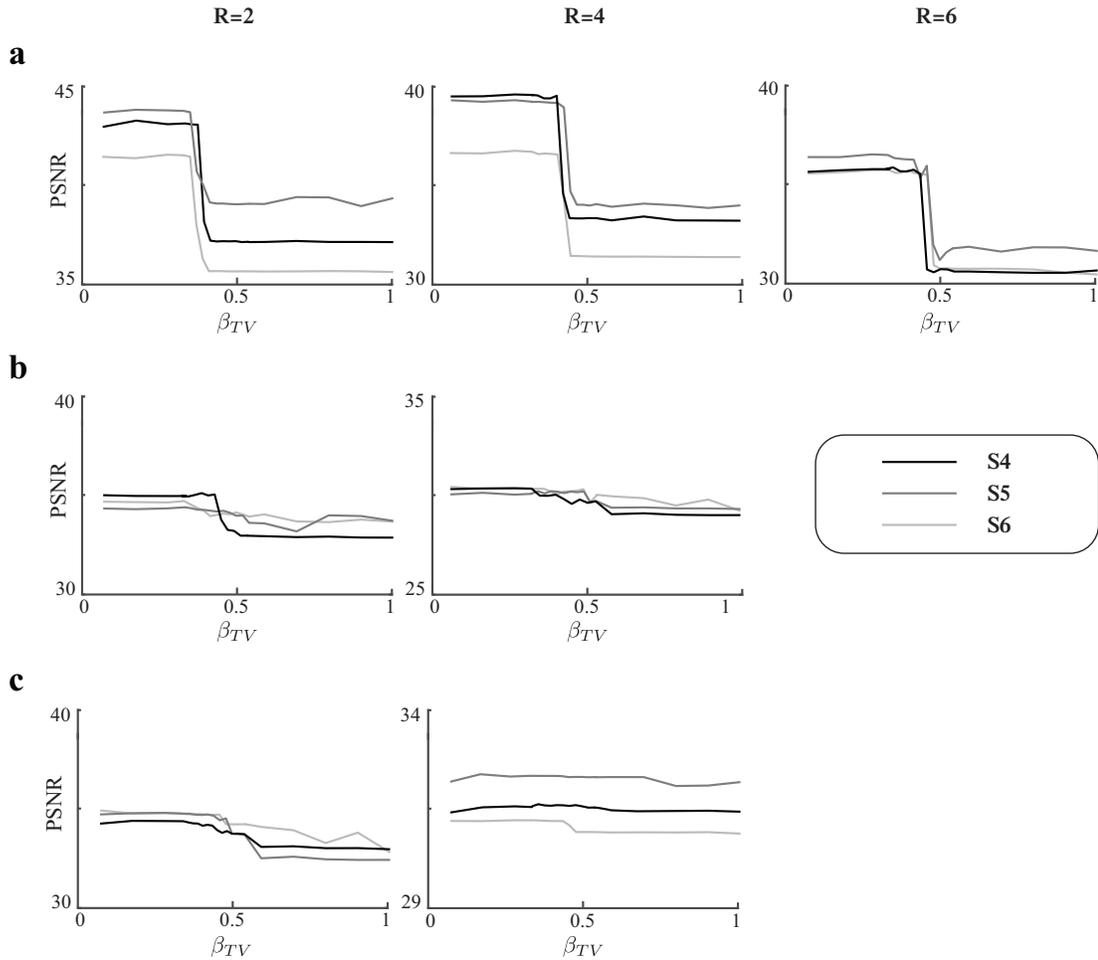


Figure 3.6. Reconstruction performance as a function of the TV-epigraph scaling constant. Reconstruction performance was assessed as a function of the epigraph scaling constant for the TV-norm function (β_{TV}). PESCiT reconstructions were performed on in vivo **(a)** bSSFP, **(b)** T1-weighted, and **(c)** ToF acquisitions of the brain. PSNR was measured between reconstructed and fully-sampled reference images. Results averaged across five cross-sections are displayed for S4, S5 and S6 at $R = 2$ (**left**), 4 (**middle**), 6 (**right column**). Consistently across subjects and across different types of acquisitions, near-optimal PSNR values are attained for $\beta_{TV} = [0.2, 0.4]$. Note that PSNR curves demonstrate a high degree of reliability against variations from the optimal β_{TV} for $\beta_{TV} < 0.4$.

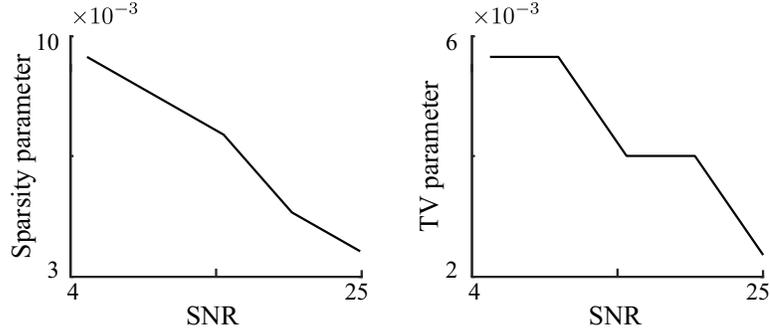


Figure 3.7. Optimal sparsity regularization parameters. Optimal sparsity (**left**) and TV (**right**) regularization parameters as a function of SNR. Reconstructions were performed on simulated bSSFP acquisitions of the brain. Bivariate Gaussian noise was added to phantom images to attain SNR levels varying in $[5, 25]$. Sparsity and TV regularization parameters were independently varied in the range $[0.001, 0.1]$ and reconstructions were performed using $\text{ReCaT}_{\text{fixed}}$. PSNR was measured between reconstructed and fully-sampled reference images. Parameter values yielding the highest PSNR were selected as the optimal values. Optimal sparsity and TV regularization parameters decline as SNR increases.

the reliability of the scaling parameters against variations in the level of detail and spatial resolution, reconstructions were obtained at low, medium and high levels of detail and resolution. Fig. 3.10 displays PSNR across β_{ℓ_1} and Fig. 3.11 displays PSNR across β_{TV} values. Again, PSNR curves as a function of β_{ℓ_1} and β_{TV} demonstrate substantial flatness, yielding near-optimal performance across the entire range of values examined.

Following these basic demonstrations, PESCaT was performed on bSSFP acquisitions of a simulated brain phantom. Representative reconstructions and error maps for PESCaT and ReCaT_{SURE} with $R=6$ are shown in Fig. 3.8. PESCaT yields reduced error across the FOV compared to ReCaT_{SURE} . This improvement with PESCaT becomes further noticeable when ReCaT_{SURE} is stopped early to match its reconstruction time to PESCaT. Quantitative assessments of image quality at $R=2, 4,$ and 6 are listed in Table 3.1. Among all techniques tested, PESCaT achieves the most similar performance to the time-consuming brute-force reconstruction. On average, PESCaT improves PSNR by 0.87 ± 0.74 dB over ReCaT_{SURE} and by 1.87 ± 0.73 dB over ReCaT_{SURE} with early stop (mean \pm std. across five cross-sections, average of $R=2, 4, 6$). Note that the proposed method

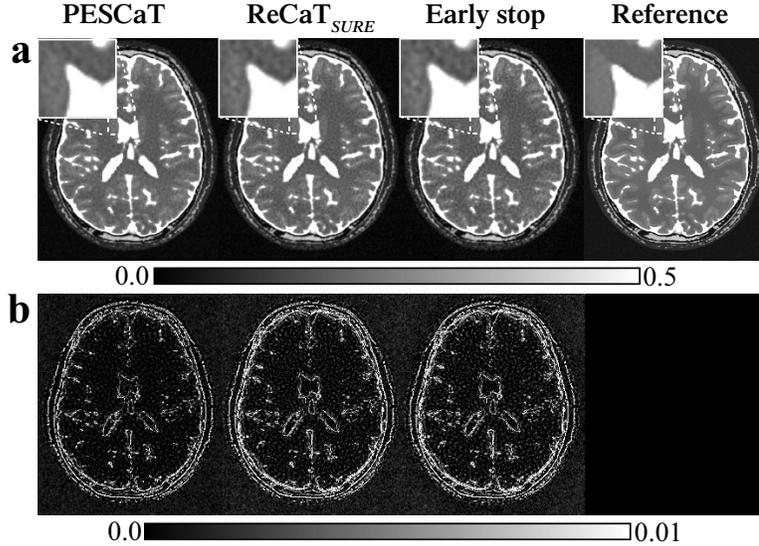


Figure 3.8. Reconstructions of phase-cycled bSSFP acquisitions of the simulated brain phantom. Simulations assumed $\alpha=45^\circ$, $TR/TE=5/2.5$ ms, a fieldmap of 0 ± 62 Hz (mean \pm std), $N = 6$ phase-cycles and $R = 6$. **(a)** PESCaT, ReCaT_{SURE}, ReCaT_{SURE} with early stop, and the fully-sampled reference images are shown. White boxes display zoomed-in portions of the images. **(b)** Mean-squared error between the reconstructed and reference images are shown. PESCaT yields visibly reduced errors compared to both ReCaT_{SURE} and ReCaT_{SURE} with early stop.

attains near-optimal performance while enabling improved computational efficiency. The average reconstruction time per slice is 27 ± 5 s for ReCaT_{SURE} and only 7 ± 4 s for PESCaT, resulting in a 4-fold gain in efficiency for the phantom dataset.

3.4.2 In vivo experiments

We first examined the evolution of the cost terms during PESCaT reconstruction of in vivo bSSFP and T1-weighted datasets (Figs. 3.12, 3.13). Both ℓ_1 and TV cost terms decrease towards later iterations indicating that the images better conform to a compressible representation.

Next, PESCaT was demonstrated for in vivo bSSFP, T1-weighted, and ToF imaging of the brain. Representative reconstructions with $R=6$ for bSSFP and $R=4$ for T1-weighted and ToF acquisitions are displayed in Figs. 3.15, 3.16, and

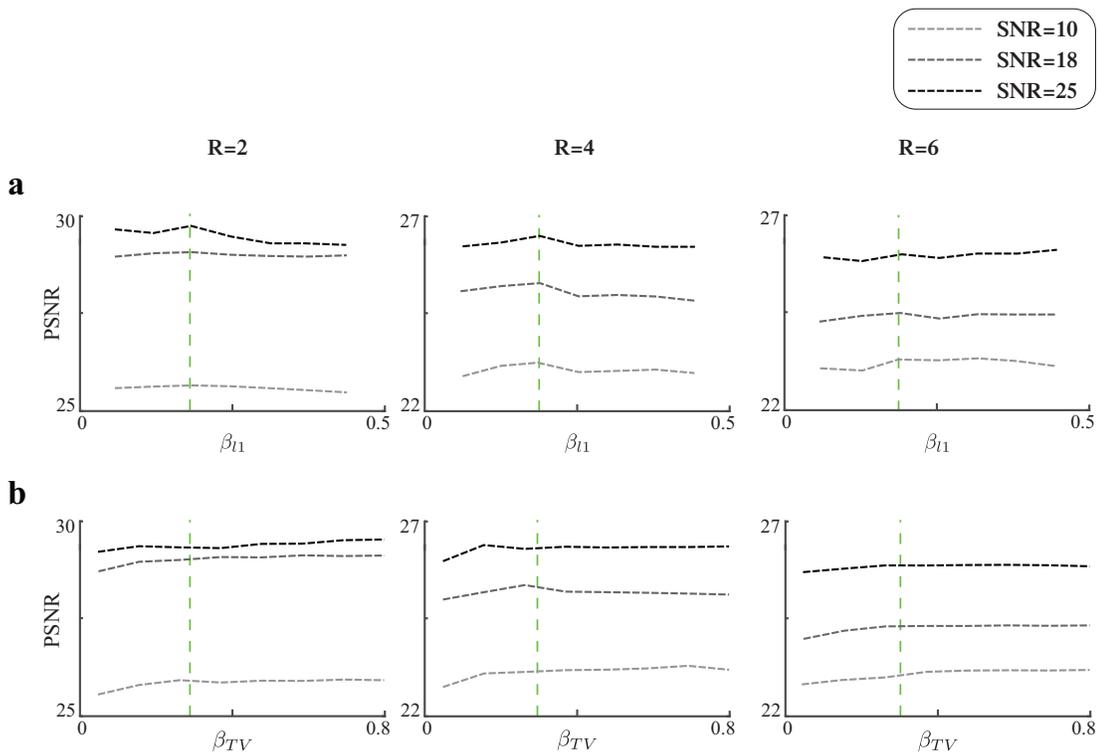


Figure 3.9. Reconstruction performance as a function of acquisition SNR. Reconstruction performance for varying (a) β_{l_1} and (b) β_{TV} as a function of acquisition SNR. PESCiT reconstructions were performed on simulated bSSFP acquisitions of the brain. PSNR was measured between reconstructed and fully-sampled reference images. Results averaged across five cross-sections are displayed for SNR = 10, 18 and 25 at R = 2 (left), 4 (middle), 6 (right column). The epigraph scaling constants selected here are marked with vertical green lines. Consistently across SNR levels, near-optimal PSNR values are attained for a broad range of β_{l_1} and β_{TV} near the selected values.

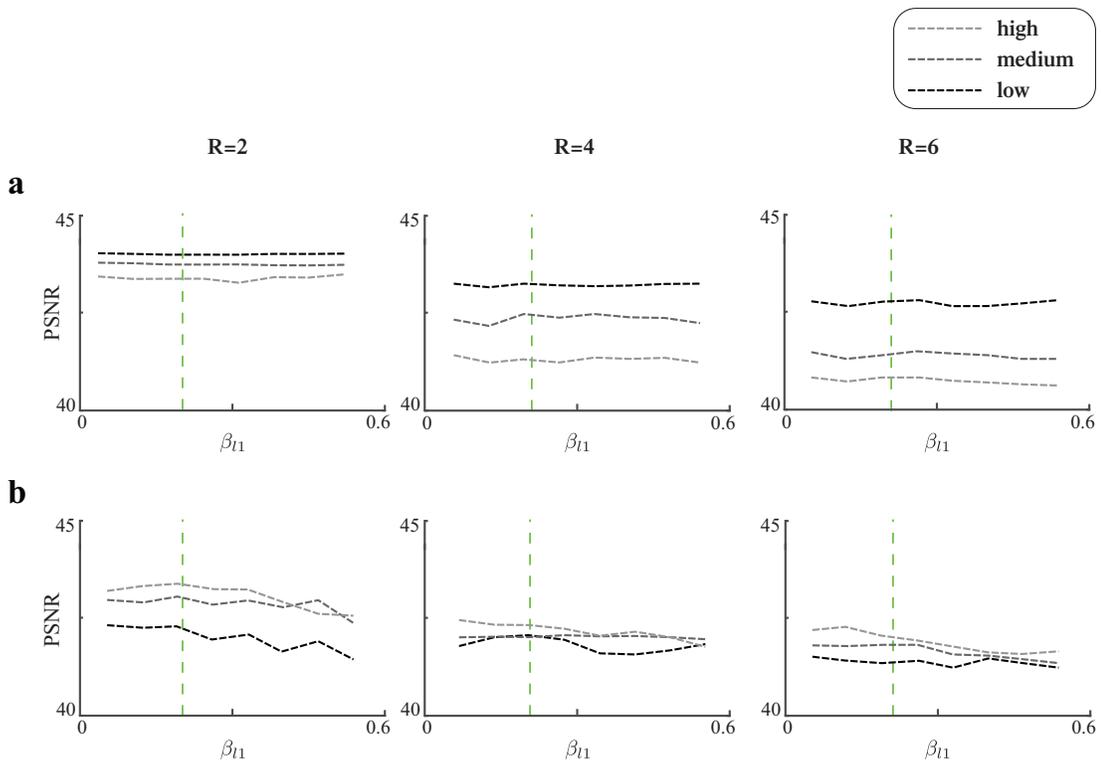


Figure 3.10. Reconstruction performance for varying β_{l_1} as a function of level of detail and resolution of the phantom. Reconstruction performance for varying β_{l_1} as a function of (a) level of detail and (b) resolution of the phantom. PESCiT reconstructions were performed on simulated bSSFP acquisitions (Fig. 3.3). PSNR was measured between reconstructed and fully-sampled reference images. Results are displayed for low, medium, and high level of detail and resolution at $R = 2$ (left), 4 (middle), 6 (right column). The epigraph scaling constant selected Here is marked with vertical green lines. Consistently across levels of detail and resolutions, near-optimal PSNR values are attained for a broad range of β_{l_1} near the selected value.

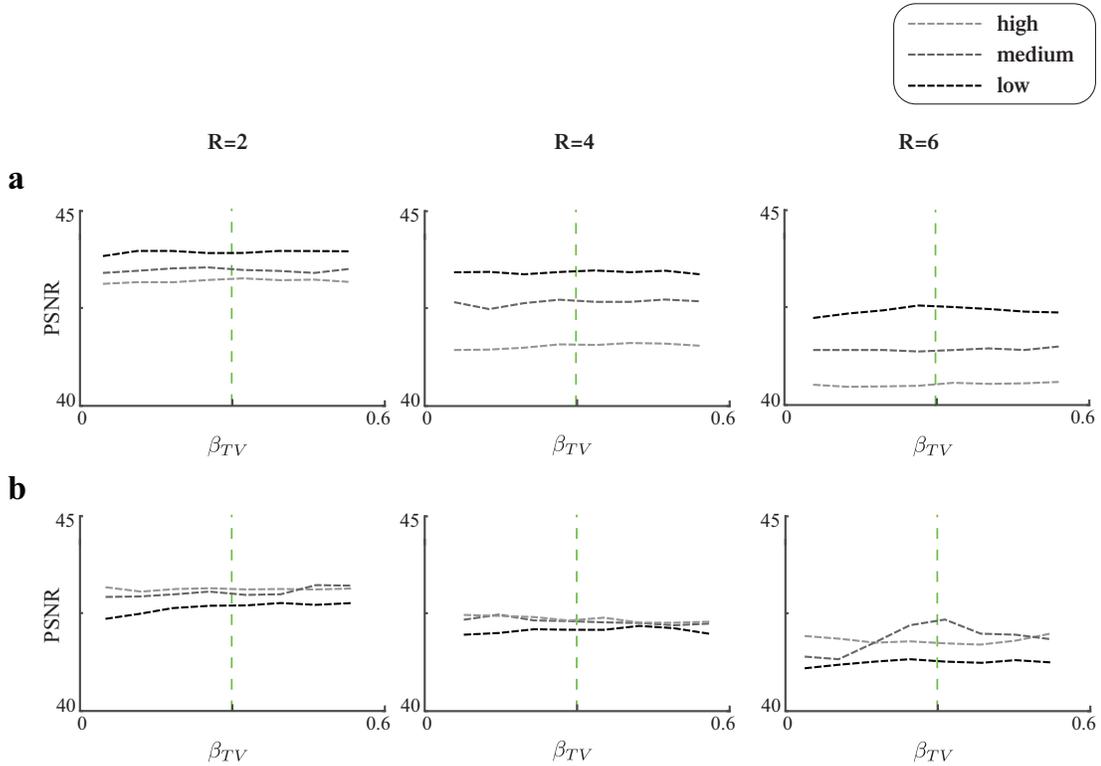


Figure 3.11. Reconstruction performance for varying β_{TV} as a function of level of detail and resolution of the phantom. Reconstruction performance for varying β_{TV} as a function of (a) level of detail and (b) resolution of the phantom. PESCiT reconstructions were performed on simulated bSSFP acquisitions (Fig. 3.3). PSNR was measured between reconstructed and fully-sampled reference images. Results are displayed for low, medium, and high level of detail and resolution at $R = 2$ (left), 4 (middle), 6 (right column). The epigraph scaling constant selected Here is marked with vertical green lines. Consistently across levels of detail and resolutions, near-optimal PSNR values are attained for a broad range of β_{TV} near the selected value.

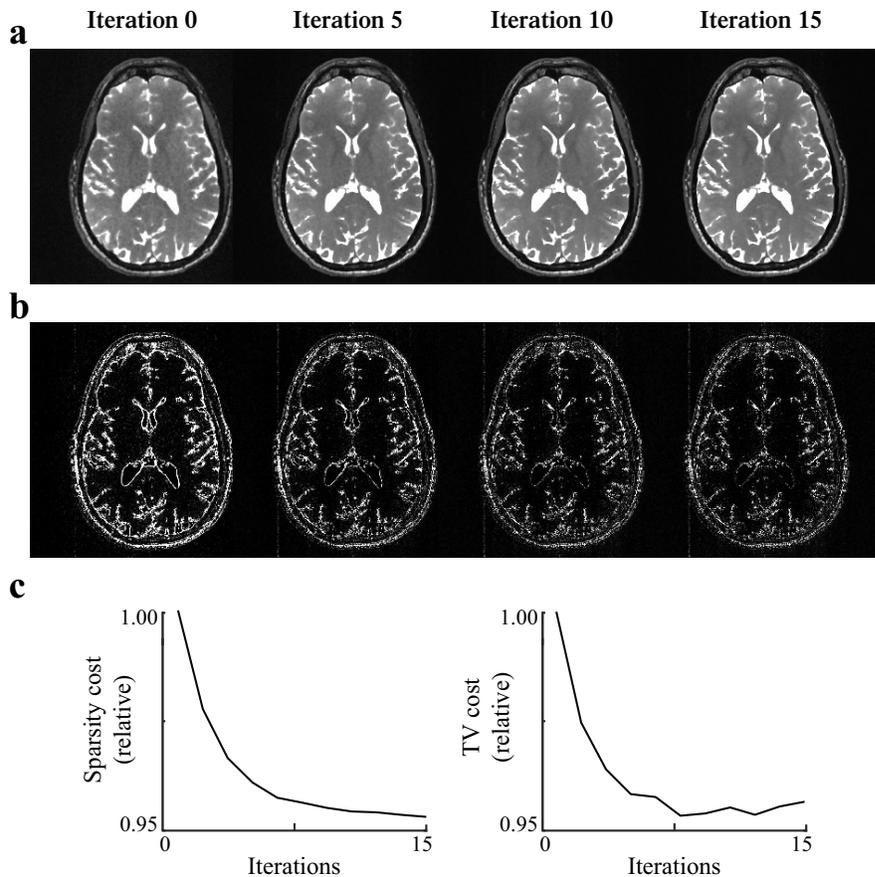


Figure 3.12. Progression of PESCiT across iterations demonstrated on an in vivo bSSFP dataset. (a) Reconstructed images of a central cross-section in a representative subject. Images are shown at the end of 0, 5, 10, and 15 iterations. (b) Respective squared-error images between the reconstructions and the fully-sampled reference. As expected, the reconstruction error diminishes towards later iterations. (c) Sparsity cost (**left panel**) and TV cost (**right panel**) are plotted across iterations. Both cost terms diminish towards later iterations, indicating convergence of the reconstruction towards a sparse representation.

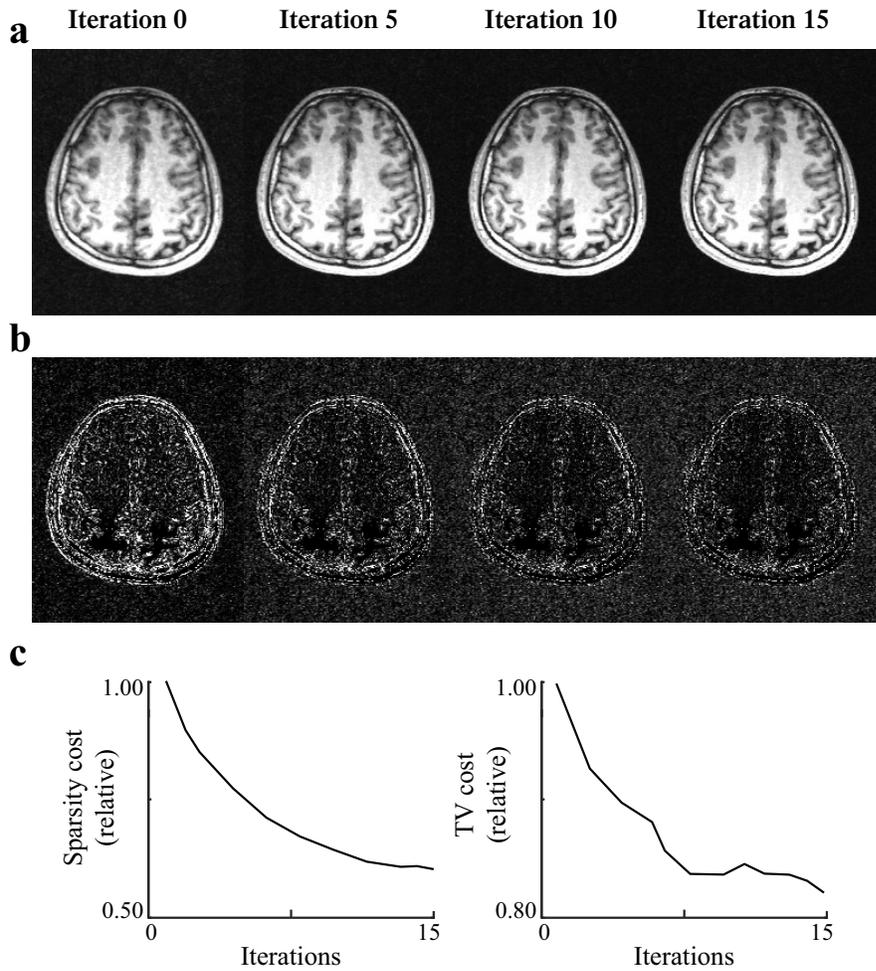


Figure 3.13. Progression of PESCiT across iterations demonstrated on an in vivo T1-weighted dataset. (a) Reconstructed images of a central cross-section in a representative subject. Images are shown at the end of 0, 5, 10, and 15 iterations. (b) Respective squared-error images between the reconstructions and the fully-sampled reference. As expected, the reconstruction error diminishes towards later iterations. (c) Sparsity cost (**left panel**) and TV cost (**right panel**) are plotted across iterations. Both cost terms diminish towards later iterations, indicating convergence of the reconstruction towards a sparse representation.

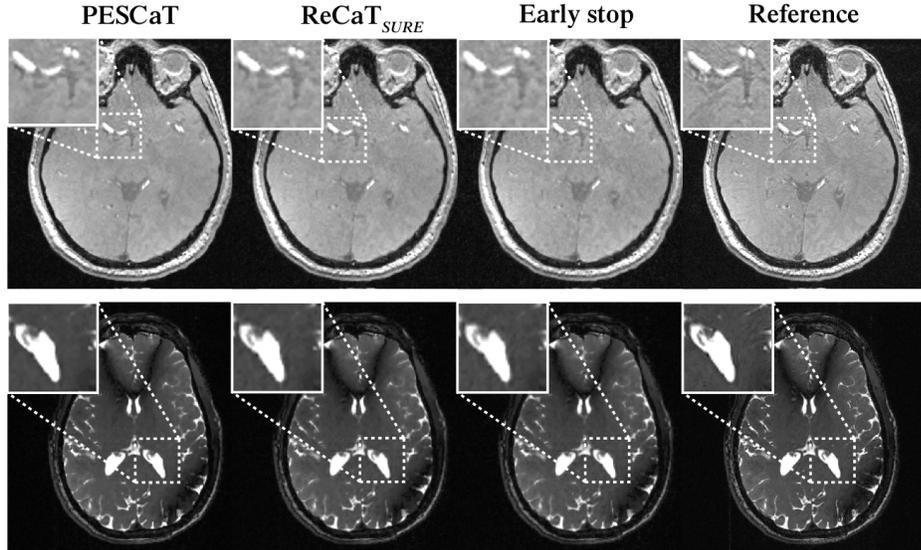


Figure 3.14. Representative cross-sections from the ToF and the phase-cycled bSSFP datasets. A representative cross-section from the ToF dataset (**top row**) and a representative cross-section from the phase-cycled bSSFP dataset (**bottom row**). Reconstructions with PESCaT, ReCaT_{SURE}, early-stop and reference images are presented. PESCaT achieves significantly improved image quality compared to ReCaT_{SURE}, and ReCaT_{SURE} with early stop that was matched to PESCaT in terms of the total reconstruction time.

3.17. Representative reconstructions of individual phase cycles in the bSSFP dataset, and of cross-sections in the ToF dataset are shown in Fig. 3.14. Overall, PESCaT and ReCaT_{SURE} reconstructions perform similar to the brute-force optimized reconstructions. Yet, PESCaT yields slightly lower levels of residual aliasing in comparison to ReCaT_{SURE}, and this difference is particularly noticeable for visualization of small vessels in ToF images (Fig. 3.17). The improvement in reconstruction quality with PESCaT is more prominent when ReCaT_{SURE} is stopped early to match its reconstruction time to PESCaT.

Quantitative assessments of the in vivo reconstructions are listed in Tables 3.2, 3.3, and 3.4. For all datasets and R, PESCaT yields the closest performance to the brute-force reconstruction among alternative self-tuning methods. For bSSFP datasets, PESCaT improves PSNR by 1.23 ± 0.29 dB over ReCaT_{SURE} and by 2.55 ± 0.51 dB over ReCaT_{SURE} with early stop (mean \pm std. across three subjects, average of R=2, 4, 6). For T1-weighted datasets, PESCaT improves PSNR by 0.71 ± 0.25 dB over ReCaT_{SURE} and by 1.21 ± 0.43 dB over ReCaT_{SURE} with early

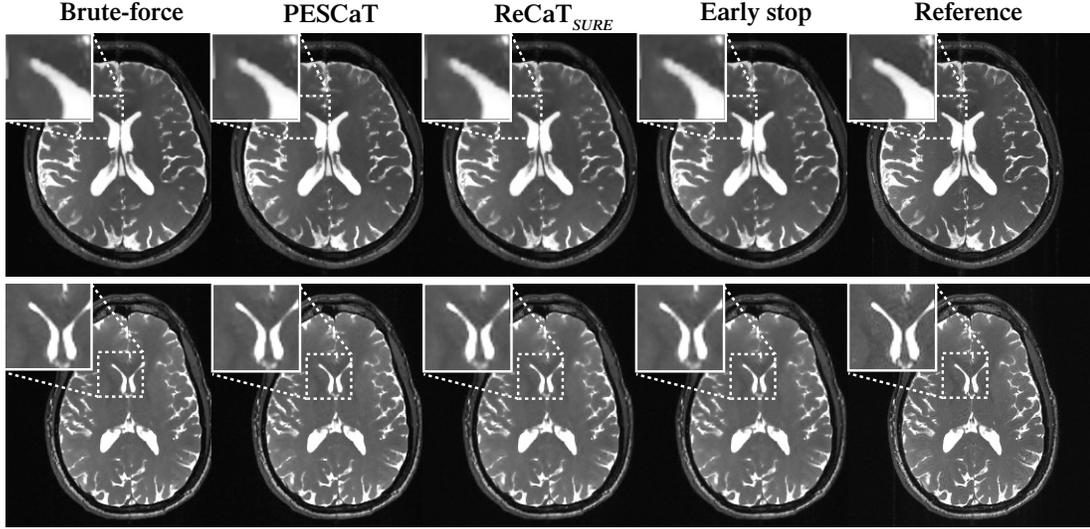


Figure 3.15. Reconstructions of in vivo bSSFP acquisitions of the brain at $R=6$. Brute-force, PESCaT, ReCaT_{SURE}, ReCaT_{SURE} with early stop, and reference images are shown in two representative subjects. White boxes display zoomed-in portions of the images. PESCaT achieves significantly improved image quality compared to ReCaT_{SURE} with early stop that was matched PESCaT in terms of the total reconstruction time. Furthermore, PESCaT yields similar image quality to ReCaT_{SURE} and brute-force methods, while also maintaining greater computational efficiency.

stop (mean \pm std. across three subjects, average of $R=2, 4$). For ToF datasets, PESCaT improves PSNR by 0.72 ± 0.46 dB over ReCaT_{SURE} and by 0.94 ± 0.51 dB over ReCaT_{SURE} with early stop (mean \pm std. across three subjects, average of $R=2, 4$). Compared to empirically-tuned ReCaT_{fixed}, PESCaT improves PSNR by 0.20 ± 0.37 dB for bSSFP datasets, by 0.45 ± 0.14 dB for T1-weighted datasets, and by 0.91 ± 0.63 dB for ToF datasets (mean \pm std. across three subjects, average of $R=2, 4, 6$ for bSSFP, average of $R=2, 4$ for T1-weighted and ToF datasets). Because both methods were allowed to optimize parameters in training subjects, these results suggest that selecting different regularization parameters for each coil/acquisition/subband/level improves reconstruction performance. Compared to PESSPIRiT, PESCaT improves PSNR by 1.16 ± 0.55 dB for bSSFP datasets, by 0.97 ± 0.78 dB for T1-weighted datasets, and by 0.76 ± 0.40 dB for ToF datasets (mean \pm std. across three subjects, average of $R=2, 4, 6$ for bSSFP, average of $R=2, 4$ for T1-weighted and ToF datasets). Performance enhancement is even more prominent compared to PESSPIRiT variants that only include sparsity or TV regularization.

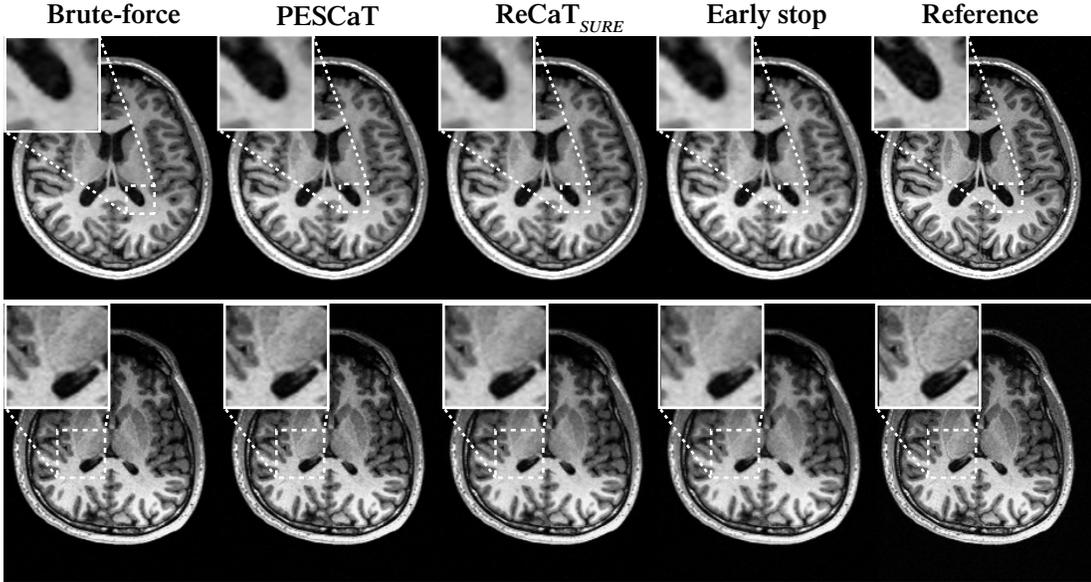


Figure 3.16. Reconstructions of in vivo T1-weighted acquisitions of the brain at $R=4$. Brute-force, PESCiT, ReCaT_{SURE}, ReCaT_{SURE} with early stop, and reference images are shown in two representative subjects. White boxes display zoomed-in portions of the images. PESCiT achieves significantly improved image quality compared to ReCaT_{SURE} with early stop that was matched PESCiT in terms of the total reconstruction time. Meanwhile, PESCiT yields similar image quality to ReCaT_{SURE} and brute-force methods.

To assess the computational efficiency of self-tuning methods, representative reconstructions were performed for a single cross-section of in vivo bSSFP, T1-weighted, and ToF acquisitions. The true MSE between the reconstructed and fully-sampled reference images were recorded across iterations of PESCiT and ReCaT_{SURE}. MSE curves across iterations are displayed in Fig. 3.18. Compared to ReCaT_{SURE}, the proposed method converges to a lower MSE value for all R and datasets. Furthermore, PESCiT reduces the number of iterations by 43.3% for bSSFP (average over $R=2, 4, 6$), 74.5% for T1-weighted (average over $R=2, 4$) and 53.2% for ToF (average over $R=2, 4$) datasets. Note that each iteration of PESCiT performs more efficient geometric projections without explicit parameter searches. The reconstruction times for PESCiT and alternative methods are listed in Table 3.5. On average, the reconstruction time of ReCaT_{SURE} was 1641 ± 45 s for bSSFP, 1799 ± 66 s for T1-weighted, and 565 ± 58 s for ToF datasets (mean \pm std. across five cross-sections, average over $R=2, 4$ for T1-weighted and ToF imaging; $R=2, 4, 6$ for bSSFP imaging). In contrast, the reconstruction time

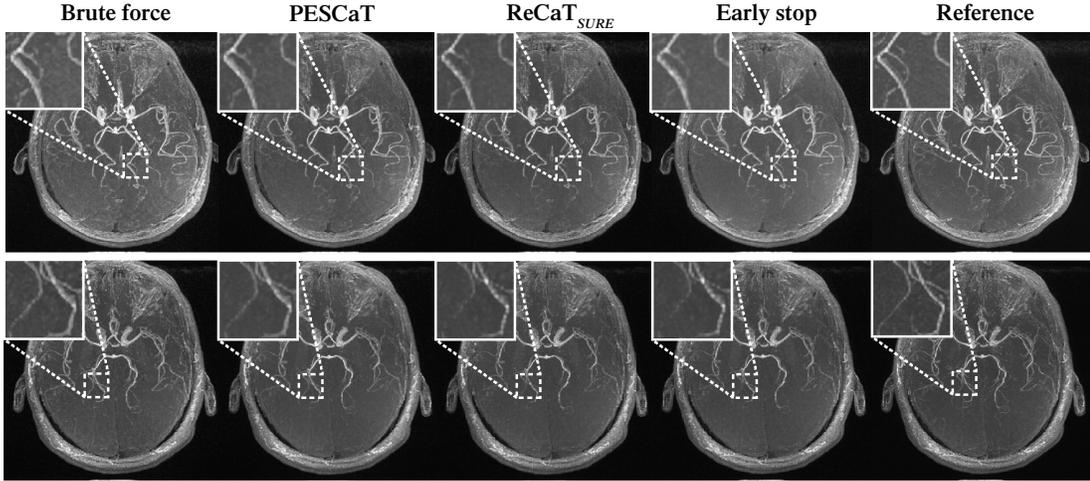


Figure 3.17. Reconstructions of in vivo ToF angiography acquisitions of the brain at $R=4$. Maximum-intensity projection (MIP) views of brute-force, PESCaT, ReCaT_{SURE}, ReCaT_{SURE} with early stop, and reference brain volumes are shown in two representative subjects. White boxes display zoomed-in portions of the MIPs. PESCaT yields superior depiction of vasculature compared to both ReCaT_{SURE} and ReCaT_{SURE} with early stop. It also maintains similar image quality to brute-force reconstructions.

of PESCaT was merely 164 ± 25 s for bSSFP, 196 ± 44 s for T1-weighted, and 159 ± 32 s for ToF datasets. These results suggest that PESCaT offers up to 10-fold gain in efficiency compared to the alternative self-tuning method ReCaT_{SURE}. While PESSPIRiT yields similar reconstruction times and ReCaT_{fixed} slightly reduces reconstruction times compared to PESCaT, both methods yield inferior reconstruction quality.

Lastly, reconstruction performance of PESCaT was demonstrated at higher acceleration rates using the 32-channel bSSFP datasets (Fig. 3.19 and Table 3.6). The proposed method improves PSNR by 0.14 ± 0.04 compared to ReCaT_{fixed}, by 1.59 ± 0.45 compared to ReCaT_{SURE}, by 3.77 ± 0.61 compared to ReCaT_{SURE} with early stop, and by 4.08 ± 0.55 over PESSPIRiT (mean \pm std. across three subjects, average of $R=8, 10$). These results help demonstrate the utility of PESCaT in enabling higher acceleration factors when using modern coil arrays.

	R=2	R=4	R=6
bSSFP			
Brute-force	3562±52	4161±33	4365±29
PESCaT	122±57	130±16	238±25
ReCaT_{SURE}	813±34	1625±37	2443±49
Early stop	122±57	130±16	238±25
ReCaT_{fixed}	81±11	84±04	89±09
PESSPIRiT	92±10	128±18	132±17
PESSPIRiT_{II}	24±02	17±03	16±03
PESSPIRiT_{TV}	76±16	21±07	22±09
T1-weighted			
Brute-force	8569±11	10167±35	
PESCaT	179±18	213±21	
ReCaT_{SURE}	1801±58	1794±53	
Early stop	179±18	213±21	
ReCaT_{fixed}	121±13	182±18	
PESSPIRiT	272±10	144±07	
PESSPIRiT_{II}	28±03	60±10	
PESSPIRiT_{TV}	48±13	28±07	
ToF			
Brute-force	3721±22	3717±23	
PESCaT	184±41	133±25	
ReCaT_{SURE}	565±27	565±27	
Early stop	184±41	133±25	
ReCaT_{fixed}	94±12	98±19	
PESSPIRiT	144±09	148±15	
PESSPIRiT_{II}	48±13	80±09	
PESSPIRiT_{TV}	128±14	128±13	

Table 3.5. Reconstruction times (seconds) for in vivo datasets. Reconstruction times were averaged over five cross sections from three subjects.

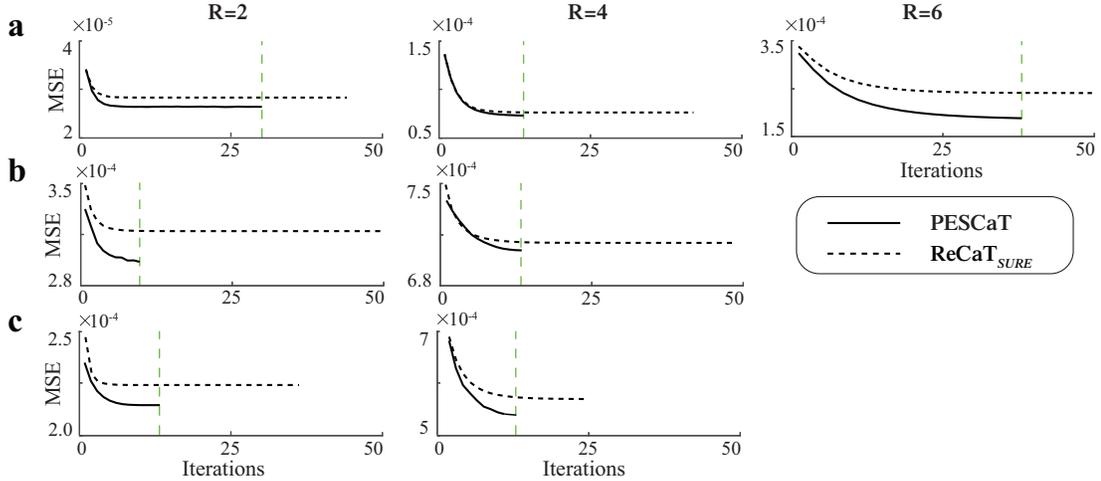


Figure 3.18. Convergence behavior of self-tuning reconstructions. Convergence behavior of self-tuning reconstructions was assessed on in vivo **(a)** bSSFP, **(b)** T1-weighted, and **(c)** ToF acquisitions of the brain. Mean-squared error (MSE) was calculated between the image reconstructed at each iteration and the fully-sampled reference image. The progression of MSE across iterations is shown for a representative cross-section reconstructed using PESCaT (**solid lines**) and ReCaT_{SURE} (**dashed lines**) at R=2 (**left**), 4 (**middle**), and 6 (**right**). Reconstructions were stopped once convergence criteria were reached (see *Methods*). The iterations at which PESCaT converged are indicated (dashed green lines). In all cases, PESCaT converges in a significantly smaller number of iterations, and it converges to a solution with lower MSE than ReCaT_{SURE}.

3.5 Discussion

Here, we have proposed a new self-tuning method for CS reconstruction of single-coil multi-acquisition, multi-coil single-acquisition, and multi-coil multi-acquisition datasets. The proposed method performs sparsity projections across coils and acquisitions to penalize the ℓ_1 -norm of wavelet coefficients, and TV projections to penalize the finite-differences gradients of image coefficients. Separate sparsity regularization parameters are selected at each wavelet subband and level, and separate TV regularization parameters are selected at each coil and acquisition. Efficient projections onto the boundary of the epigraph sets of the ℓ_1 -norm and TV-norm functions are used to simultaneously calculate the projections themselves and automatically determine the relevant regularization parameters. PESCaT does not have any constraints regarding the number of acquisitions or

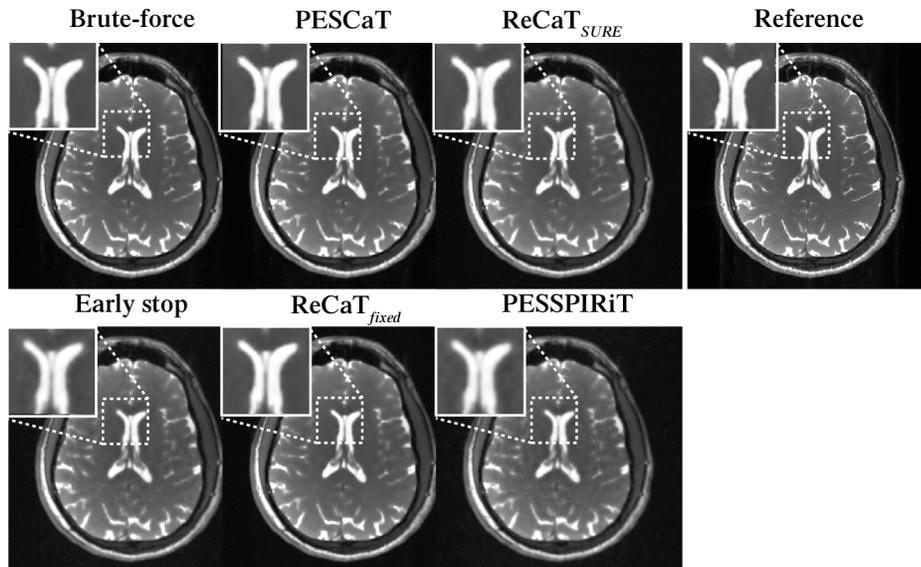


Figure 3.19. Reconstructions of 32-channel in vivo bSSFP acquisitions of the brain at $R=10$. Brute-force, PESCiT, ReCaT_{SURE}, ReCaT_{SURE} with early stop, ReCaT_{fixed}, PESSPIRiT, and reference images are shown in a representative subject. White boxes display zoomed-in portions of the images. PESCiT achieves significantly improved image quality compared to the alternative methods. Furthermore, PESCiT yields similar image quality to the brute-force method, while also maintaining greater computational efficiency.

Table 3.6. Numerical assessment of bSSFP reconstructions at high acceleration rates.

	$R=8$		$R=10$	
	PSNR	NRMSE $\times 10^3$	PSNR	NRMSE $\times 10^3$
Brute-force	38.45 \pm 0.55	14.11 \pm 0.89	37.15 \pm 0.41	17.45 \pm 0.85
PESCiT	37.48\pm1.03	16.94\pm1.57	36.08\pm0.86	19.11\pm1.53
ReCaT _{fixed}	37.37 \pm 1.01	17.12 \pm 1.57	35.89 \pm 0.92	19.41 \pm 1.66
ReCaT _{SURE}	35.65 \pm 1.13	19.75 \pm 1.13	34.72 \pm 1.07	21.27 \pm 1.23
Early stop	33.38 \pm 0.90	22.48 \pm 1.19	32.63 \pm 0.91	23.25 \pm 1.44
PESSPIRiT	33.10 \pm 1.47	23.21 \pm 2.70	32.29 \pm 1.20	25.49 \pm 2.51
PESSPIRiT _{l₁}	31.92 \pm 2.01	26.02 \pm 2.19	31.30 \pm 1.90	27.98 \pm 2.23
PESSPIRiT _{TV}	31.46 \pm 1.96	27.57 \pm 2.24	30.78 \pm 1.76	29.79 \pm 1.96

bSSFP acquisitions with 32 channels and 8 phase-cycles were reconstructed. PSNR and NRMSE were measured between the reconstructed image and a fully-sampled reference image. Measurements were obtained for all implemented methods. Results are averaged across three subjects, and reported as mean \pm std across five cross-sections.

coils. Therefore, it can be readily applied to both single-acquisition and multi-acquisition datasets regardless of the number of coils available. PESCiT also offers flexibility regarding the inclusion of regularization terms. Because the algorithm has a modular structure with respect to individual calibration, sparsity, and TV projections, it is possible to omit either TV or sparsity regularization. The proposed method will still work towards a solution at the intersection of the remaining projection sets.

A recent study proposed a reconstruction for multi-coil multi-acquisition bSSFP imaging, named ReCaT [96]. Here, we have implemented a self-tuning version of ReCaT (ReCaT_{SURE}). Similar to PESCiT, ReCaT_{SURE} uses sparsity projections implemented via soft-thresholding and TV projections implemented via iterative clipping. However, in ReCaT_{SURE}, the sparsity regularization parameter was selected via a SURE-based method to minimize the expected reconstruction error. TV regularization parameter was selected in a data-driven manner based on the local standard deviations within the reconstructed image. Since parameter selection in ReCaT_{SURE} involves line searches over a relevant range of parameters, it can be computationally expensive. In contrast, PESCiT leverages highly efficient geometric projections onto epigraph sets to simultaneously select the optimal parameters and calculate the projections. Hence, PESCiT enables significant savings in reconstruction time compared to self-tuning methods based on line searches. Meanwhile, the main advantage of PESCiT over an empirically-tuned reconstruction that optimizes regularization parameters on training data is that it allows for independent selection of regularization parameters for each coil/acquisition/subband/level. The superior reconstruction quality of PESCiT compared to ReCaT_{fixed} confirms this prediction quantitatively.

The proposed method includes two epigraph scaling constants β_{ℓ_1} and β_{TV} as free parameters. Here we have empirically demonstrated that the optimal scaling constants are highly consistent across individual subjects, across different noise levels and across multiple imaging contrasts of the same anatomy. These observations are also complemented by prior work that suggests that the solutions of epigraph sets projections are robust against deviations from optimal scaling constants [198, 31]. It remains to be demonstrated whether the scaling constants

are also similar across different anatomies. Still, we expect that PESCaT shows improved robustness against variability in datasets compared to the empirically-tuned ReCaT_{fixed}. The optimal regularization parameters for ReCaT_{fixed} showed relatively high variability across the datasets examined Here (not shown). Thus, ReCaT_{fixed} might require more careful tuning of regularization parameters, resulting in relatively higher computational overhead.

Further performance improvements might be attained by addressing some limitations of the proposed method. For multi-acquisition datasets, significant motion among acquisitions can reduce reconstruction quality. A motion-correction projection can be incorporated into the PESCaT algorithm to mitigate artifacts due to the residual motion. Second, the proposed method uses a fully-sampled central region in k-space to estimate the tensor interpolation kernel. In applications where the acquisition of calibration data is impractical such as spectroscopic and dynamic imaging, calibrationless approaches could be incorporated for improved performance [199, 184]. Third, although the epigraph scaling constants β_{ℓ_1} and β_{TV} were optimized over a held-out dataset, it might be possible to automatically select them using parameter selection via SURE or GCV. This remains an important future research direction toward fully-automated reconstructions.

Here, the alternating projections onto sets algorithm was used to find a solution at the intersection point of the sets corresponding to calibration, sparsity, and TV projections. Rapid convergence was observed in all examined cases. However, in situations where the intersection between these sets is sparsely populated, more sophisticated algorithms such as alternating direction method of multipliers (ADMM) could be used for fast and effective optimization [20]. PESCaT employs projections onto epigraph sets to concurrently select regularization parameters and perform projections. As such, it is non-trivial to efficiently adapt the proposed parameter selection to an ADMM-based reconstruction. It remains an important future work to benchmark PESCaT against ADMM coupled with an appropriate parameter-selection strategy.

Projections onto epigraph sets were used to penalize ℓ_1 -norm and TV-norm functions Here. Note that the projection onto convex sets formulation allows

penalization of any convex function. Thus, the proposed technique could be generalized to include alternative regularizers such as filtered variation or total generalized variation [108]. These modifications might allow performance enhancements in applications where standard TV regularization yields undesirable block artifacts.

In conclusion, PESCaT enables near-optimal image quality while automatically selecting regularization parameters in reconstructions of undersampled MRI datasets. Parameter selection for ℓ_1 -norm and TV-norm regularizers and projections onto the ℓ_1 and TV-balls are performed simultaneously. PESCaT was demonstrated to outperform alternative self-tuning approaches based on SURE in bSSFP, T1-weighted and time-of-flight angiographic imaging. The results presented here demonstrate that PESCaT is a promising method for CS-MRI in routine practice.

3.6 Publications

This chapter of the thesis have been partially presented and published in the following conferences and journals:

- Mohammad Shahdloo, Efe Ilicak, Mohammad Tofghi, Emine U Saritas, A Enis Cetin, and Tolga Çukur. Projection onto Epigraph Sets for Rapid Self-Tuning Compressed Sensing MRI. *IEEE Transactions on Medical Imaging*, 38(7):1677–1689, July 2019.
- Mohammad Shahdloo, Efe Ilicak, Mohammad Tofghi, Emine U Saritas, A Enis Cetin, and Tolga Çukur. Rapid Self-Tuning Compressed-Sensing MRI Using Projection onto Epigraph Sets. In *International Society for MR in Medicine (ISMRM)*, page 0251, Paris, 2018.
- Mohammad Shahdloo, Efe Ilicak, Mohammad Tofghi, Emine U Saritas, A

Enis Cetin, and Tolga Çukur. Adaptive Wavelet Thresholding for Profile-
Encoding Reconstruction of Balanced Steady-State Free Precession Acqui-
sitions. In *European Society for MR and Medicine (ESMRMB)*, page 391,
Barcelona, 2017.

Chapter 4

Attentional Modulations of Action-Category Representation in the Brain

Summary

Humans are adept in perceiving others' actions. Action perception in natural scenes, however, relies on efficient allocation of limited brain resources to prioritize processing the task-relevant perceptual visual inputs. It has been suggested that during visual search for objects, distributed semantic representation of hundreds of object categories get modulated to expand the representation of target. Yet, it is unknown whether and where in the brain visual search for action categories modulates semantic representations. To investigate this dynamic attentional process, we analyzed brain activities recorded via functional magnetic resonance imaging while subjects viewed natural movies and searched for “communication” or “locomotion” actions. We used encoding models to measure selectivity for each of the 109 action categories in the movies. We derived an embedding space that encoded semantic variability among action categories. We then investigated attentional modulation of semantic representation during search for action categories. We find that attention modulates semantic representation of actions widely across neocortex. Moreover, the degree of attentional modulation interacts with intrinsic selectivity for target action categories. These results suggest

that attention dynamically modulates semantic representation to optimize perception of the task-relevant actions during natural vision.

4.1 Introduction

The ability to reliably perceive observed actions along with intentions of actors is critical for survival of most species and for fulfillment of their behavioral demands. Recent studies have attributed this ability to the action observation network (AON), a network of occipitotemporal, parietal and premotor areas in humans [148, 27], and homologously in other primates [139, 140, 205]. Several reports suggest that various types of information pertaining to actions, ranging from shape and kinematics to action-effector interactions and action categories are represented hierarchically [64, 70, 220, 201] and at various locations across the AON. For instance, shape and kinematics of actions are represented in lateral occipitotemporal cortex and in the posterior bank of inferior temporal cortex [100]. Effector type (e.g. foot, hand) is represented in ventral premotor cortex (101, 34), while parietal and cingulate cortices represent action categories [1, 55, 150]. These studies suggest a modular functional organization for the AON.

Furthermore, there is evidence that selective attention to low-level action features modulates action representation across AON. Early psychophysical studies suggest that attending to low-level visual features of objects involved in actions modulates the perception of action cues [196] and the amount of attentional modulation depends on relevance of the target feature to the observed action [32]. More recent reports provide evidence for top-down influences of attending to action kinematics on population responses across AON [134, 175]. Muthukumaraswamy and Singh [134] presented video clips of a hand that performed various sequences of finger movements. They reported enhanced responses across the AON, as reflected in attenuation of the beta band (15-35Hz) oscillation over sensorimotor cortex, while subjects were attending to the sequence of finger movements compared to passive viewing. Furthermore, Schuch et al. [175] used short

video clips of actors' hands reaching a cup followed by grasping either the top of the cup (i.e. power grip), or the handle (i.e. precision grip). Enhanced responses across the AON were reported while subjects attended to action kinematics (i.e. grip type) versus to the color of the cup.

In addition to these top-down influences, recent neuroimaging studies extend the evidence for attentional modulations to higher-level action features ([44, 174, 81, 142], see [195] and [160] for reviews). Safford et al. [174] presented overlapping moving tools and moving humans via simplistic point-light dots [104]. They reported that attending to animate moving actors (i.e. humans) enhanced blood oxygen level dependent (BOLD) responses in superior temporal sulcus (STS), whereas attending to inanimate moving objects enhanced responses in inferior temporal sulcus (ITS) and middle temporal gyrus (MTG). Further to this, Nicholson et al. [142] presented short action movies in controlled scenes and subjects attended to action kinematics, to objects involved in actions, or to action goals in different runs. They reported enhanced responses in inferior frontal gyrus (IFG), lateral occipitotemporal cortex (LOTTC), and medial frontal gyrus (MFG) as a result of attention to action goals.

Despite these lines of evidence suggesting that attending to a hierarchy of features ranging from low-level actor animacy and kinematics, to higher-level action features such as involved objects involved and action goals modulates responses across AON, action categories also convey a significant portion of information within observed actions. Crucially, there is evidence suggesting that hundreds of object and action categories in natural scenes are semantically represented across cortex [91], and visual search for object categories warps semantic representations in favor of the target [37]. Yet, only a few studies have investigated modulations of semantic representations due to attention to action categories [38, 137, 138]. In a recent study, Nastase et al. [137] presented movie clips of animals from five taxonomies (primates, ungulates, birds, reptiles, and insects) performing actions belonging to four categories (eating, fighting, running, and swimming). Participants were instructed to attend either to the animal taxonomy or to actions. The authors fit general linear models (GLMs) for all combinations of taxa and

actions and modeled the representational geometry by assuming individual template representational dissimilarity matrices (RDMs [112, 111, 80]) for taxonomy and for actions. Using multivariate analysis on RDMs across tasks, the authors reported that during search for actions RDMs in LOTC and in premotor cortex were biased toward the template RDM for actions. These previous studies provide evidence for global attentional modulations in semantic representation of a handful of actions. Yet, it is currently unknown whether and how visual search for specific action categories modulates semantic representation of the multitude of actions in natural scenes.

We hypothesized that visual search for action categories during natural vision should expand the representation of targets by causing voxelwise semantic tuning shifts (Fig. 4.1 [37]). To test our hypothesis, we analyzed the recorded whole-brain BOLD responses from five subjects while they viewed 60 min of natural movies. Data were collected at the University of California, Berkeley. Subjects maintained steady fixation and covertly searched for 15 “communication” actions or 28 “locomotion” actions among 109 action categories. We used natural movie stimuli since they have greater ecological validity [54], lead to more reliable inter-subject neural responses [77, 79], and present action categories in their natural context, a factor that is important in action perception [94, 219, 49]. Voxelwise modeling with spatial regularization [28] was used to measure category responses for hundreds of objects and actions in the movies separately for each individual subject and for each attention condition [143, 136]. Principal component analysis was used to estimate a semantic space underlying action category responses. Semantic tuning was then assessed by projecting action category responses onto this semantic space. Finally, the voxelwise semantic tuning profiles during the two attention conditions were compared to quantify the magnitude and direction of tuning shifts.

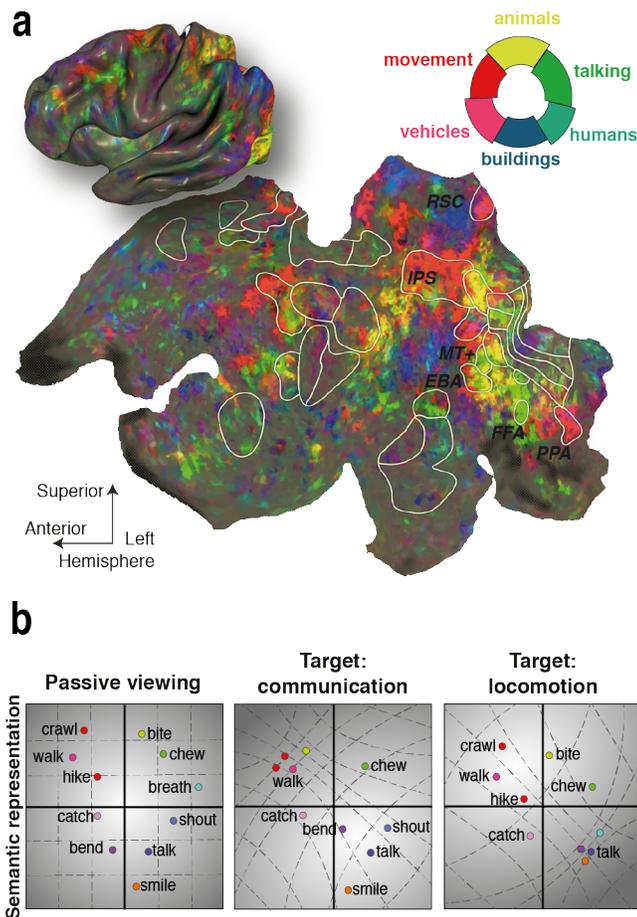


Figure 4.1. Hypothesized changes in semantic representation of action categories.

a Recent studies have proposed that the human brain organizes hundreds of object and action categories in a semantic space that is distributed systematically across cortex. Under this view, functional areas that are classically known to represent certain object-action categories are peaks within this continuous semantic space [91]; PPA, parahippocampal place area; FFA, fusiform face area; EBA, extrastriate body area; MT+, human MT; IPS, intraparietal sulcus; RSC, retrosplenial cortex. **b.** In this semantic space, action categories that are semantically similar to each other are mapped to nearby points and semantically dissimilar categories are mapped to points far from each other. We hypothesized that visual search for specific action categories should expand the representation of the target and semantically similar categories, while compressing the representation of dissimilar categories [37].

4.2 Methods

4.2.1 Subjects

Five healthy male adult volunteers with normal or corrected-to-normal vision participated in this study: S1 (age 31), S2 (age 27), S3 (age 32), S4 (age 33), S5 (age 27). Data were collected at the University of California, Berkeley. Experiment protocols were approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley. All participants gave written informed consent before scanning.

4.2.2 fMRI data collection

Data were collected using a 3T Siemens Tim Trio MRI scanner (Siemens Medical Solutions) via a 32-channel receiver coil. Functional data were collected using a T2*-weighted gradient-echo echo-planar-imaging pulse sequence with the following parameters: TR = 2 sec, TE = 33 msec, water-excitation pulse with flip angle = 70° , voxel size = 2.24 mm \times 2.24 mm \times 4.13 mm, field of view = 224 mm \times 224 mm, 32 axial slices. To construct cortical surfaces, anatomical data were collected using a three-dimensional T1-weighted magnetization-prepared rapid-acquisition gradient-echo (MPRAGE) sequence with the following parameters: TR = 2.3 sec, TE = 3.45 msec, flip angle = 10° , voxel size = 1 mm \times 1 mm \times 1 mm, field of view = 256 mm \times 212 mm \times 256 mm. Surface flattening and visualization were done via Freesurfer and PyCortex [41, 167, 59].

4.2.3 Stimuli and experimental design

Data for the main experiment were collected in six 10 min 50 s runs in a single session. Continuous natural movies were used as the stimulus in the main experiment. Three distinct 10 min movie segments were compiled from short movie clips

(10-20 secs) without sound. Movie clips were selected from a diverse set of natural movies (see [143] for details). Movie clips were cropped into a square frame and downsampled to 512×512 pixels. The movie stimulus was displayed at 15 Hz on a MRI-compatible projector screen that covered $24^\circ \times 24^\circ$ visual angle. Subjects were instructed to covertly search for target categories in the movies while maintaining fixation. A set of instructions regarding the experimental procedure and exemplars of the search targets were provided to the subjects before the experiment. A color square of $0.16^\circ \times 0.16^\circ$ at the center with color changing at 1 Hz was used as the fixation spot. A cue word was displayed before each run to indicate the attention target: “communication” or “locomotion”. The “communication” target contained actions with the intent of communication, including both verbal communication actions and nonverbal gestural communication actions (e.g. talking, shouting, smirking). The “locomotion” target contained locomotion-related actions with the intent of moving animate entities, including humans and anthropomorphized animals (e.g. moving, running, driving). The order of search tasks was interleaved across runs to minimize subject expectation bias. This resulted in presentation of 1800 sec of movies without repetition in each attention task. Data from the the first 20 secs and last 30 secs of each run were discarded to minimize effects of transient confounds. Following these procedures, 900 data samples for each attention task were obtained.

A separate set of functional data were collected while subjects passively viewed 120 min of natural movies (i.e. passive-viewing data). This dataset was used in constructing the semantic space and in voxel selection. Data for the passive-viewing experiment were collected in twelve 10 min 50 s runs in which 12 separate movie segments were displayed. Presentation procedures were the same between the main experiment and passive-viewing experiment, save for the number of runs. The passive-viewing data contained 3600 data samples.

4.2.4 fMRI data preprocessing.

Motion correction was performed using Statistical Parameter Mapping toolbox (SPM12 [56]). Functional volumes were aligned to the first image from the first run in each subject. Brain tissue was identified using the brain extraction tool (BET) from the FSL software package [187]. Low-frequency response components were detected using a third order Savitzky-Golay low-pass filter with 240 sec temporal window and were removed from voxel responses. Voxel responses were then z-scored to attain zero mean and unit variance. Voxels within the 2 mm neighborhood of the cortical sheet were identified as cortical voxels in each subject (S1, 37791 voxels; S2, 32671 voxels; S3, 36942 voxels; S4, 42090 voxels; S5, 39254 voxels).

4.2.5 Definition of regions of interest

To define anatomical regions of interest (ROIs) in each subject, the cortical surface was segmented into 156 regions of the Destrieux atlas [46] via Freesurfer. Segmentation results were projected from the anatomical space onto the functional space using PyCortex, and each voxel was assigned an anatomical label based on the projections. Functional ROIs were identified in each subject using visual category and retinotopic localizers [91]. Localizer experiments for visual category-selective areas (occipital face area, OFA; retrosplenial cortex, RSC) were performed in six 4.5 min runs of 16 blocks [91]. Subjects passively viewed 20 random static images from one of the objects, scenes, body parts, faces, or spatially scrambled objects groups in each block. Each image was shown for 300 ms following a 500 ms blank period. RSC was identified as voxels with positive scene versus objects contrast (t -test, $p < 10^{-4}$, uncorrected). OFA was defined using face-versus-object contrast (t -test, $p < 10^{-4}$, uncorrected). The boundaries of these areas were hand drawn on the cortical surfaces along the contours at which the contrast level reached half of the maximum. Localizer experiment for early visual areas (RET: V1, V2, V3) contained four 9 min runs. Subjects viewed

clockwise and counterclockwise rotating polar wedges in two runs. In the remaining two runs, subjects viewed expanding and contracting rings. Visual angle and eccentricity maps were used to define visual areas V1-3. Finally, ROIs were refined to voxels inside the drawn boundaries near a 2 mm neighborhood of the cortical sheet.

4.2.6 Head motion, eye movement, and physiological noise

To prevent head motion and physiological noise confounds, estimates of these nuisance factors were regressed out of the BOLD responses. Six affine motion time courses estimated during the motion-correction stage were taken as the head-motion regressors. The cardiac and respiratory activity during the main experiment were recorded using a pulse oximeter and a pneumatic belt. These data then were used to estimate two regressors to capture respiration and nine regressors to capture cardiac activity [208].

To ensure that eye-movements did not unduly bias the results, several control analyses were performed. ViewPoint EyeTracker (Arrington Research) was used to monitor subjects' eye positions at 60 Hz, after getting calibrated at the beginning of each experimental run. Kruskal-Wallis tests were used to detect systematic differences in the distribution of eye position and movement. The distribution of eye position during search for "communication" and search for "locomotion" tasks were examined. The distribution of eye position is not affected by attention condition ($p = \dots$), or by target presence or absence ($p = \dots$), and no significant interactions are present between these two factors ($p = \dots$). Next, the distribution of eye position during a 1 sec window around target onset and target offset was studied to test whether eye movement is affected by target or distractor detection. The eye position distribution is not affected by target onset ($p = \dots$) or offset ($p = \dots$), and there is no significant interaction between the aforementioned factors ($p = \dots$). Furthermore, the moving-average standard

deviation of eye position was studied in a 200 ms window to determine systematic differences in rapid moment-to-moment variations in eye position across the two tasks. There are no significant effects of attention condition ($p = \dots$), target presence or absence ($p = \dots$), target onset ($p = \dots$), or target offset ($p = \dots$), and there are no significant interactions between these factors ($p = \dots$). Finally, moving-average standard deviation of eye position was included in the model as a nuisance regressor and was regressed out of the BOLD responses.

4.2.7 Category features

A category feature space was constructed to encode the information pertaining to object and action categories in the movies. Each second of the movie stimulus was manually labeled using the WordNet lexicon [133] to find the time course for presence of 922 different object and action categories in the movie stimulus. This yielded an indicator matrix where each row represents a one-second clip of the movie stimulus and each column represents a category. Finally, category features were obtained by downsampling the indicator matrix to 0.5 Hz in order to match the acquisition rate of fMRI.

4.2.8 Motion-energy features

To infer cortical selectivity for low-level scene features, local spatial frequency and orientation information of each frame of the movie stimulus were quantified using a motion-energy filter bank. The filter bank contained 2139 Gabor filters that were computed at eight directions (0 to 315°, in 45° steps), three temporal frequencies (0, 2, and 4 Hz), and six spatial frequencies (0, 1.5, 3, 6, 12, and 24 cycles/image). Filters were placed on a square grid spanning the 24° × 24° field of view. The luminance channel was extracted from the movie frames and passed through the filter bank. The outputs were then passed through a compressive nonlinearity to yield the motion-energy features [143]. Finally, the motion-energy features were temporally downsampled to match the fMRI acquisition rate.

4.2.9 Space-Time Interest Points (STIP) features

Intermediate-level kinematic information of the movies were quantified by constructing the Space-time Interest Point (STIP) features using STIP toolbox [114]. STIP features have been successfully leveraged in many computer vision applications to recognize human actions [115]. The STIP toolbox first segmented the video into distinct spatiotemporal patches. Then each patch was scanned to identify patches with high spatial or temporal variance as interest points. Histograms of oriented gradients (HoG [40]) and histograms of optical flow (HoF [85]) for these interest points was taken as the collection of 162 STIP features. Finally, STIP features were downsampled to match the acquisition rate of fMRI.

4.2.10 Model estimation and testing

Linearized models were fit in each voxel to estimate model weights that map each set of features (i.e. category, motion-energy, or STIP features) to the measured BOLD responses in each attention condition in individual subjects. Spatially-informed regularized linear regression [28] with separate regularization terms across feature (λ_f) and neighborhood (λ_n) dimensions was used to fit the models. To capture the hemodynamic response, delayed feature time-courses were concatenated. Delays of two, three, and four samples, corresponding to 4, 6, and 8 sec were used. To account for potential correlations between target detection and BOLD responses, a nuisance target-presence regressor was included in the model. The target-presence regressor contained category regressor for “communication” during search for “communication” task and the category regressor for “locomotion” during search for “locomotion” task. Model fitting for the two attention tasks was performed concurrently by concatenating the features and BOLD responses across tasks (Fig. 4.2). This procedure ensured consistency between the assigned regularization parameters across tasks, and enabled employing the target regressor [179].

A nested cross-validation (CV) procedure was used to choose the regularization

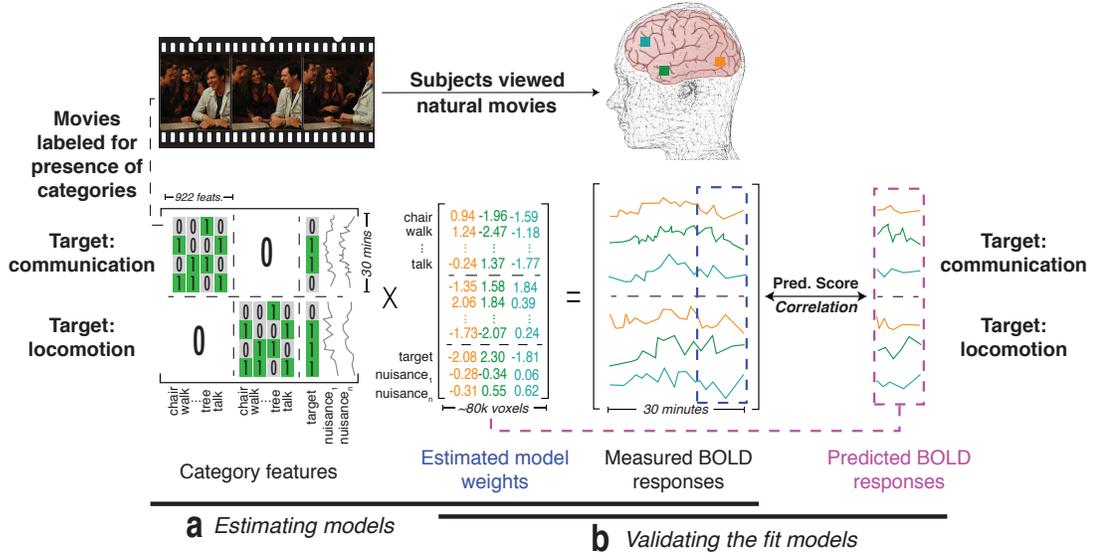


Figure 4.2. Model fitting and validation procedure. While undergoing fMRI, five subjects viewed 60 mins of natural movies and covertly searched for “communication” or “locomotion” action categories while fixating. **a.** An indicator matrix was constructed that identified presence of each of the 922 object and action categories in each 1-sec clip of the movies. Nuisance regressors were included to account for head-motion, physiological noise, and eye-movement confounds. A target regressor was included to account for target detection confounds. Category model weights (i.e. category responses) were estimated that mapped each category feature to the BOLD responses recorded in each attention task. **b.** Accuracy of the fit models were assessed by measuring their performance in predicting held-out BOLD responses, after discarding the nuisance regressors and the target regressor. Prediction score of the fit models was taken as Pearson’s correlation coefficient between estimated and measured responses.

parameters and estimate model weights. Data from the main experiment were segmented into 60 30-second blocks. In each of the 10 outer folds, 4 randomly chosen blocks were held-out as validation data. Then, in each of the 10 inner folds, 54 randomly chosen blocks were used as training data and the 2 remaining blocks were used as test data. To fit models for the passive-viewing data, data were segmented into 144 50-second blocks. In each fold, 8 randomly chosen blocks were held-out as validation data, 132 randomly chosen blocks were used as training data and the 4 remaining blocks were used as test data. We used 10 regularization parameters across the feature dimension in the range $\lambda_f \in [2^5, 2^{17}]$.

Similarly, 10 regularization parameters across the neighborhood dimension in the range $\lambda_n \in [2^{10}, 2^{20}]$ were used. Training data were used to fit models for each (λ_f, λ_n) pair independently. Model weights were then used to predict responses in the test data and prediction scores of the fit models were assessed. Prediction scores were taken as Pearson’s correlation coefficient between actual and predicted voxel responses. The pair of (λ_f, λ_n) maximizing the average prediction score across inner CV folds was chosen in each voxel. Finally, the optimal pair of parameters were used to fit models on the union of training and test data in each outer fold and model weights were averaged across the outer folds.

Prediction performance of the fit models were then evaluated. In each outer fold, after discarding the nuisance regressors, responses were predicted for the validation data using the fit models and prediction scores were averaged across the attention tasks. Finally, prediction scores were averaged across the outer folds.

4.2.11 Action category responses

Category features were used to fit category models (i.e. to estimate category responses) in each voxel for the passive-viewing data and the for data recorded during the two search tasks. Category responses reflect the voxel tuning for each of the 922 object and action categories in the movie stimulus. To infer tuning for action categories, 922-dimensional category responses were masked to select only the 109 action categories. This yielded the voxelwise 109-dimensional action category responses.

4.2.12 Variance partitioning analysis

Object-action categories can be correlated with low-level visual features of natural movies [119]. Moreover, there is evidence for representation of intermediate-level features of actions (e.g. action kinematics) across the cortex [101]. Therefore,

there is a possibility that the estimated category responses be confounded by selectivity for low- and intermediate-level scene features. To control for this, we performed a variance partitioning analysis. This analysis estimates the response variance that is uniquely explained by the category model after accounting for variance that can be attributed to low- and intermediate-level features captured by the motion-energy and STIP models. To do this, we separately measured the variance explained when all three models (category, motion-energy, and STIP) are fit simultaneously (i.e. “combined model”), and variance explained when motion-energy and STIP models are fit simultaneously (i.e. “control model”). The explained variance (R^2) was calculated as square of the prediction scores, separately for the combined and control models. Note that from a model fitting perspective, negative prediction scores correspond to zero explained variance. Finally, unique variance explained by the category model was calculated as

$$\hat{R}_{cat}^2 = R_{comb}^2 - R_{cont}^2 \quad (4.1)$$

where \hat{R}_{cat}^2 is the variance uniquely explained by the category model after accounting for low- and intermediate-level features, R_{comb}^2 is the variance explained by the combined model, and R_{cont}^2 is the variance explained by the combination of motion-energy and STIP models in each voxel.

4.2.13 Semantic representation of actions

Passive-viewing data were used to construct a continuous semantic space for action category representation. In this space, semantically similar action categories would project to nearby points, whereas semantically dissimilar categories would project to distant points [91]. Voxelwise models were fit and action category responses during passive viewing were estimated. A group semantic space was then obtained using principal component analysis (PCA) on the action category responses of cortical voxels pooled across all subjects. To maximize the quality of the semantic space, voxels in which the category model predicted unique response variance after accounting for the variance attributed to low- and intermediate-level scene features were selected. These voxels were further refined to include the

ones that were among the top 3,000 best predicted voxels in each subject. The collection of the first 13 principal components (PCs) that explained more than 90% of the variance in responses was selected. Semantic tuning profile for each voxel under each attention task was then obtained by projecting the respective action category responses onto the PCs.

To ease the visualization of the semantic space, action categories were manually clustered and cluster centers were projected onto the semantic space (see Fig. 4.4a, Supp. Table 4.1). To identify cluster assignments action categories were visualized on the WordNet hierarchical tree, resulting in nine separated action clusters. Projections of cluster centers onto the semantic space were then obtained as the average of the projections of action categories within each cluster.

Table 4.1. Clusters of action categories.

Cluster label	Actions
communication	smile, communicate, nod, indicate, grimace, read, cellphone, talk, write, argue, shout, pray, gesticulate, yell, smirk
locomotion	fly, climb, ride horseback, chase, hike, ride, crash, drive, scurry, dive, swim, descend, walk, rise, travel, edge, rush, gallop, pounce, pass, fall, canter, ski, skid, run, crawl
body-change	chew, change, grow, lean, arise, break, bend, crouch, struggle
breathe	breathe, yawn, huff, sigh
touch	punch, drive, lick, snog, strike, touch, caress, slam, hit
change-shape	affect, change, stretch, shape, raise
consume	eat, drink, consume
object-move	decant, put, kick, lift, push, propel, stand, jab, turn, drag, slide, pull, clap, move, pour, fasten, connect, act
self-move	dance, turn, tumble, exit, reach, spin, revolve, jump, flip, hop, bounce, sneak

To better visualize the distribution of actions across semantic space, 109 action categories in the movies were manually clustered. Action categories were visualized on the WordNet hierarchy, yielding 9 segregated action clusters (Fig. 4.4a).

4.2.14 Characterizing tuning shifts

Attentional tuning shifts toward or away from targets would be reflected in modulation of semantic tuning strengths for “communication” or “locomotion” action categories. Thus, the magnitude and direction of tuning shifts can be assessed by comparing the semantic tuning strengths for these categories between the two attention condition. Tuning strengths for the two target categories were quantified as the similarity between semantic tuning profiles and idealized templates tuned solely for “communication” or “locomotion” action categories. First, idealized category responses were constructed as 109-dimensional vectors containing ones for target categories (i.e. communication or locomotion categories) and zeros elsewhere. Idealized templates were then obtained by projecting these idealized category responses onto the semantic space. Tuning strength for each target category was then quantified as Pearson’s correlation coefficient between voxel semantic tuning profile and the corresponding template

$$T_{i,C} = \text{corr}(s_i, s'_C) \quad (4.2)$$

$$T_{i,L} = \text{corr}(s_i, s'_L) \quad (4.3)$$

where $T_{i,C}$ and $T_{i,L}$ are the tuning strength for “communication” and “locomotion” during task $i \in \{C, L\}$ denoting attend to “communication” or attend to “locomotion”, s_i is the semantic tuning profile during task i , and s'_C and s'_L denote the idealized semantic tuning templates for “communication” and “locomotion”, respectively. Finally, voxelwise tuning shift index (TSI) was quantified as

$$TSI = \frac{(T_{C,C} - T_{C,L}) + (T_{L,L} - T_{L,C})}{2 - \text{sign}(T_{C,C} - T_{C,L})T_{C,L} - \text{sign}(T_{L,L} - T_{L,C})T_{L,C}} \quad (4.4)$$

Tuning shifts toward the attended category would yield positive TSIs where a TSI of 1 indicates a complete match between voxel semantic tuning and idealized templates, whereas negative TSIs would indicate shifts away from the attended category where a TSI of -1 indicates a complete mismatch between voxel tuning and idealized templates. Moreover, a TSI of 0 would indicate that the voxel tuning did not shift between the two attention conditions. Furthermore, individual TSIs

during each attention task were quantified as

$$TSI_C = \frac{T_{C,C} - T_{C,L}}{1 - \text{sign}(T_{C,C} - T_{C,L})T_{C,L}} \quad (4.5)$$

$$TSI_L = \frac{T_{L,L} - T_{L,C}}{1 - \text{sign}(T_{L,L} - T_{L,C})T_{L,C}} \quad (4.6)$$

where TSI_C and TSI_L denote the TSI during search for communication and locomotion actions respectively. To study the tuning shifts in an ROI, TSIs were averaged across uniquely explained voxels within the ROI.

4.3 Results

Recent studies suggest that representations of thousands of object and action categories in natural scenes are embedded in a semantic space the axes of which are mapped continuously across the cortex [91] and search for object categories warps this space in favor of targets [37]. Yet, it is currently unknown whether and where in the brain natural visual search for action categories alters semantic representations. To investigate this question, we estimated voxelwise tuning for hundreds of object and action categories across cortex. While undergoing fMRI, five human subjects viewed 60 minutes of natural movies and attended to “communication”, or “locomotion” action categories in separate runs. Category regressors were constructed that labeled presence of 922 distinct object and action categories in the movies. Separate category models for each attention task were then fit in each voxel. This enabled us to measure single-voxel category responses during the two attention tasks, reflecting the proportional contribution of each object and action category to the evoked responses (Fig. 4.2, see *Experimental Procedures*).

4.3.1 Attention alters category tuning profiles

Natural stimuli contain correlations among various levels of features and there is a possibility that estimated category responses be confounded by voxel tuning

for low- and intermediate-level scene features. To address this issue, we identified voxels in which the category model explained unique response variance after accounting for these alternative features by performing a variance partitioning analysis on a separate dataset collected for this purpose (i.e. “passive-viewing dataset”, see *Experimental Procedures*). We quantified low-level and intermediate-level scene information by constructing the motion-energy and spatio-temporal interest point (STIP) features respectively. We find that the category model explains unique response variance after accounting for low- and intermediate-level features in $38.9 \pm 0.38\%$ of cortical voxels (mean \pm sem across five subjects), yielding 12,248-16,845 voxels in individual subjects that were used in the subsequent analyses (i.e. “uniquely predicted voxels”). Note that using the passive-viewing dataset to perform the variance partitioning analysis prevents confounding the voxel selection procedure by potential attentional modulations of the category responses.

Functional inferences on the estimated category responses will be valid, only if the fit category models successfully predict BOLD responses that were held-out during model fitting. To assess model performance, we measured average prediction scores across the two attention tasks, taken as Pearson’s correlation coefficient between the predicted and measured held-out responses. Category models have high prediction scores (greater than 1 std above the mean) in $39.3 \pm 0.2\%$ of the uniquely predicted voxels. These include many voxels across the AON including occipitotemporal, parietal, and premotor cortices, as well as areas in prefrontal and cingulate cortices (Fig. 4.3).

Previous reports evidenced for modulations of single-voxel tuning profiles during search for object categories [37]. This raises the possibility that visual search for action categories could also modulate category tuning profiles in single voxels. Thus, the category responses estimated for individual attention conditions should yield greater prediction scores compared to a null model fit by pooling data across conditions. To test this possibility, we compared the prediction scores obtained from the estimated category responses to those obtained using the null model. We find that the category model significantly outperforms the null model in $44.7 \pm 1.3\%$ of uniquely predicted voxels (bootstrap test, $q(\text{FDR}) < 0.05$).

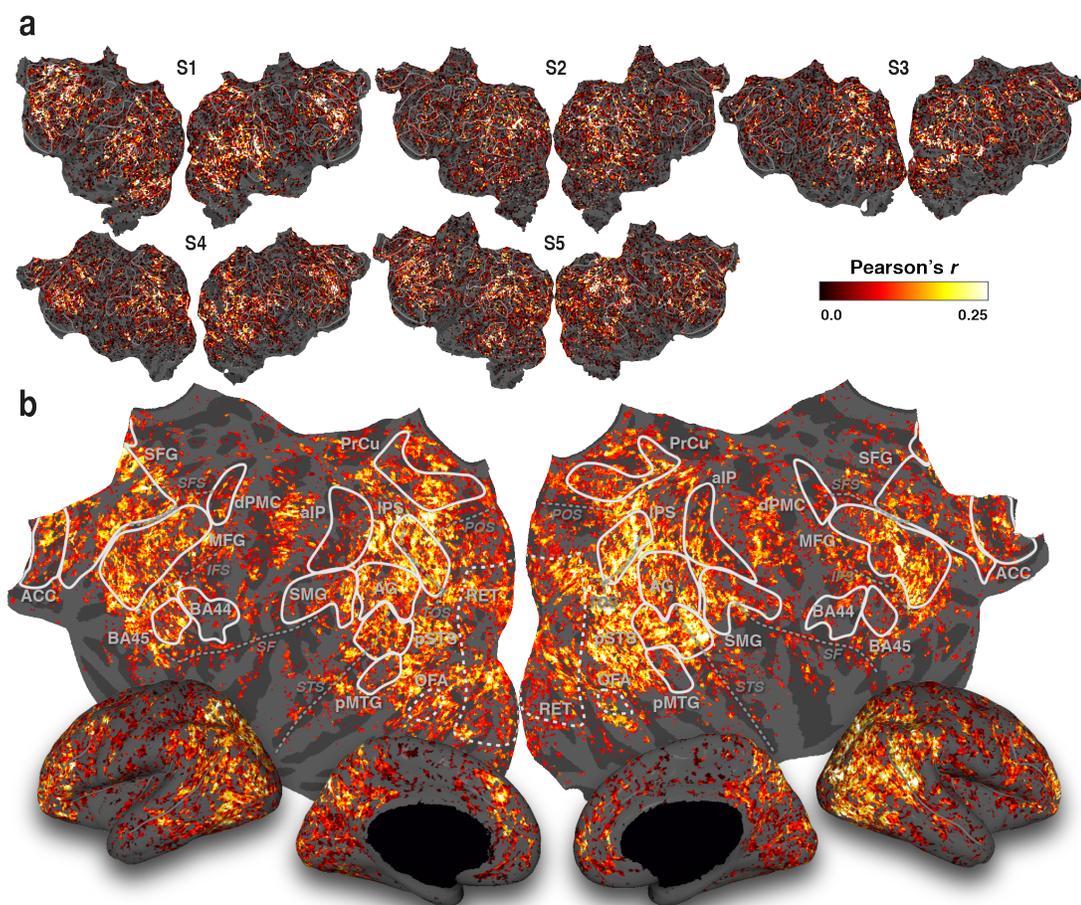


Figure 4.3. Prediction performance of the category model. **a.** Category model performance (i.e. prediction score) plotted on the flattened cortical surface of individual subjects. Model prediction score was taken as the Pearson's correlation coefficient between the held-out measured BOLD responses and the responses predicted by the model. A variance partitioning analysis was used to quantify the variance uniquely explained by the category model after accounting for low- and intermediate-level stimulus features (see *Experimental Procedures*). Voxel for which unique explained response variance of the category model after accounting for these features was not different than zero are hidden, revealing the cortical curvature below. **b.** Average prediction score across subjects. Prediction score in individual subjects was projected onto the standard brain template from Freesurfer. Projected prediction scores were then averaged across subjects. The category model well predicts response across ventral-temporal, parietal, and frontal cortices. Regions of interest are illustrated by white lines: pMTG, posterior middle temporal gyrus; pSTS, posterior superior temporal sulcus; AG, angular gyrus; SMG, supramarginal gyrus; IPS, intraparietal sulcus; aIP, anterior intraparietal cortex; PrCu, precuneus; dPMC, dorsal premotor cortex; BA44/45, Brodmann area 44/45; MFG, middle frontal gyrus; SFG, superior frontal gyrus; [continued on next page]

[continued] ACC, anterior cingulate cortex; RET, early visual areas V1-3; OFA, occipital face area; RSC, retrosplenial cortex. Several important sulci are illustrated by dashed gray lines: TOS, temporo-occipital sulcus; STS, superior temporal sulcus; SF, Sylvian fissure; IFS, inferior frontal sulcus; MFS, middle frontal sulcus; SFS, superior frontal sulcus. Data sets are available at <http://icon.bilkent.edu.tr/brainviewer/shahdlooetal2020>

Moreover, control analyses revealed that these attentional modulations cannot be attributed to eye-movements, head-motion, physiological noise, and target detection biases (see *Experimental Procedures*). These results can be taken to infer that many voxels across cortex encode high-level information pertaining to object and action categories, and that single-voxel category tuning profiles get significantly modulated by search for action categories.

4.3.2 Attention warps semantic representation of actions

The fit category models identify tuning for 813 object and 109 action categories. To investigate semantic representation of these 109 action categories, we estimated a continuous semantic space via principal component analysis (PCA) on action category responses. To prevent overfitting, category responses estimated via the passive viewing dataset were used to estimate the semantic space. Action categories that are semantically similar to each other would project to nearby points in this space, whereas semantically dissimilar categories would project to distant points. To examine semantic information embedded within this space, we visualized it by projecting action categories onto the semantic dimensions (i.e. principal components, PCs; Fig. 4.4a). The first PC seems to distinguish between self-movements (e.g. “chew”, “yawn”, “eat”) and actions involving objects or other humans (e.g. “communicate”, “touch”, “drive”). The second PC seems to distinguish between dynamic versus static actions (e.g. “raise”, “propel”, “dance” versus “breath”, “drink”). The third PC seems to distinguish between human-related actions (e.g. “communicate”, “jump”, “crouch”) and dynamic actions that involve objects (e.g. “drive”, “punch”, “drag”). These observations suggest that this estimated space well captures the semantic variance in action

categories.

We expect that modulation of single-voxel category responses during visual search for action categories be reflected in modulation of semantic tuning profiles. To test this prediction, we obtained single-voxel semantic tuning profiles by projecting action category responses onto the semantic space. The first and third semantic dimensions maximally differentiated between actions belonging to the two target categories (i.e. “communication” versus “locomotion” categories, see Fig. 4.4a). To maximize the sensitivity in visualization of attentional modulations, we compared the projections onto these semantic dimensions across the two attention tasks (Fig. 4.4b, Supp. Figs. 4.5-4.9). We observe that attention causes semantic tuning shifts in many cortical voxels. Specifically, many voxels in inferior posterior parietal cortex (PPC), cingulate cortex, and anterior inferior prefrontal cortex shift their tuning toward communication during search for communication actions, whereas voxels in superior PPC, and medial parietal cortex shift their tuning toward locomotion during search for locomotion actions. Moreover, many voxels in areas with generic selectivity for actions, such as angular gyrus and supramarginal gyrus (AG, SMG), shift their tuning toward targets irrespective of the search target. Previous studies suggest a functional segregation along the superior-inferior PPC, such that the inferior PPC gets activated during observation of communication actions, while observation of locomotion actions activates the superior PPC [171, 1, 34]. Consistent with this view, our finding here suggests that during search for a given action category, tuning shift toward the target category is most prominent in voxels that are selective for the target. Whereas, cortical areas with generic action selectivity shift their semantic tuning toward targets irrespective of the target identity.

4.3.3 Distribution of tuning shifts across cortex

To quantitatively assess semantic tuning changes during search for action categories, we first measured tuning strengths for “communication” and “locomotion” categories. Then, we calculated a tuning shift index ($TSI \in [-1, 1]$) for each voxel

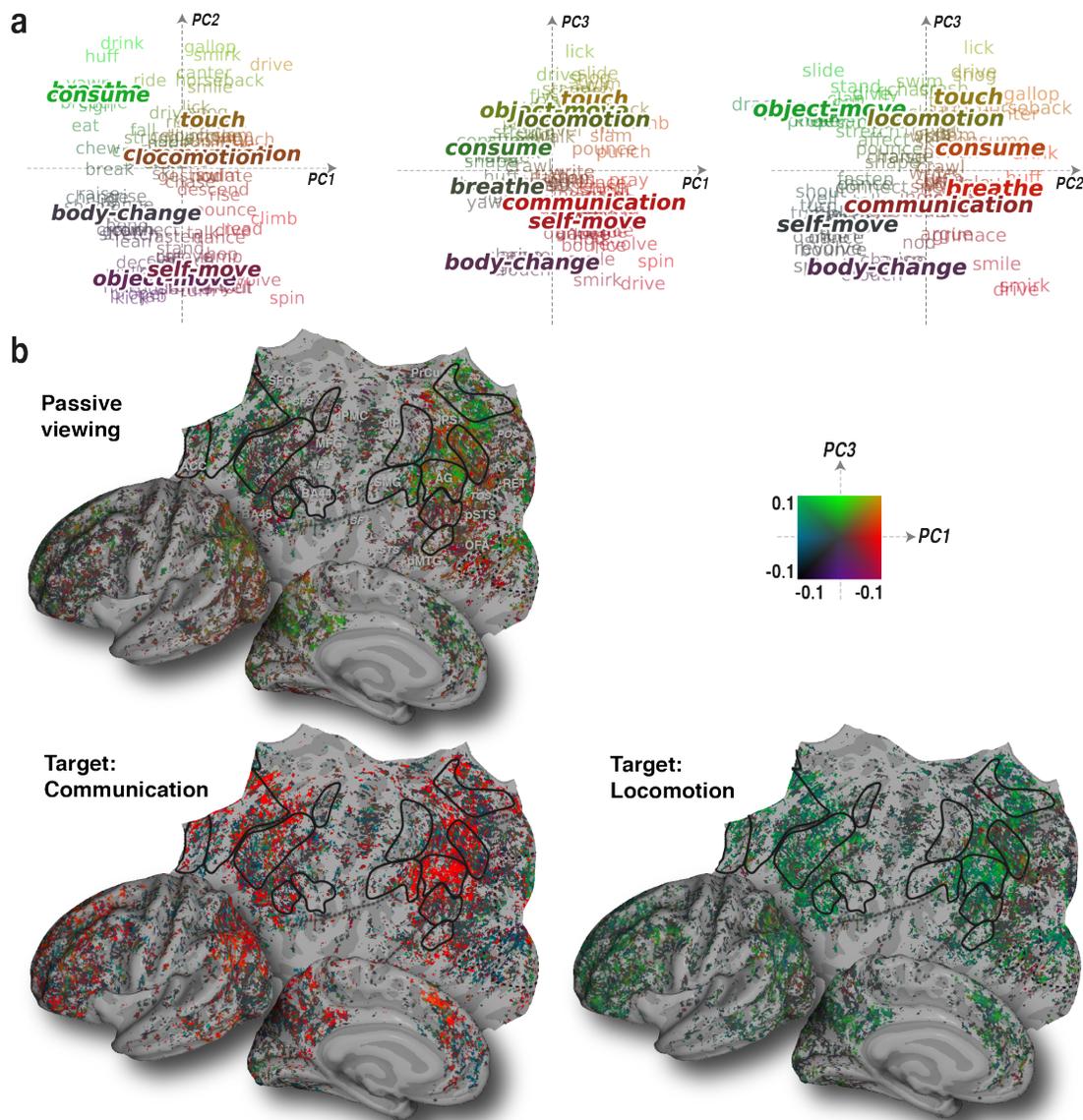


Figure 4.4. Semantic space underlying action category responses. Principal components analysis on action category responses during passive viewing was performed to estimate a semantic space. **a.** To illustrate the semantic information embedded within semantic space, action categories, along with centers of action category clusters, were projected onto the semantic space (see *Methods*). The semantic space well captures the semantic variability among action categories. The “communication” and “locomotion” actions are best separated across the first and third semantic dimensions (i.e. principal components, PCs). **b.** Action category responses during passive viewing and during the two attention tasks in individual subjects were projected onto the semantic space and a two-dimensional colormap was used to color each voxel based on the projection values across first and third semantic dimensions. The projection values were then averaged across subjects (see Figs. 4.5-4.9 [continued on next page])

[continued] for individual subjects). Voxels where category model did not explain unique response variance after accounting for low- and intermediate stimulus features are hidden, revealing the cortical curvature below. Many voxels across cortex occipitotemporal, parietal, and prefrontal cortices shift their tuning toward targets. Specifically, voxels in inferior posterior parietal cortex shift their tuning toward “communication” during search for “communication” categories. Whereas, voxels in superior posterior parietal cortex shift their tuning toward “locomotion” during search for “locomotion” categories.

to quantify the difference in tuning strengths for target versus distractor categories. Tuning shift toward the attended category would yield a positive TSI, whereas tuning shift toward the distractor category would yield a negative TSI, and a TSI of 0 means that the voxel tuning is not modulated at all under the two attention conditions (see *Experimental Procedures*). We observe that voxels across many different cortical regions shifted their tuning toward the attended category (Fig. 4.10a,b). Figure 4.10c shows tuning shifts in several common regions of interest (ROIs). Tuning shifts are significantly greater than zero in many areas across AON including lateral occipitotemporal cortex (pSTS, pMTG), posterior parietal cortex (IPS, AG, SMG), and premotor cortex (BA44/45; bootstrap test $p < 0.05$). Previous studies provide evidence for attentional modulations in lateral occipitotemporal and premotor cortices during visual search for action categories [137, 138]. Extending these reports, here we find more broadly spread attentional tuning shifts at all levels of the AON hierarchy. Further to this, we find that average tuning shift in AG and SMG is significantly higher than TSI in occipitotemporal (pSTS, pMTG) and premotor cortex (dPMC, BA44/45; $p < 0.05$). Recent studies suggest AG and SMG to be central nodes across a cortical network involved in semantic processing [35, 15, 192, 51, 120]. Thus, the reported tuning shifts here proposes a distinctive account for AG and SMG among other AON nodes, according which these areas facilitate action perception by maintaining semantic representation of task-relevant action categories.

We recently reported that object-category-based visual search causes tuning shifts toward targets in prefrontal cortex [37]. In line with that study, here we find that visual search for action categories causes significant tuning shifts in middle frontal and superior frontal cortices. Moreover, we find that tuning

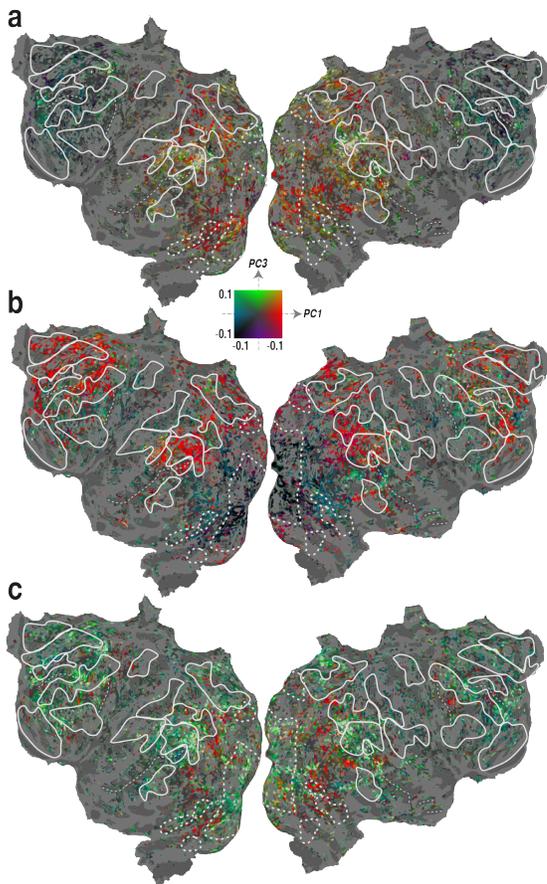


Figure 4.5. Cortical flat maps of projections onto the semantic space for subject S1. Action category responses during (a.) passive viewing, (b.) search for “communication”, and (c.) search for “locomotion” categories for subject S1 were projected onto the semantic space and a two-dimensional colormap was used to color each voxel based on the projection values across first and third semantic dimensions. Voxels where category model did not explain unique response variance after accounting for low- and intermediate stimulus features are hidden, revealing the cortical curvature below. Many voxels across cortex occipitotemporal, parietal, and prefrontal cortices shift their tuning toward targets.

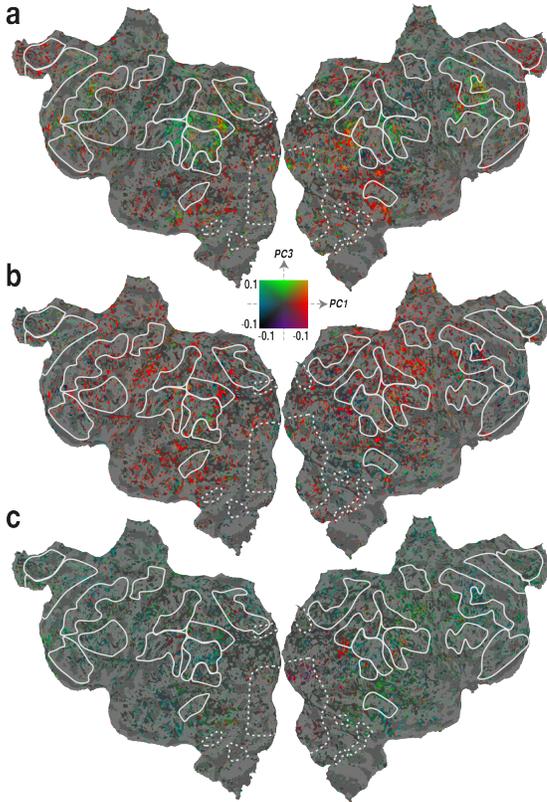


Figure 4.6. Cortical flat maps of projections onto the semantic space for subject S2. Action category responses during (a.) passive viewing, (b.) search for “communication”, and (c.) search for “locomotion” categories for subject S2 were projected onto the semantic space and a two-dimensional colormap was used to color each voxel based on the projection values across first and third semantic dimensions. Voxels where category model did not explain unique response variance after accounting for low- and intermediate stimulus features are hidden, revealing the cortical curvature below. Many voxels across cortex occipitotemporal, parietal, and prefrontal cortices shift their tuning toward targets.

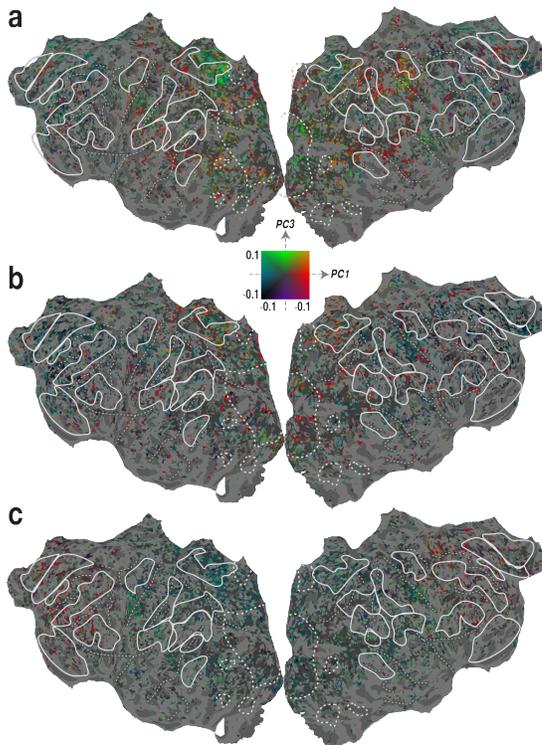


Figure 4.7. Cortical flat maps of projections onto the semantic space for subject S3. Action category responses during (a.) passive viewing, (b.) search for “communication”, and (c.) search for “locomotion” categories for subject S3 were projected onto the semantic space and a two-dimensional colormap was used to color each voxel based on the projection values across first and third semantic dimensions. Voxels where category model did not explain unique response variance after accounting for low- and intermediate stimulus features are hidden, revealing the cortical curvature below. Many voxels across cortex occipitotemporal, parietal, and prefrontal cortices shift their tuning toward targets.

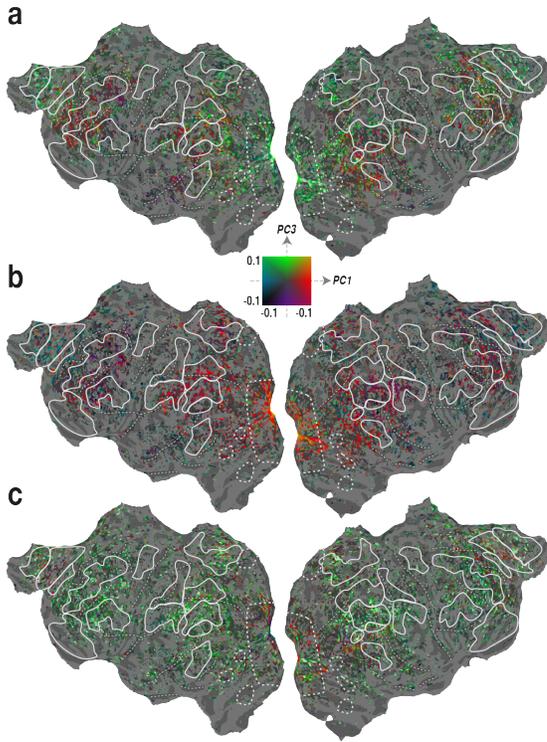


Figure 4.8. Cortical flat maps of projections onto the semantic space for subject S4. Action category responses during (a.) passive viewing, (b.) search for “communication”, and (c.) search for “locomotion” categories for subject S4 were projected onto the semantic space and a two-dimensional colormap was used to color each voxel based on the projection values across first and third semantic dimensions. Voxels where category model did not explain unique response variance after accounting for low- and intermediate stimulus features are hidden, revealing the cortical curvature below. Many voxels across cortex occipitotemporal, parietal, and prefrontal cortices shift their tuning toward targets.

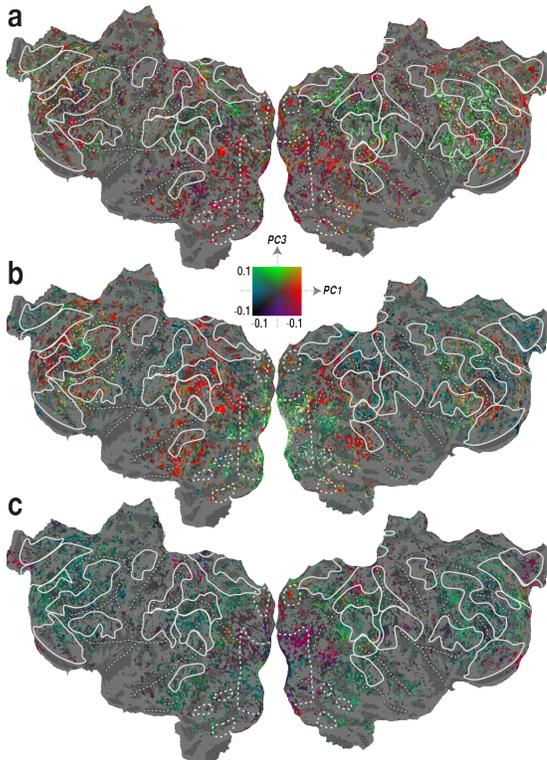


Figure 4.9. Cortical flat maps of projections onto the semantic space for subject S5. Action category responses during (a.) passive viewing, (b.) search for “communication”, and (c.) search for “locomotion” categories for subject S5 were projected onto the semantic space and a two-dimensional colormap was used to color each voxel based on the projection values across first and third semantic dimensions. Voxels where category model did not explain unique response variance after accounting for low- and intermediate stimulus features are hidden, revealing the cortical curvature below. Many voxels across cortex occipitotemporal, parietal, and prefrontal cortices shift their tuning toward targets.

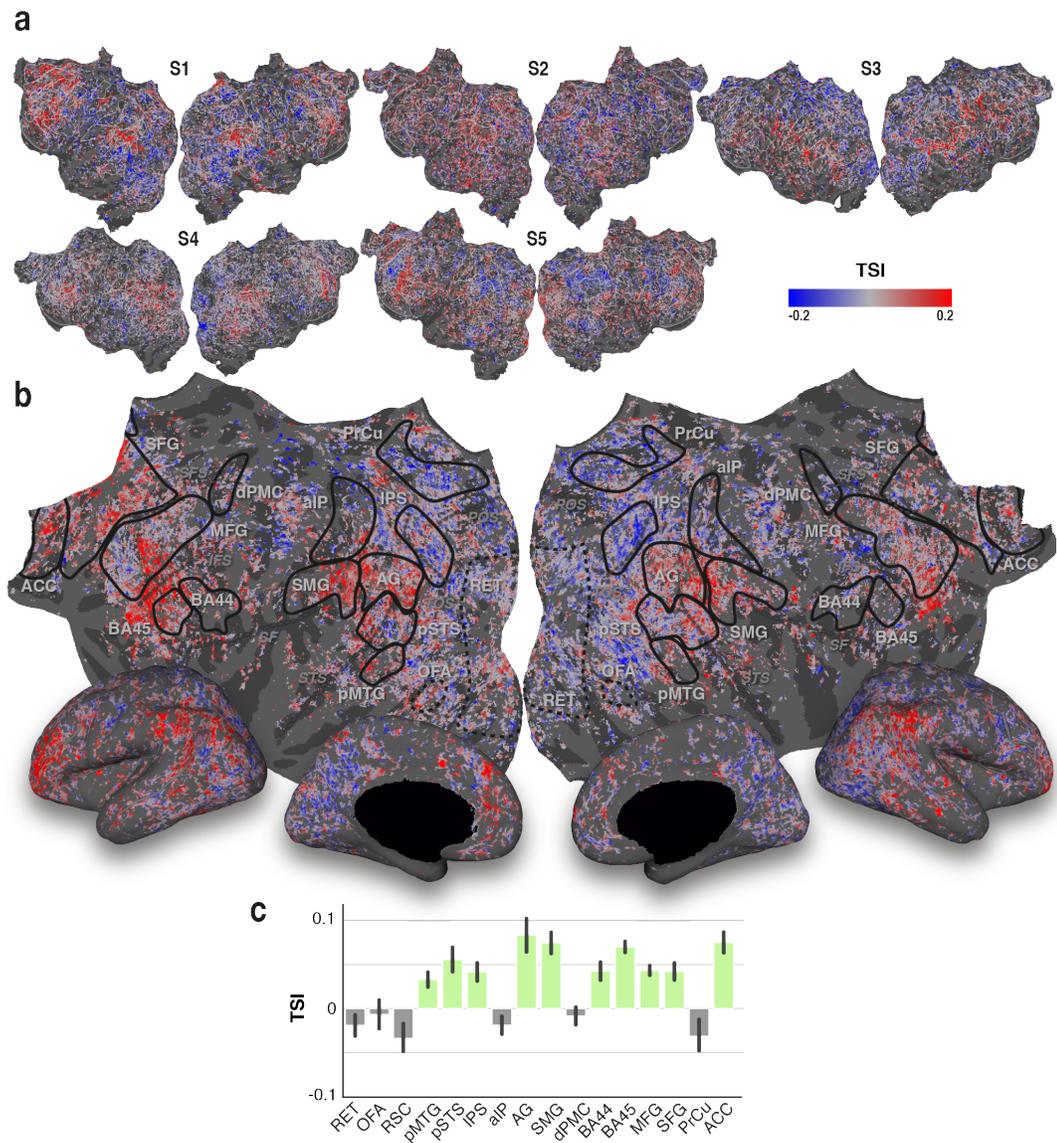


Figure 4.10. Cortical distribution of tuning shifts. For each voxel, a tuning shift index (TSI) was calculated to quantify the attentional changes in tuning for the attended versus unattended categories. Tuning shifts toward the attended category would yield positive $TSI \in (0, 1]$ where a TSI of 1 indicates a complete match between voxel semantic tuning and idealized templates tuned solely for targets, whereas negative $TSI \in [-1, 0)$ would indicate shifts away from the attended category where a TSI of -1 indicates a complete mismatch between voxel tuning and idealized templates. A TSI of 0 would indicate that the voxel tuning did not shift between the two attention conditions. **a.** TSI values plotted on the flattened cortical surfaces of the five subjects. Voxels where category model did not explain unique response variance after accounting for low- and intermediate stimulus features are hidden, revealing the cortical curvature below. Cortical distribution of TSI values is consistent across subjects. **b.** TSI values from individual subjects were projected onto the standard *[continued on next page]*

[continued] brain template and averaged across subjects. Voxels across many cortical regions shifted their tuning toward the attended category. These include regions across AON (occipitotemporal cortex, posterior parietal cortex, and premotor cortex), lateral prefrontal cortex, and anterior cingulate cortex. **c.** TSI in cortical areas (mean \pm sem across five subjects). Significant mean values are denoted by green bars and gray bars denote non-significant mean values (bootstrap test, $p > 0.05$). Tuning shift is significantly greater than zero in regions across AON including lateral occipitotemporal cortex (pSTS, pMTG), posterior parietal cortex (IPS, AG, SMG), and premotor cortex (BA44, BA45), and in regions across prefrontal and cingulate cortices (SFG, ACC). Data sets are available at <http://icon.bilkent.edu.tr/brainviewer/shahdlooetal2020>

shift in anterior cingulate cortex (ACC) is significantly greater than zero ($p < 0.05$). This is in accord with reports suggesting that ACC facilitates detection of targets among distractors in cluttered scenes by enhancing representation of targets [42, 216, 215]. In contrast, we observe that many voxels across precuneus (PrCu), medial parietal cortex, and temporoparietal junction shift their tuning toward the distractor category. This observation is consistent with the view that these areas enhance visual search by error monitoring and distractor detection [33].

4.3.4 Tuning shifts for unattended categories

Previous evidence suggest that natural visual search for object categories shifts semantic tuning for both attended and unattended object categories [37]. Thus, we asked whether visual search also shifts semantic tuning for unattended action categories. To answer this question, we calculated TSI for unattended categories. Note that the modeling framework used here allows us to estimate category responses for 109 distinct action categories. Thus, by masking the 109-dimensional action category response vectors to select the unattended categories, TSI can be calculated specifically for these categories. We observe that nearly across all the cortex tuning shifts for unattended categories are smaller than the overall tuning shifts (Fig. 4.11a,b). Specifically, TSI is non-significant in all of the studied ROIs, save for AG, SMG, and BA45 (Fig. 4.11c). Moreover, in these areas (AG, SMG, BA45), overall tuning shifts are significantly greater than tuning shifts for

unattended categories (bootstrap test, $p < 0.01$). These results suggest that representation of action categories in parietal and premotor AON nodes is relatively more dependent on the search task than that at lateral occipitotemporal areas. Moreover, they suggest that tuning shifts for attended categories accounted for a relatively larger fraction of the overall tuning shifts compared to unattended categories.

4.3.5 Tuning shifts are influenced by intrinsic selectivity for action categories

Recent studies suggest that voxels that are intrinsically selective for specific object categories during passive viewing retain their tuning for the preferred category (e.g. “humans” for fusiform face area, and “vehicles” for parahippocampal place area) even when a non-preferred category is the search target [163, 164, 179]. Thus, we expect that tuning shifts during search for individual action categories should be influenced by intrinsic selectivity for targets. To investigate this, we calculated separate tuning shift indices during search for communication actions (TSI_C) and during search for locomotion actions (TSI_L). We then projected TSI_C and TSI_L values onto cortical flat maps (Fig. 4.12a,b), and examined them in several common ROIs (Fig. 4.12c). We observe that voxels across posterior parietal and prefrontal cortices shift their semantic tuning toward the attended category irrespective of the target identity. Yet, many voxels across anterior parietal, occipital, and cingulate cortices shift their tuning toward the attended category only during search for communication or during search for locomotion categories.

4.3.5.1 Areas where both TSI_C and TSI_L are non-significant

TSI_C and TSI_L are non-significant in early visual areas (RET; bootstrap test, $p > 0.05$), and in face- and place-selective areas (OFA, RSC; $p > 0.05$). Furthermore, TSI_C and TSI_L are non-significant in anterior intraparietal cortex (aIP; $p > 0.05$).

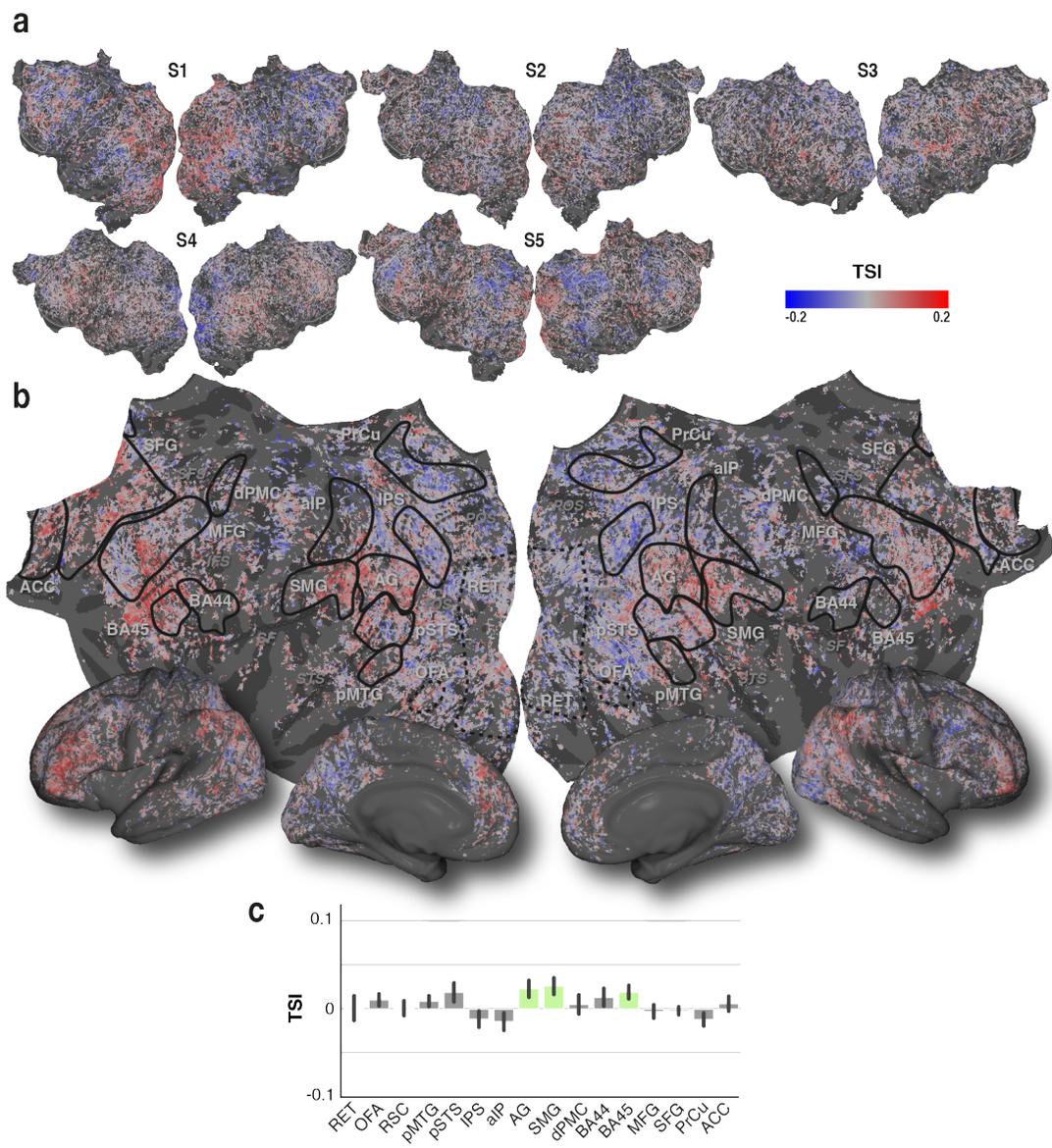


Figure 4.11. Tuning shifts for unattended categories. To examine how representation of unattended action categories changes during visual search, we investigated the tuning shift for these categories. **a.** TSI values plotted on the flattened cortical surfaces of the five subjects. Voxels where category model did not explain unique response variance after accounting for low- and intermediate stimulus features are hidden, revealing the cortical curvature below. Cortical distribution of TSI values is consistent across subjects. **b.** TSI values from individual subjects were projected onto the standard brain template and averaged across subjects. Tuning shift for unattended categories is lower compared to the the tuning shift for all action categories, shown in Fig. 4.10. **c.** TSI in cortical areas (mean \pm sem across five subjects). Significant mean values are denoted by green bars and gray bars denote non-significant mean values (bootstrap test, $p > 0.05$). Tuning shifts in most of the cortical areas [continued on next page]

[continued] are non-significant. Tuning for unattended categories in voxels in posterior parietal cortex (AG, SMG) and anterior inferior frontal cortex (BA45) shifts toward targets.

Previous studies suggest that aIP is selective for manipulative actions and not for communication or locomotion actions [170, 146]. Thus, negligible tuning changes in aIP during search for communication or locomotion actions can be attributed to the lack of tuning for these action categories.

4.3.5.2 Areas where only TSI_C or TSI_L are significant

TSI_C is significantly greater than zero in anterior inferior frontal gyrus (BA44/45), in superior frontal gyrus (SFG), and in ACC ($p < 0.05$). On the other hand, TSI_L is significantly greater than zero in intraparietal sulcus (IPS) and in angular gyrus (AG; $p < 0.05$). Previous studies suggest that activity in anterior inferior frontal gyrus enhances the representation of interpersonal communicative actions [93, 186]. Moreover, lateral and medial prefrontal cortices have been suggested to be causally involved in communicative cognition [204, 217, 103]. Furthermore, previous reports provide evidence for representation of space and animate locomotion actions in superior PPC [21, 7, 95, 1]. These lines of evidence suggest that in areas that are selective for specific action categories, visual search for the preferred action category alters semantic representation in favor of the target. Furthermore, these attentional effects are not limited to the areas belonging to AON, but rather they extend to higher-order cognitive areas that facilitate action perception.

In dorsal premotor cortex (dPMC), TSI_L is significantly less than zero, whereas TSI_C is non-significant ($p > 0.05$). Several previous studies report that dPMC enhances representation of locomotion actions by facilitating the detection of distractors [2, 197, 225]. In accord with these reports, our finding here suggests that dPMC enhances the representation of distractors during search for locomotion categories.

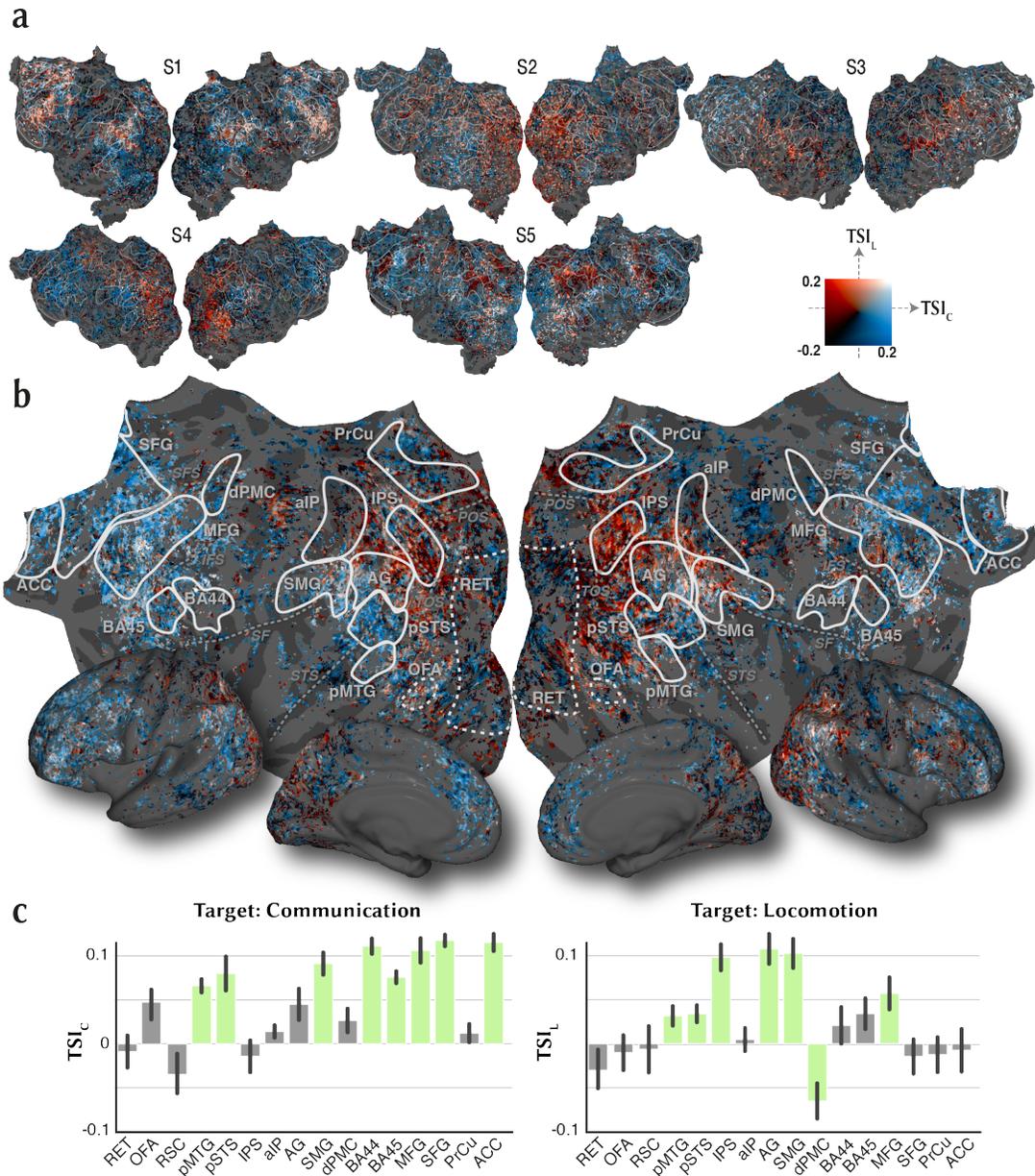


Figure 4.12. Tuning shifts during search for individual targets. Separate tuning shift indices were calculated during search for “communication” (TSI_C), and “locomotion” (TSI_L) categories. **a.** TSI_C and TSI_L values plotted on the flattened cortical surfaces of the five subjects. A two-dimensional colormap was used to color each voxel based on TSI_C and TSI_L values. Voxels where category model did not explain unique response variance after accounting for low- and intermediate stimulus features are hidden, revealing the cortical curvature below. **b.** TSI_C and TSI_L values were projected onto the standard brain template and averaged across subjects. Cortical distribution of TSI_C and TSI_L values is consistent across [continued on next page]

[continued] subjects. **c.** TSI_C (**left**) and TSI_L (**right**) values in cortical areas (mean±sem across five subjects). Significant mean values are denoted by green bars and gray bars denote non-significant mean values (bootstrap test, $p > 0.05$). TSI_C is significantly greater than zero in BA44/45, SFG, and ACC. Whereas, TSI_L is significantly greater than zero in IPS and AG, and is significantly less than zero in dPMC. Both TSI_C and TSI_L are significantly greater than zero in pSTS, pMTG, SMG, and MFG. Data sets are available at <http://icon.bilkent.edu.tr/brainviewer/shahdlooetal2020>

4.3.5.3 Areas where both TSI_C and TSI_L are significant

TSI_C and TSI_L are significantly greater than zero in pSTS, pMTG, supramarginal gyrus (SMG), and middle frontal gyrus (MFG; $p < 0.05$). Posterior STS, pMTG, and SMG are considered as AON nodes that represent actions regardless of their semantic identity [124, 27, 102]. Thus, the observed positive tuning shifts in in these areas could be accounted to their generic selectivity for actions. Further to this, the significant tuning shift in MFG reported here can be associated with accounts suggesting that dorsolateral prefrontal cortex, as a node across dorsal attention network, enhances visual search by maintaining the representation of targets [33, 160, 128, 152]. Overall, these results suggest that in voxels with generic action selectivity and voxels in high-level cortical areas, visual search for action categories is facilitated by tuning shifts toward targets irrespective of target identity.

4.4 Discussion

We used brain activity evoked by natural movies to investigate how visual search for action categories modulates semantic representation of a large and diverse set of observed actions across cortex. Our results show that category-based attention during natural vision causes widespread changes in semantic representation of actions. Furthermore, these attentional modulations are not restricted to the action observation network, but are rather extended to higher-level cognitive processing areas. These results expand on prior work demonstrating attentional modulations in representation of object categories [155, 37, 73, 50]. Our results

suggest that attention facilitates action perception by modulating semantic tuning for action categories in favor of the targets. Our results also suggest an interaction between tuning shifts and intrinsic selectivity of neural populations for action categories; tuning shift toward a given category during search for a target is most prominent in cortical areas that are selective for that target.

Here, we derived an embedding space that encodes semantic variability among action categories. The first, and most important, dimension in this semantic space distinctively represents self-actions (e.g. chew, lean, yawn; i.e. intransitive actions) versus actions that involve distal objects or people (e.g. hit, talk, ride; i.e. transitive actions). Several recent neuroimaging studies report that observing actions that involve distal objects versus actions without objects leads to spatially distinctive representations [70, 218, 193]. Specifically, recent meta-analyses [27, 66] report that observing actions that involve distal objects leads to increased activation, compared to observing actions without objects, across parietal and prefrontal AON nodes [23, 141], areas that have been suggested to encode information pertaining to action goals [69, 64, 162]. Our results here can thus be taken to suggest that transitivity is a central feature in semantic representation of action categories in the brain, likely because it helps to disambiguate action goals.

Several previous studies have reported response modulations as a result of attention to high-level action features in parietal and prefrontal cortices, but not in occipitotemporal areas [142, 137, 138]. In contrast, here we observe attentional tuning shifts in occipitotemporal cortex, albeit to a lower degree compared to higher cognitive areas. The rich natural movie stimulus used here contains a large set of action categories that are performed in their natural context, by various actors, and are viewed at different angles, in contrast to most previous controlled experiments where a handful of actions are observed from a fixed angle on a homogeneous background. Moreover, the performed actions in natural movies usually involve faces or body-parts that evoke responses across occipitotemporal areas selective for these visual entities. Thus, the observed attentional modulations in occipitotemporal cortex can be accounted to semantic richness of our natural stimuli, which likely leads to attentional modulations at early levels of

action perception hierarchy.

We find significant tuning shifts toward targets in AG and SMG, higher than that in occipitotemporal and premotor AON nodes, irrespective of the search target. Previous studies have emphasized the role of these areas in amodal semantic representation of actions; while observing visual actions, hearing action sounds, or reading action words [156, 122, 202, 8]. Furthermore, there is evidence suggesting that during semantic processing, these areas act as central connectivity hubs, linking low-level perceptual processing areas to higher-level cognitive areas in prefrontal cortex [84, 51]. Complementing these lines of evidence, our results might be taken to suggest a distinctive account for AG/SMG during action observation among other AON nodes. According to this account, semantic tuning shifts toward target categories in these areas serves to enhance task-relevant semantic representations in higher cognitive areas

Recent studies suggest that during visual search, cortical areas that are selective for a given object category retain their tuning for the preferred category even when a non-preferred category is the search target [163, 164, 179]. Here, by investigating tuning shifts during search for individual targets we find that voxels in superior parietal cortex –which is suggested to be selective for locomotion actions– shift their tuning toward the target only during search for locomotion. While, voxels in communication-selective anterior prefrontal cortex shift their tuning toward the target only during search for communication. This suggests that semantic tuning shifts during search for action categories interact with the intrinsic selectivity for target categories. According to this account, semantic tuning in neural populations that are intrinsically selective for specific action categories is robust to selective attention to non-preferred actions.

The natural movie stimuli used here have greater ecological validity compared to simplified controlled movie clips used in most previous action-perception studies. However, action categories might be correlated with stimulus features at different levels and these correlations can contribute to population responses to actions. Thus, low-level features such as global motion-energy [214, 143] and

intermediate-level features such as scene dynamics [67] can confound the estimated action category responses, which can then lead to biased assessments of tuning shifts. We employed several procedures to control for these confounds. First, to minimize correlations between estimated action category responses and global motion-energy of the movie clips, we used a motion-energy regressor in our modeling procedure [143, 119]. Second, we restricted analyses voxels to the voxels in which the category model explained unique response variance after accounting for low-level motion-energy features and intermediate-level STIP features. However, we do not rule out the possibility that there might be residual influences due to other high-level action features such as expected action goals [89, 90], and actors’ perceived attitude [4]. Further work is needed to functionally dissociate potential contributions of these high-level features and attentional modulations in action representation.

In conclusion, we found that natural visual search for a given action category modulates semantic representations in favor of the target, within and beyond the action observation network. Moreover, these modulations interact with intrinsic selectivity of neural populations for search targets. This dynamic mechanism serves to facilitate action perception by efficiently allocating neural resources to accentuate the representation of task-relevant action categories. Overall, these results help explain humans’ astounding ability to perceive others’ actions in dynamic, cluttered natural visual scenes.

4.5 Publications

This chapter of the thesis have been partially presented and published in the following conferences and journals:

- Mohammad Shahdloo, Burcu A Urgan, and Tolga Çukur. Semantic Representation of Actions is Modulated by Category-Based Natural Visual Search. *in prep.*

- Mohammad Shahdloo, Burcu A Urgan, and Tolga Çukur. Attention to Action Categories Shifts Semantic Tuning Toward Targets Across the Brain. In *Organization for Human Brain Mapping (OHBM)*, page T661, Rome, 2019.

Chapter 5

Temporal Receptive Windows in the Brain Mapped via Deep Language Models

Summary

Natural language unfolds over the course of time, and to perceive it the brain has to integrate past perceptual information. There is evidence suggesting that virtually all neural populations maintain and process the past perceptual information. Moreover, extent of the information integration, formalized as the temporal receptive window (TRW), progressively increases along the language processing hierarchy. Yet, it is unknown whether and how auditory search for object categories modulates TRWs along this hierarchy. To investigate this process, we analyzed the data from an experiment where five human subjects were asked to listen to over two hours of natural stories while undergoing functional magnetic resonance imaging (fMRI), during passive listening and while performing two tasks: attending to “humans” or attending to “places”. We used multiple long short-term memory networks sensitive to different extents of past information integration to map TRWs during passive listening across the cortex. We then investigated attentional modulation of TRWs by comparing TRWs during passive listening and during the two attention tasks. We find that attention expands TRWs across cortex, and this attentional modulation become progressively stronger toward

later stages of language processing hierarchy. Furthermore, the modulations in areas that selectively process targets are modulated in favor of the targets. These results propose a dynamic attentional mechanism that modulates TRWs across the brain to facilitate representation of targets, by integrating the task relevant contextual past information.

5.1 Introduction

Spoken language perception relies on processing a continuous stream of auditory stimuli, hence processing the past information critically influences perception of the present sensory inputs. Classical view holds that recent past sensory information is maintained in neural populations across the working memory network [6, 52, 53], segregated from sensory cortices [5, 158, 190]. Yet, a more modern view that better accommodates recent evidence, namely the process memory framework, suggests that sensory cortices maintain and process past perceptual inputs to represent the current stimulus [74, 60, 78, 154]. It is well established that in sensory cortical areas that represent visual inputs, visual information get pooled across spatial receptive fields (SRFs [88]). Analogously, past perceptual inputs get pooled across temporal receptive windows (TRWs), to affect the cortical response to the current stimulus [75, 9, 83].

Visual spatial receptive fields progressively expand along a hierarchy [131, 22, 157, 169, 213]; early visual areas are selective for localized visual features, covered by small SRFs, whereas higher-level ventral-temporal areas are selective for broader spatial extents of visual stimulus covered by larger SRFs, likely to accommodate contextual information pertaining to the sensory inputs [30, 43, 172, 15, 209]. This postulates that an analogous hierarchy organizes TRWs across the language processing stream, as supported by several recent neuroimaging and electrocorticography (ECoG) studies [75, 118, 212, 86, 30]. Lerner et al. examined temporal receptive windows across the auditory cortex by measuring inter-subject correlation (ISC [76, 77]) of functional MRI responses to a story

scrambled at the level of words, sentences, and paragraphs. Variability of ISC as a function of increasing levels of stimulus scrambling was taken as a latent measure of TRW. ISC would be insensitive to the scrambling level in voxels with short TRW, whereas ISC in voxels with long TRW would decline when the stimulus was scrambled at the sentence and paragraph levels. They reported relatively short TRWs in early auditory cortex, whereas higher-level auditory and cognitive areas such as angular gyrus (AG), temporoparietal junction (TPJ), and medial prefrontal cortex exhibited long TRWs. So as to confirm, ECoG measurements while an audio-visual movie scrambled at different time scales was presented revealed similar hierarchy of TRWs, with short TRWs in early auditory areas and long TRWs in lateral prefrontal cortex [86].

There is further functional imaging evidence suggesting that SRFs are not static but are rather influenced by top-down attentional modulations. Specifically, recent evidence suggest that spatial attention expands SRFs to accommodate larger spatial extents surrounding target’s location [105, 168], and these attentional influences become more prominent toward later stages of visual processing stream [17]. Thus, it is likely that selective attention can also modulate TRWs during natural listening. As evidence that support this possibility, recent studies reported top-down influences on population responses during auditory language perception as a result of selective attention to perceptual [144, 149, 194, 185] and lexical [182, 211, 173, 129] features. Critically, in a recent study Regev et al. investigated top-down effects of selective attention in representation of multi-modal linguistic inputs in cortical areas with short and long TRWs. They simultaneously presented a written and a spoken story and subjects were asked to attend to one story and ignore the other, or to do passive reading and passive listening in runs where only one of the modalities was presented. As a result of attending to the stories, they reported enhanced ISC in parietal and prefrontal cortices, suggesting that attention enhances task-relevant activity in areas that have been previously suggested to exhibit long TRWs [118]. Yet, it is not specifically known whether and how selective attention modulates TRWs across different levels of the language processing hierarchy.

Here we sought to investigate attentional modulations of single-voxel TRWs

during category-based auditory attention. To this end, we analyzed the data from an experiment where five human subjects were asked to listen to over two hours of stories from *The Moth Radio Hour* [118] while performing passive listening, attending to “humans”, or attending to “places” in different runs. Whole-brain blood-oxygen-level-dependent (BOLD) responses were recorded using functional MRI (fMRI). Data were collected at the University of California, Berkeley. We used a rich contextual representation of the stories derived from a long short-term memory (LSTM) neural network to fit voxelwise language models for each task and in each individual subject. Then, we systematically estimated TRW in single voxels by fitting separate language models using features from separate LSTM networks trained to have a range of memory extents. We compared TRWs between the passive listening task and the two attention tasks to estimate the sensitivity of TRWs to attentional modulations. Furthermore, we compared TRWs between the two attention tasks to assess the interaction of TRW length and intrinsic semantic selectivity across neocortex. Our findings suggest that attention modulates TRWs in many voxels across the language processing network. Moreover, the amount and direction of attentional modulations depends on intrinsic semantic selectivity for targets.

5.2 Methods

5.2.1 Subjects

Five healthy male adult volunteers with normal hearing participated in this study: S1 (age 31), S2 (age 27), S3 (age 32), S4 (age 33), S5 (age 27). Data were collected at the University of California, Berkeley. Experiment protocols were approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley. All participants gave written informed consent before scanning.

5.2.2 fMRI data collection

Data were collected using a 3T Siemens Tim Trio MRI scanner (Siemens Medical Solutions) using a 32-channel receiver coil. Functional data were collected using a T2*-weighted gradient-echo echo-planar-imaging pulse sequence. The following sequence parameters were prescribed: TR = 2 sec, TE = 33 msec, water-excitation pulse with flip angle = 70°, voxel size = 2.24 mm×2.24 mm×4.13 mm, field of view = 224 mm×224 mm, 32 axial slices. Anatomical scans were performed using a three-dimensional T1-weighted magnetization-prepared rapid-acquisition gradient-echo (MPRAGE) sequence with the following parameters: TR = 2.3 sec, TE = 3.45 msec, flip angle = 10°, voxel size = 1 mm×1 mm×1 mm, field of view = 256 mm×212 mm×256 mm. Anatomical data were used to automatically construct flattened cortical surfaces via Freesurfer [41, 167]. PyCortex was used for visualization of the flattened surfaces [59].

5.2.3 Stimuli and experimental design

Natural stories were used as the stimulus in the experiment. Ten stories were taken from *The Moth Radio Hour* [118]. Each 10-15 min autobiographical story was narrated by a single male or female speaker before a live audience. The stories covered a wide range of topics and were semantically rich. Each story was manually transcribed and time-aligned (as detailed in 92) to yield the onset time of each of the 11,220 story words. The stimulus was played on Sensimetrics S14 in-ear piezoelectric headphones that were calibrated to yield a flat frequency response. The audio stimulus was normalized to have peak loudness of -1 dB relative to maximum and was played at 44,1 kHz. Each fMRI run consisted of a single story, complemented by 10 sec of silence before the story and 10 sec of silence after the story.

Two separate sets of functional data were collected. A passive-listening dataset was collected during two experimental sessions, each consisting of five runs. Subjects passively listened to the ten stimulus stories in separate runs. This resulted

in presentation of 124 min 34 sec of stories without repetition, yielding 3,737 data samples. Data for the attention tasks were collected during two experimental sessions, each consisting of six runs. The six stories used for the attention tasks were common with the stories in passive-listening task. Subjects were instructed to attend to “humans” or to “places” in the stories. Exemplars of targets were provided to the subjects before the experiment. The order of attention task was interleaved across runs to minimize subject expectation bias. This resulted in presentation of 75 min 24 sec of stories without repetition in each attention task, yielding 2,262 data samples.

5.2.4 fMRI data preprocessing

Statistical Parameter Mapping toolbox (SPM12 Friston et al. 56) was used to perform motion correction. Functional volumes were aligned to the first image from the first run in each subject. Non-brain tissues were identified and removed using the brain extraction tool (BET) from the FSL software package [187]. Low-frequency response components were filtered out of the voxel responses using a third order Savitzky-Golay low-pass filter with 240 sec temporal window. Finally, voxel responses were normalized to attain zero mean and unit variance. Voxels within the 2 mm neighborhood of the cortical sheet were identified as cortical voxels in each subject and were used for analyses (S1, 34,269 voxels; S2, 35,979 voxels; S3, 37,226 voxels; S4, 42,090 voxels; S5, 36,942 voxels).

5.2.5 Definition of regions of interest

To define the anatomical regions of interest (ROIs), cortical surfaces were segmented into 156 regions of the Destrieux atlas [46] via Freesurfer. Segmentation results were projected from anatomical space onto the functional space using Py-Cortex and each voxel was given one anatomical label based on the projections. Finally, ROIs were refined to voxels near a 2 mm neighborhood of the cortical sheet.

5.2.6 Semantic embedding of stories

To investigate the semantic information pertaining to the stories, a word co-occurrence matrix was constructed [92]. First, a large corpus of English text was collected by scraping 9,412,972 user comments from <http://reddit.com>, that comprised of 21,542,312 words. The set of 985 common English words taken from *Wikipedia's List of 1000 Basic Words* was used as the set of basis words. Then, log-transformed co-occurrence frequency between each of the 11,220 story words and the 985 basis words within a 15-word window were assessed by iterating through the text corpus, yielding a 985 by 11,220 co-occurrence matrix. To account for differences in story word frequencies, each 985-dimensional co-occurrence vector was z-scored. Moreover, to account for differences in basis word frequencies, each row of the co-occurrence matrix was z-scored.

5.2.7 Language model features

Because the co-occurrence statistics of words are calculated independently, the constructed co-occurrence matrix does not account for the temporal correlations among story words. To address this issue, we used a long short-term memory (LSMT) deep neural network as the language model to incorporate temporal information into the semantic embedding [98]. The network comprised of 3 layers of 985 LSTM units. The network was trained on the text corpus to predict the 985-dimensional co-occurrence vector of each training sample using co-occurrence vectors of 20 preceding words. This resulted in 985 by 20 dimensional training samples, namely the context matrices. Features from the output layer of the network were used to assess a 985-dimensional feature vector for each stimulus word. Language model features were then obtained by downsampling the time-course of feature vectors using a Lanczos filter to match the acquisition rate of fMRI.

5.2.8 Model estimation and validation

Regularized linear regression (i.e. ridge regression) was used to fit models in each voxel that mapped language model features to the measured BOLD responses in individual subjects in each task (i.e. passive viewing, attending to humans, attending to places). Finite impulse response (FIR) predictors at lags of 2, 4, 6, and 8 secs were used to account for the hemodynamic response. A nested cross-validation (CV) procedure was used to choose the regularization parameters and to estimate model weights. Data collected during the attention tasks were segmented into 56 80-second blocks. In each of the 10 outer folds, 5 randomly chosen blocks were held-out as validation data. Then, in each of the 10 inner folds, 48 randomly chosen blocks were used as training data and the 3 remaining blocks were used as test data. To fit models for the passive-listening data, data were segmented into 93 80-second blocks. In each fold, 10 randomly chosen blocks were held-out as validation data, 75 randomly chosen blocks were used as training data and the 8 remaining blocks were used as test data. We used 10 regularization parameters in the range $\lambda_f \in [2^5, 2^{17}]$. In each task, training data were used to fit models for each λ_f independently. Model weights were then used to predict responses in the test data and prediction scores of the fit models were assessed, taken as Pearson’s correlation coefficient between actual and predicted voxel responses. The value of λ_f maximizing the average prediction score across inner CV folds was chosen in each voxel. Finally, the optimized parameters were used to fit models on the union of training and test data in each outer fold, and model weights were averaged across the outer folds.

To validate the fit models, we evaluated their performance in predicting held-out responses, separately in each task. In each outer fold, responses were predicted for the validation data using the fit models and prediction score in each voxel was calculated. Prediction scores were averaged across outer folds. Statistical significance of the fit models during passive listening task was assessed by comparing the prediction scores to 10,000 bootstrap samples from a null distribution of prediction scores. A null hypothesis was considered where prediction

score is not greater than the correlation between two independent Gaussian random vectors of the same length as the predicted BOLD responses. The p -value was taken as the fraction of the bootstrap samples for which the prediction score was lower than the correlation of random vectors. Finally, statistical significance levels were corrected for multiple comparisons using false discovery rate (FDR) control [10].

5.2.9 Physiological noise controls

To prevent head motion and physiological noise confounds, these nuisance factors were estimated and regressed out of the BOLD responses. Six head-motion regressors were assessed by estimating affine motion time courses during the motion-correction stage. During experimental runs, cardiac and respiratory activity were recorded using a pulse oximeter and a pneumatic belt. These data were then used to estimate two regressors to capture respiration and nine regressors to capture cardiac activity [208].

5.2.10 Temporal receptive windows

The estimated language features were obtained from the LSTM language network trained on context matrices, where the 985 by 20 dimensional context matrix for a given word represents contextual information pertaining to 20 preceding words. To manipulate the temporal extent of contextual information used to obtain the language features we trained separate LSTM networks using randomized training samples, where randomization at level ℓ yields training samples that reflect semantic information pertaining to ℓ previous words. This was achieved by replacing the ℓ^{th} to 20^{th} left-most columns in each training sample with co-occurrence vectors corresponding to random words from the text corpus. This is equivalent to replacing the $\ell - 20$ trailing context words with random words from the corpus. Separate LSTM networks were trained using $\ell \in [1, 20]$ separate randomization levels. This yielded 20 separate sets of language features. Each set of

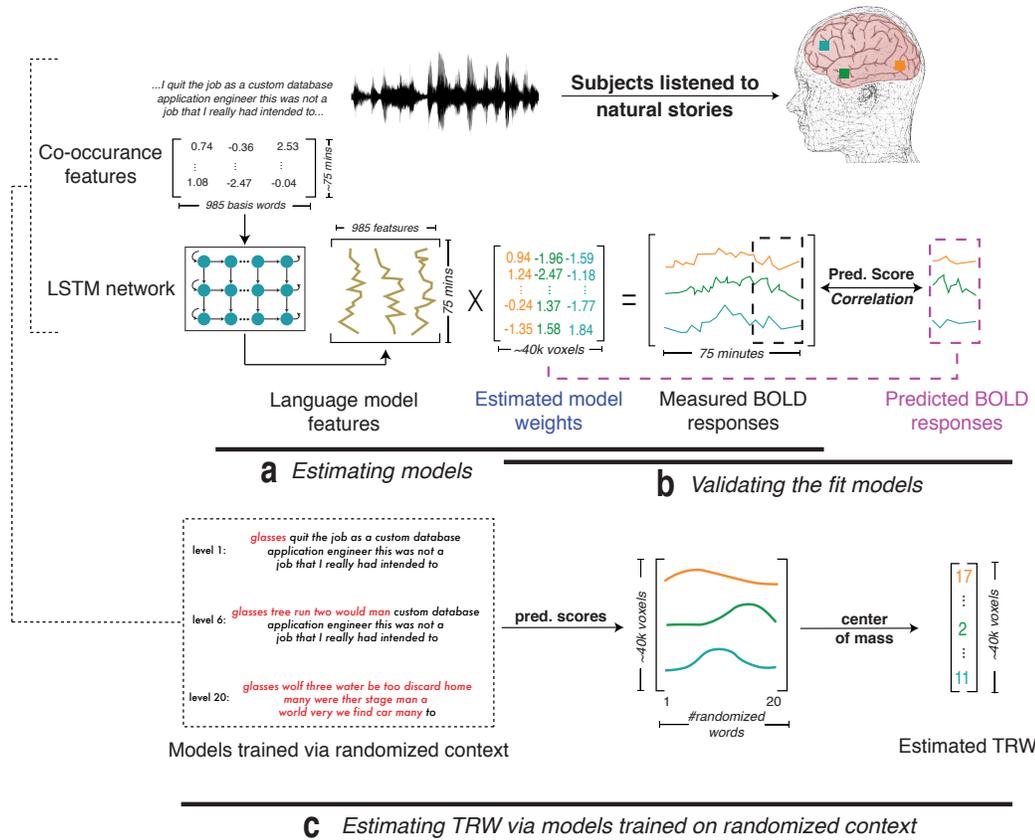


Figure 5.1. Model fitting, validation, and estimation of TRWs. While undergoing fMRI, five human subjects listened to over 2 hours of natural stories during passive listening, and during attention to “humans” or “places”. **a.** Co-occurrence statistics was used to obtain a semantic representation of the stories. To incorporate lingual temporal information into the semantic representation, a long short-time memory (LSTM) language model was used, yielding the time course of 985 language features. The LSTM network was trained via a large English text corpus, to predict co-occurrence vector of a sample word using co-occurrence vectors of 20 preceding words. Language model weights were estimated that mapped each language feature to the BOLD responses recorded in each task. **b.** Accuracy of the fit models were assessed by measuring their performance in predicting held-out BOLD responses. Prediction score of the fit models was taken as Pearson’s correlation coefficient between estimated and measured responses. **c.** To emulate variability of integrated information in the language features, separate LSTM networks were trained using contextual samples that were randomized at various levels, yielding 20 separate sets of language features. Voxelwise models were fit that mapped each set of language features to the recorded responses, and prediction scores of the models were calculated, yielding 20-dimensional prediction score profiles. TRW in each voxel was taken as center of mass of the prediction score profiles.

language features were used to fit language models in single voxels during each task (i.e. passive viewing, attending to humans, attending to places) and in individual subjects, and prediction scores of the models were assessed. This yielded a 20-dimensional vector of prediction scores for each voxel in each task. Finally, the randomization level yielding the maximum prediction score was obtained by calculating the center-of-mass of the prediction score vectors, and was taken as the TRW in each voxel during each task in individual subjects.

5.2.11 Alternative approach to measure TRW

We implemented an alternative approach to incorporate different extents of contextual information integration into the fit models. To this end, instead of randomizing the tailing context words used to train LSTM networks we changed the size of the context matrices; samples with context size ℓ are constructed by removing the feature vectors corresponding to $\ell - 20$ tailing columns from the context matrix. The remaining steps in calculating TRWs are the same as the randomization-based approach.

5.2.12 Sensitivity and bias of TRWs

In this study, we sought to investigate whether and how attention modulates TRWs across cortex. To test sensitivity of TRWs to attentional modulation we compared TRW between the passive listening task and the two attention tasks. We calculated an attentional modulation index (AMI) in single voxels as

$$AMI = \frac{1}{2}(|TRW_0 - TRW_H| + |TRW_0 - TRW_P|) \quad (5.1)$$

where TRW_0 , TRW_H , and TRW_P are the TRW during passive listening, attention to humans, and attention to places. In a voxel with AMI of 0 attention does not modulate TRW, whereas TRW in a voxel with AMI of 1 gets maximally modulated by category-based attention. Finally, we quantified an attentional bias

index (ABI) as

$$ABI = TRW_H - TRW_P \quad (5.2)$$

Maximized TRW during attention to humans versus attention to places yields positive versus negative ABI in the range $[-1, 1]$.

5.3 Results

It is currently unknown whether and how search for specific object categories during natural listening modulates temporal receptive windows (TRWs) across the language processing hierarchy (Fig. 5.2). To address this question, we analyzed the data from an experiment where five human subjects were asked to listen to over two hours of narrative stories while undergoing functional magnetic resonance imaging (fMRI), and perform three separate tasks: passive listening, attending to “humans”, or attending to “places”. We used a long short-term memory (LSTM) network to obtain a latent representation of the lingual information pertaining to the presented stories and modeled brain responses evoked by natural stories. This network integrated past semantic information over a 20-word window to output timecourse of 985 language features. Language models were fit in single voxels that mapped the obtained language features to the recorded responses in individual subjects and for each task (Fig. 5.1a,b).

5.3.1 Language model well predicts responses to narrated stories

The fit language models would be functionally important only if they can successfully predict responses to stimuli not used for model fitting. To address this issue, we used the language models estimated during passive listening to predict held-out responses. Prediction score of the fit models were taken as Pearson’s correlation coefficient between single voxel predicted responses and the measured

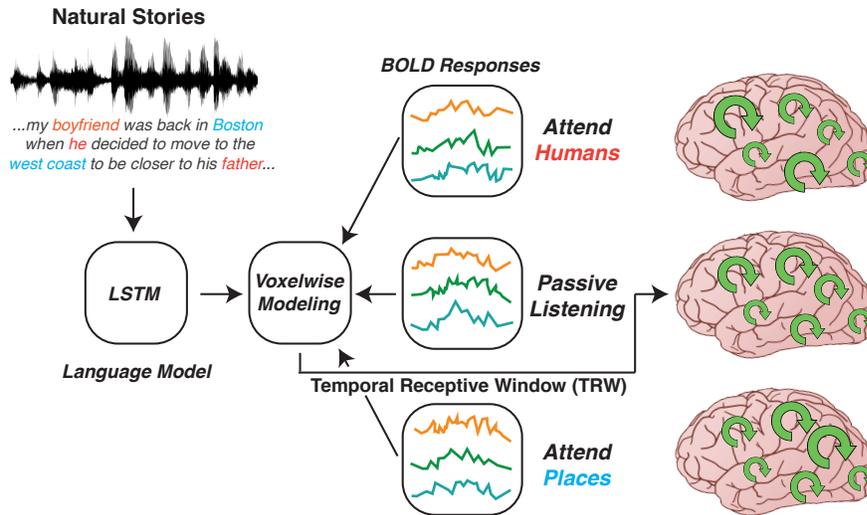


Figure 5.2. Hypothesized changes in process memory timescales. Previous studies suggest that virtually all neural populations maintain and process past perceptual information to represent current inputs. Moreover, timescale of this process memory mechanism (formalized as the temporal receptive window (TRW)) is hierarchically organized across cortex. During natural auditory listening, a language model can be employed to estimate TRWs via a model-based approach. There is evidence suggesting that selective attention modulates language representation in the brain. Thus, auditory object-based search should modulate TRWs.

responses (see *Methods*). We then assessed statistical significance of the calculated prediction scores. A null hypothesis was considered where prediction scores are not different than zero. Under this null hypothesis, prediction scores are not different than the correlation between two Gaussian random vectors. We find that the language models significantly predict responses in $11.0 \pm 4.2\%$ of cortical voxels (bootstrap test, $q(\text{FDR}) < 0.05$). These significantly predicted voxels were used in all subsequent analyses. Moreover, the language models have high prediction scores (greater than 1 std above the mean) in $27.7 \pm 6.4\%$ of the significantly predicted voxels. Many voxels across primary and non-primary auditory cortex (AC), parietal cortex, and inferior frontal cortex are among these well-predicted voxels. These results suggest that many voxels across auditory-processing cortex, and higher-level parietal and prefrontal cortices encode lingual information in natural stories, that can be significantly characterized via the LSTM language model.

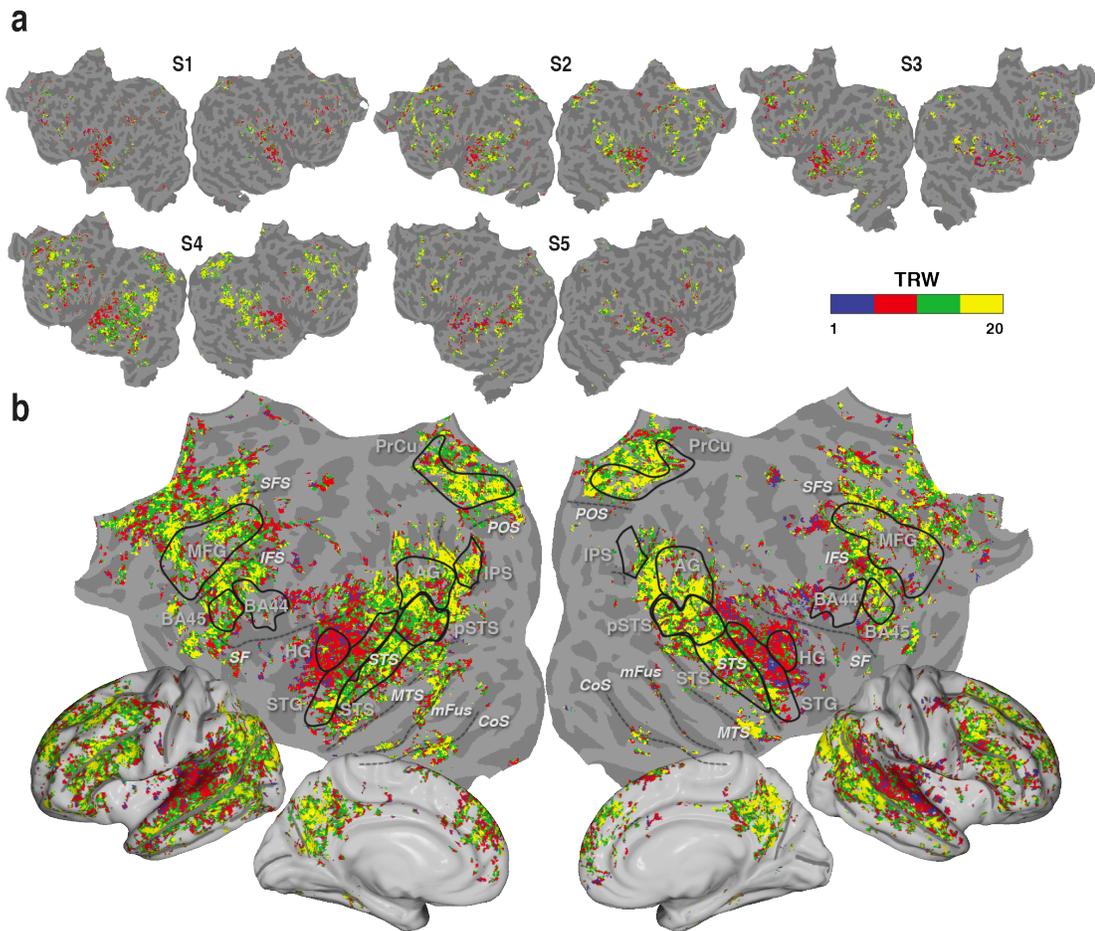


Figure 5.3. TRW values estimated during passive listening. **a.** TRW values estimated during passive listening were plotted on the flattened cortical surfaces of the five subjects. Voxels where prediction scores of the language model during passive listening were not significant are hidden, revealing the cortical curvature below. **b.** TRW values were projected onto the standard brain template and averaged across subjects. Anatomical regions of interest and important sulci are identified on the flat map; HG, Heschl's gyrus; pSTS, posterior superior temporal sulcus; STG, superior temporal gyrus; AG, angular gyrus; IPS, intraparietal sulcus; PrCu, precuneous; BA44/BA45, Brodmann area 44/45; IFS, inferior frontal sulcus; SFS, superior frontal sulcus; SF, Sylvian fissure; STS, superior temporal sulcus; MTS, middle temporal sulcus; mFus, mid-fusiform sulcus; CoS, collateral sulcus; POS, parieto-occipital sulcus. TRWs are short in primary auditory cortex (HG). Voxels in non-primary auditory cortex (STG, STS) exhibit low to intermediate TRWs. Many voxels across parietal and prefrontal cortices exhibit long TRWs.

5.3.2 Distribution of TRWs across cortex

To emulate the sensitivity of language models to variations in the information integration extent, separate networks were trained that used different extents of past semantic information, yielding 20 separate sets of language features. In each voxel, language models were fit using each of the 20 feature sets, and prediction scores of the fit models were calculated, yielding 20-dimensional voxelwise prediction score profiles. The prediction score profile of a given voxel would indicate selectivity of that voxel for different extents of semantic information integration. The TRW width in single voxels was then taken as the center of mass of prediction score profiles (Fig. 5.1c), yielding $\text{TRW} \in [1, 20]$. Under this scheme, a voxel with short TRW would be selective for short extent of past semantic information, whereas a voxel with wide TRW would favor integrating information over a longer extent. We visualized the distribution of TRWs by projecting them onto flattened cortical surfaces (Fig. 5.3). Many voxels in primary AC (near Heschl’s gyrus (HG)) exhibited short TRWs (Fig. 5.3, blue and red). Moving rostrally and caudally from HG to non-primary AC voxels preferred increasingly wider TRWs, where superior temporal gyrus (STG) voxels exhibited intermediate TRWs (Fig. 5.3, red and green). Voxels in superior temporal sulcus (STS), temporo-parietal junction (TPJ), precuneus (PrCu), and prefrontal cortex exhibited the widest TRWs (Fig. 5.3, yellow). The distribution of TRWs found here very well replicates the results from previous studies that estimated TRWs by temporally modifying the stimulus [75, 118]. Figure 5.7a shows TRW widths in several common regions of interest (ROIs). Previous studies also report progressively increasing TRW width from primary AC toward non-primary AC and inferior parietal cortex. Consistent with these reports, we find that TRW widths significantly increase from HG to STG, and from STG to STS (bootstrap test, $p < 0.01$). We also find that TRWs are significantly wider in prefrontal cortex (BA45, IFS, MFG), and in parietal cortex (AG, IPS, PrCu) compared to primary AC ($p < 0.01$). These results can be taken to suggest that representation of narrated stories in primary AC does not rely on the recent history of semantic inputs, whereas higher-level cortical areas integrate semantic information to represent the stories.

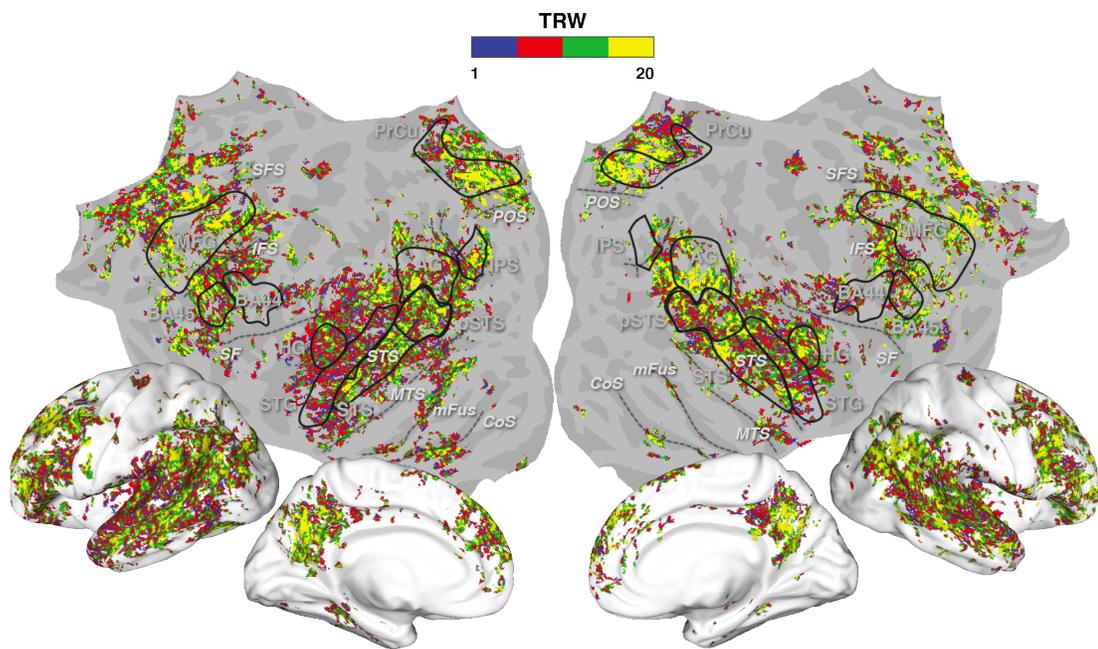


Figure 5.4. TRW values estimated via the alternative approach. TRW values were assessed via an alternative approach where—instead of randomizing the tailing context words—the tailing context words were removed from the training samples. TRW values were projected onto the standard brain template and averaged across subjects. The map of TRW distribution assessed via the alternative method is similar to the original map of TRW distribution shown in Fig. 5.3.

We compared the TRWs assessed via an alternative method to validate the distribution of TRWs. In this alternative method, instead of randomizing the tailing context words, we removed the tailing context words from training samples. The distribution of TRWs assessed via this alternative approach is shown in Fig. 5.4. This alternative assessment of TRWs is similar—but noisier—compared to the TRWs assessed via the randomization approach (Fig. 5.3 versus Fig. 5.4).

5.3.3 Attention modulates TRWs across cortex

Recent reports suggest that attentional modulations vary across areas along the auditory and language processing hierarchies [99, 25, 227, 151]. This raises the possibility that TRWs modulations by category-based auditory search might also vary across the auditory and lingual processing hierarchies. To test this possibility, we compared TRWs between passive listening task and the two attention tasks. We quantified attentional modulations of TRWs by defining an attentional modulation index ($AMI \in [0, 1]$). TRW width in a voxel with AMI of 1 gets maximally modulated by attention, while in a voxel with AMI of zero TRW does not get modulated by attention. To visualize the distribution of attentional modulations of TRW across cortex, we visualized AMI values onto flattened cortical surfaces (Fig. 5.5). We observe that TRWs are modulated in many voxels across non-primary AC (STG, STS), temporal pole, parietal cortex (AG, IPS, PrCu), and anterior inferior frontal cortex (BA45). This observation is aligned with previous findings suggesting that selective attention modulates representation of natural stories in parietal, and prefrontal cortices [165]. Moreover, we studied AMI in several common ROIs (Fig. 5.7b). AMI is significantly higher in non-primary AC (STS, STG; bootstrap test, $p < 0.01$) and in prefrontal cortex (BA45, MFG, IFS; $p < 0.05$) compared to primary AC. These results suggest that TRW in voxels across non-primary auditory cortex and high-level language areas get expanded during attention to auditory object categories.

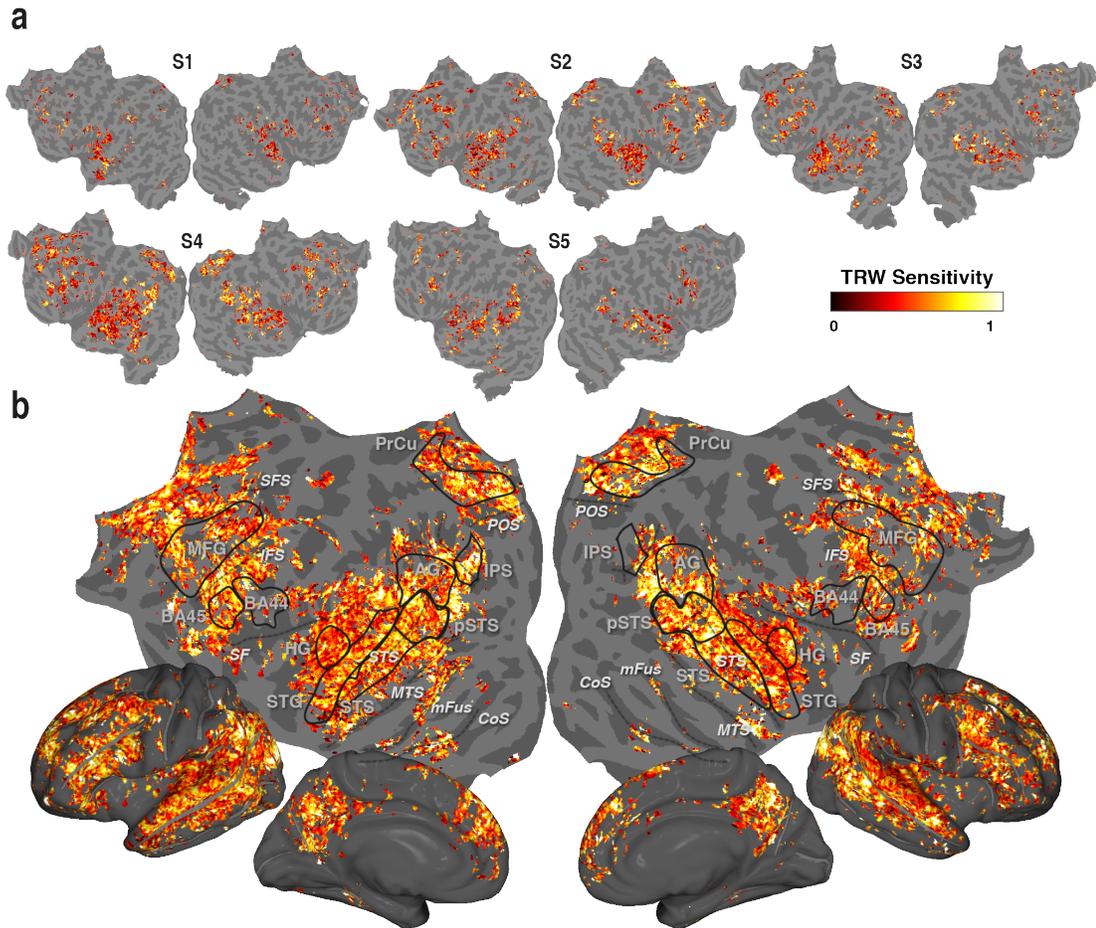


Figure 5.5. Attentional modulation of TRWs. **a.** For each voxel, an attentional modulation index ($AMI \in [0, 1]$) was calculated to quantify the attentional shrinkage or expansion of TRWs. Attention maximally modulates TRW in a voxel with an AMI of 1, whereas an AMI of zero indicates that voxel TRW is not modulated by attention. AMI values plotted on the flattened cortical surfaces of the five subjects. Voxels where prediction scores of the language model during passive listening were not significant are hidden, revealing the cortical curvature below. Cortical distribution of AMI values is consistent across subjects. **b.** AMI values from individual subjects were projected onto the standard brain template and averaged across subjects. TRW in many voxels across non-primary auditory cortex (AC), parietal cortex, and prefrontal cortex are modulated by attention.

5.3.4 Auditory search for object categories selectively modulates TRWs

Previous evidence suggest that attending to a given object category (e.g. birds, tools) in an auditory scene enhances BOLD responses in cortical areas that are selective for that target [116]. This postulates that TRW modulations during category-based auditory search should be influenced by selectivity of neural populations for targets. To test this hypothesis, we compared TRWs between the “attend to humans” and “attend to places” tasks. To quantify the TRW modulations across these task, an attentional bias index ($ABI \in [-1, 1]$) was formalized. TRW in a voxel with ABI of 1 is maximized during search for humans, while an ABI of -1 indicates maximized TRW during search for places, and an ABI of zero indicates that voxel TRW is not modulated across tasks. We visualized this attentional bias by projecting ABI values onto cortical surfaces (Fig. 5.6). TRW bias was most prominent in posterior temporal, parietal, and inferior prefrontal cortices. Figure 5.7c shows ABI in several common ROIs. ABI is not significant (bootstrap test $p > 0.05$) in AC (HG, STG, STS), in angular gyrus, and in most areas in the prefrontal cortex (BA44/45, MFG). Yet, ABI is greater than zero in pSTS and PrCu. This suggests that during search for humans, TRWs in pSTS and PrCu get expanded. Previous studies suggest that precuneous facilitates representation of humans by encoding high-level information such as other’s feelings and intentions [145, 117]. However, perception of these features while attending to humans during natural audition intrinsically relies on the semantic context preceding target onset. Furthermore, previous reports suggest that pSTS maintains semantic information pertaining to humans during visual and auditory perception [3]. Thus, positive ABI in PrCu and pSTS can be accounted to significance of semantic context in perception of the high-level semantic features represented in these areas. We find that TRWs in IPS and IFS get expanded during search for places. Previous evidence suggest that IPS maintains spatial information that are later recalled during both visual [21] and auditory [109, 132] search for places. Moreover, inferior prefrontal cortex has been suggested to maintain representation of spatial information during natural audio-visual perception [24], and modulate place perception through its connection with the hippocampus [113].

These lines of research suggest that during search for places, TRWs in IPS and in IFS get expanded to accommodate spatial contextual information pertaining to targets.

5.4 Discussion

Using fMRI responses to over two hours of natural stories, we employed a model-based approach to estimate temporal receptive windows (TRWs) in the brain, and their modulation during auditory search for “human” or “place” categories. Our results show that TRWs are short in early auditory cortex (AC), but they get increasingly wider toward higher stages of cognitive processing, replicating the distribution of TRWs that were previously revealed by alternative methods based on stimulus scrambling. Furthermore, our results suggest that category-based auditory search variably modulates TRWs along the auditory and lingual processing hierarchies; attentional modulations are stronger in higher-level auditory and lingual processing areas.

There is neurophysiological evidence showing that visual search can expand visual spatial receptive field (SRF) of a neuron to accommodate more of the information surrounding the target [183, 58]. Our results here suggest a similar flexibility for TRWs; auditory category-based search expands TRWs to accommodate contextual semantic information pertaining to targets. Specifically, analogous to the previous findings suggesting that SRFs in later stages of visual processing exhibit stronger attentional modulations [147], here we find increasingly stronger attentional modulations for TRWs toward later stages of cognitive processing.

We find that semantic identity of the search target interacts with TRW modulations in different areas across the cortex. Auditory search for humans expands TRWs in areas that selectively represent humans and inter-personal communications [135, 226], whereas search for places expands TRWs in areas selective for places [39, 130]. These evidence support the hypothesis that selective attention facilitates auditory language processing by expanding TRWs to integrate greater

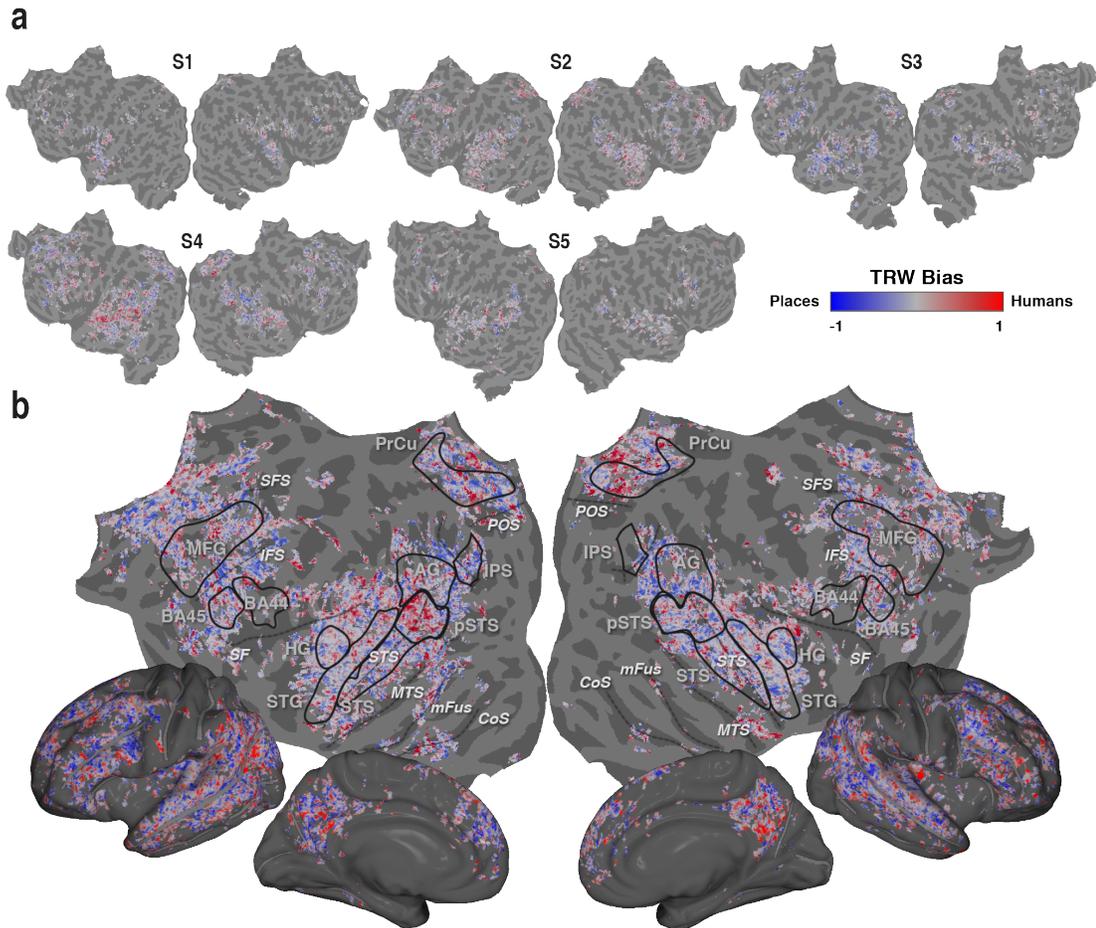


Figure 5.6. Attentional bias of TRWs. **a.** For each voxel, an attentional bias index ($ABI \in [-1, 1]$) was calculated to quantify the attentional modulation of TRWs during attention to humans versus attention to places. TRW maximally expands during attention to humans in a voxel with an ABI of 1, whereas an ABI of -1 indicates that voxel TRW is maximally expanded during attention to places. ABI values plotted on the flattened cortical surfaces of the five subjects. Voxels where prediction scores of the language model during passive listening were not significant are hidden, revealing the cortical curvature below. Cortical distribution of ABI values is consistent across subjects. **b.** ABI values from individual subjects were projected onto the standard brain template and averaged across subjects. TRWs in voxels across pSTS and PrCu expand during attention to humans. While, attention to places expands TRWs in IPS and IFS.

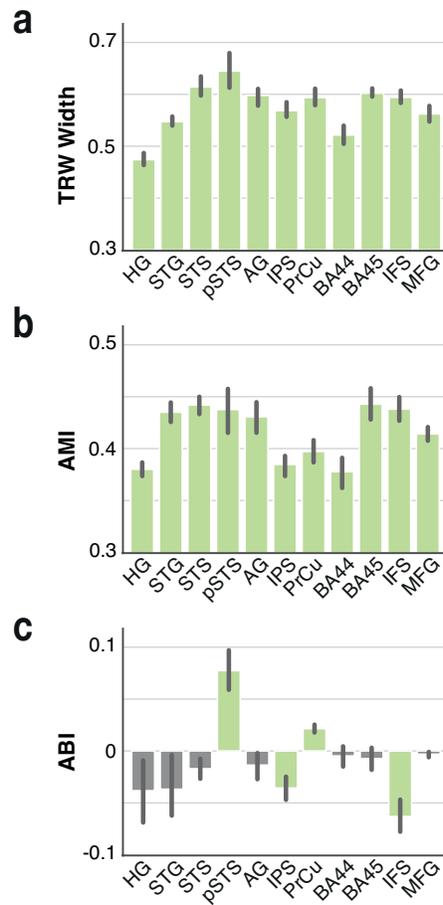


Figure 5.7. TRWs and their attentional modulations in anatomical ROIs. **a.** TRW during passive viewing, in anatomical areas (mean±sem across five subjects). TRWs are significantly wider in non-primary AC (STG, STS) compared to primary AC (HG; bootstrap test, $p < 0.01$). TRWs are also wider in parietal and prefrontal cortex compared to HG ($p < 0.05$). **b.** Attentional modulation index (AMI) in anatomical ROIs. AMI is significantly higher in non-primary AC (STS, STG; $p < 0.01$) and in prefrontal cortex (BA45, MFG, IFS; $p < 0.05$) compared to primary AC. **c.** Attentional bias index (ABI) in anatomical ROIs. TRWs are significantly wider during attending to humans in pSTS and PrCu ($p < 0.05$), whereas it is significantly. While, they are wider during attending to places in IPS and IFS ($p < 0.05$).

extents of contextual semantic information pertaining to targets. These evidence also parallel an analogous dynamic modulation mechanism for SRFs during visual search, where spatial visual search expands SRFs to accommodate greater extent of spatial context pertaining to targets [168].

Previous studies have employed temporal manipulation of the stimulus to map TRWs across the cortex [75, 118]. Here, we employed a more systematic technique to estimate TRWs, using multiple long short-term memory (LSTM) networks that each were trained to maintain a certain amount of past semantic information. This approach has several advantages compared to previous methods that manipulate the stimulus. Comparing responses recorded during multiple stimulus presentation sessions would suffer from physiological and experimental variations between sessions. Furthermore, several presentation of the same audio stimulus, although with variation of the scrambling extent, can lead to subject expectation bias. Yet, the approach used here is immune to these potential confounds since it incorporates the manipulation into the model instead of manipulating the stimulus.

The neural network used as the language model well explains variance in the responses. Yet, it has the limitation of using a fixed context length of 20 preceding words to predict the embedding vector for each input word. Recent progress in natural language processing have led to emergence of novel architectures based on Transformers [206, 47]. These architectures employ self-attention units that results in non-uniform variable-length integration of past contextual information to predict the output to the current input instance. Employing these self-attention network architectures –that are more ecologically valid compared to LSTMs– in the future can increase precision and statistical power in TRW estimations.

In summary, we systematically mapped TRWs across the cortex and found that category-based attention during natural audition modulates TRWs across auditory and lingual processing hierarchies. Moreover, these modulations interact with semantic selectivity of neural populations for search targets. This dynamic attentional mechanism enhances language representation by expanding

the amount of represented contextual information pertaining to task-relevant categories. Overall, these results help explain human’s ability to perceive complex streams of auditory lingual information in daily life.

5.5 Publications

This chapter of the thesis have been partially presented and published in the following conferences and journals:

- Mohammad Shahdloo, Mert Acar, and Tolga Çukur. Attention During Story Listening Modulates Temporal Receptive Windows Across Human Cortex. In *Conference on Cognitive Computational Neuroscience (CCN)*, page PS1A.52, Berlin, 2019.

Chapter 6

Concluding Remarks and Future Directions

The broad goal of this dissertation was to develop novel computational tools to reconstruct MRI images and analyze the rich information underlying large-scale fMRI acquisitions. With this aim, we proposed a novel self-tuning compressed sensing reconstruction for generic multi-coil multi-acquisition MRI datasets. Our proposed method outperforms conventional self-tuning reconstruction methods while being up to an order of magnitude computationally efficient. Furthermore, we analyzed fMRI data gathered during hours of natural stimuli presentation to uncover novel insights into brain’s cognitive functions during visual and auditory category-based attention. The key contributions of the thesis are:

1. Developing PESCaT, a new self-tuning reconstruction method to optimize regularization parameters in CS-MRI
2. Investigating semantic representation of actions and corresponding attentional modulations during category-based visual search
3. Developing a model-based approach to estimate temporal receptive windows and their attentional modulations during natural audition

These contributions propose several new research questions and directions. First, PESCiT reconstruction imposes fidelity to the calibration data collected from the central region of k-space. Yet, in some MRI applications, such as MR spectroscopy and dynamic imaging, acquisition of calibration data is impractical [199, 184]. An important and promising research direction would be to extend PESCiT to be compatible with calibration-less reconstruction scenarios. Second, here we investigated attentional modulation of semantic representation of actions while controlling for spurious correlations among action categories and low- and intermediate-level stimulus features. Yet, we do not rule out the possibility that there might be residual influences due to other high-level action features such as expected action goals, and actors' perceived attitude [89, 90, 4]. Further work would be needed to functionally dissociate potential contributions of these high-level features and attentional modulations in action representation. Moreover, further work using high-resolution fMRI sequences can lead to uncovering detailed dynamics of action representation and its attentional modulations. Finally, we estimated temporal receptive windows (TRWs) by training separate long short-term memory (LSTM) networks that were aware of separate extents of past contextual information. However, more modern network architectures for natural language perception have been proposed in the literature. Thus, it is an open research direction to investigate estimates of TRW using alternative networks such as self-attention networks based on Transformers.

Bibliography

- [1] Rouhollah O Abdollahi, Jan Jastorff, and Guy A Orban. Common and Segregated Processing of Observed Actions in Human SPL. *Cerebral Cortex*, 23(11):2734–2753, August 2012.
- [2] Alan Anticevic, Grega Repovs, and Deanna M Barch. Resisting emotional interference: Brain regions facilitating working memory performance during negative distraction. *Cognitive, Affective and Behavioral Neuroscience*, 10(2):159–173, June 2010.
- [3] Bashar Awwad Shiekh Hasan, Mitchell Valdes-Sosa, Joachim Gross, and Pascal Belin. "hearing faces and seeing voices": Amodal coding of person identity in the human brain. *Scientific Reports*, 6:37494, November 2016.
- [4] Patric Bach and Kimberley C Schenke. Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass*, 11(7), July 2017.
- [5] Alan Baddeley. The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423, November 2000.
- [6] Alan Baddeley. Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, 4(10):829–839, January 2003.
- [7] Lorella Battelli, Patrick Cavanagh, and Ian M Thornton. Perception of biological motion in parietal patients. *Neuropsychologia*, 41(13):1808–1816, January 2003.

- [8] Marina Bedny and Alfonso Caramazza. Perception, action, and word meanings in the human brain: The case from action verbs. *Annals of the New York Academy of Sciences*, 1224(1):81–95, January 2011.
- [9] Tristan A Bekinschtein, Stanislas Dehaene, Benjamin Rohaut, François Tadel, Laurent Cohen, and Lionel Naccache. Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences of the United States of America*, 106(5):1672–1677, February 2009.
- [10] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, August 2001.
- [11] Berkin Bilgic, Vivek K Goyal, and Elfar Adalsteinsson. Multi-contrast reconstruction with Bayesian compressed sensing. *Magnetic Resonance in Medicine*, 66(6):1601–1615, December 2011.
- [12] Berkin Bilgic, Itthi Chatnuntawech, Audrey P Fan, Kawin Setsompop, Stephen F Cauley, Lawrence L Wald, and Elfar Adalsteinsson. Fast image reconstruction with L2-regularization. *Journal of Magnetic Resonance Imaging*, 40(1):181–191, November 2013.
- [13] Berkin Bilgic, Itthi Chatnuntawech, Christian Langkammer, and Kawin Setsompop. Sparse methods for Quantitative Susceptibility Mapping. In Manos Papadakis, Vivek K Goyal, and Dimitri Van de Ville, editors, *Proceedings of SPIE - The International Society for Optical Engineering*, page 959711. Massachusetts General Hospital, Boston, United States, SPIE, January 2015.
- [14] Berkin Bilgic, Tae Hyung Kim, Congyu Liao, Mary Kate Manhard, Lawrence L Wald, Justin P Haldar, and Kawin Setsompop. Improving parallel imaging by jointly reconstructing multi-contrast data. *Magnetic Resonance in Medicine*, 80(2):619–632, August 2018.
- [15] Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. Where is the semantic system? A critical review and meta-analysis of 120

- functional neuroimaging studies. *Cerebral Cortex*, 19(12):2767–2796, December 2009.
- [16] Erdem Biyik, Efe Ilicak, and Tolga Çukur. Reconstruction by calibration over tensors for multi-coil multi-acquisition balanced SSFP imaging. *Magnetic Resonance in Medicine*, 79(5):2542–2554, May 2018.
- [17] Mart Bles, Jens Schwarzbach, Peter De Weerd, Rainer Goebel, and Bernadette M Jansma. Receptive field size-dependent attention effects in simultaneously presented stimulus displays. *NeuroImage*, 30(2):506–511, April 2006.
- [18] Kai Tobias Block, Martin Uecker, and Jens Frahm. Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint. *Magnetic Resonance in Medicine*, 57(6):1086–1098, 2007.
- [19] Thierry Blu and Floirán Luisier. The SURE-LET approach to image denoising. *IEEE Transactions on Image Processing*, 16(11):2778–2786, November 2007.
- [20] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, December 2010.
- [21] Frank Bremmer, Anja Schlack, Jean René Duhamel, Werner Graf, and Gereon R Fink. Space coding in primate posterior parietal cortex. In *NeuroImage*, pages S46–51. Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany, January 2001.
- [22] Kenneth H Britten and Hilary W Heuer. Spatial Summation in the Receptive Fields of MT Neurons. *Journal of Neuroscience*, 19(12):5074–5084, June 1999.
- [23] G Buccino, Ferdinand Binkofski, G R Fink, L Fadiga, L Fogassi, V Gallese, R J Seitz, K Zilles, G Rizzolatti, and H J Freund. Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. *European Journal of Neuroscience*, 13(2):400–404, February 2001.

- [24] Neil Burgess, Eleanor A Maguire, Hugo J Spiers, and John O’Keefe. A temporoparietal and prefrontal network for retrieving the spatial context of lifelike events. *NeuroImage*, 14(2):439–453, January 2001.
- [25] Emily Caporello Bluvas and Timothy Q Gentner. Attention to natural auditory signals. *Hearing Research*, 305(1):10–18, November 2013.
- [26] John D Carew, Grace Wahba, Xianhong Xie, Erik V Nordheim, and M Elizabeth Meyerand. Optimal spline smoothing of fMRI time series by generalized cross-validation. *NeuroImage*, 18(4):950–961, April 2003.
- [27] Svenja Caspers, Karl Zilles, Angela R Laird, and Simon B Eickhoff. ALE meta-analysis of action observation and imitation in the human brain. *NeuroImage*, 50(3):1148–1167, April 2010.
- [28] Emin Çelik, Salman Ul Hassan Dar, Özgür Yılmaz, Ümit Keleş, and Tolga Çukur. Spatially informed voxelwise modeling for naturalistic fMRI experiments. *NeuroImage*, 186:741–757, February 2019.
- [29] A Enis Cetin, A Bozkurt, O Gunay, Y H Habiboglu, Kivanc Kose, I Onaran, Mohammad Tofghi, and R A Sevimli. Projections onto convex sets (POCS) based optimization by lifting. In *2013 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 623–623. IEEE, 2013.
- [30] Hsiang-Yun Sherry Chien and Christopher J Honey. Constructing and Forgetting Temporal Context in the Human Cerebral Cortex. *bioRxiv*, 95: 761593, September 2019.
- [31] G Chierchia, N Pustelnik, J C Pesquet, and B Pesquet-Popescu. Epigraphical projection and proximal tools for solving constrained convex optimization problems. *Signal, Image and Video Processing*, 9(8):1737–1749, November 2015.
- [32] Trevor T J Chong, Ross Cunnington, Mark A Williams, and Jason B Mattingley. The role of selective attention in matching observed and executed actions. *Neuropsychologia*, 47(3):786–795, February 2009.

- [33] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3): 215–229, March 2002.
- [34] Daniele Corbo and Guy A Orban. Observing others speak or sing activates spt and neighboring parietal cortex. *Journal of Cognitive Neuroscience*, 29(6):1002–1021, June 2017.
- [35] Tamara C Cristescu, Joseph T Devlin, and Anna C Nobre. Orienting attention to semantic categories. *NeuroImage*, 33(4):1178–1187, December 2006.
- [36] Tolga Çukur. Accelerated phase-cycled SSFP imaging with compressed sensing. *IEEE Transactions on Medical Imaging*, 34(1):107–115, January 2015.
- [37] Tolga Çukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6):763–770, April 2013.
- [38] Tolga Çukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. Visual search for action categories alters semantic representation in the human brain. In *Society for Neuroscience*, page 156.08, Washington, DC, 2014.
- [39] Rhodri Cusack. The intraparietal sulcus and perceptual organization. *Journal of Cognitive Neuroscience*, 17(4):641–651, April 2005.
- [40] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, pages 886–893. INRIA Rhone-Alpes, Saint-Ismier, France, IEEE, December 2005.
- [41] Anders M Dale, Bruce Fischl, and Martin I Sereno. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, 9(2): 179–194, January 1999.

- [42] K D Davis, W D Hutchison, A M Lozano, R R Tasker, and J O Dostrovsky. Human anterior cingulate cortex neurons modulated by attention-demanding tasks. *Journal of Neurophysiology*, 83(6):3575–3577, June 2000.
- [43] Wendy A de Heer, Alexander G Huth, Thomas L Griffiths, Jack L Gallant, and Frédéric E Theunissen. The Hierarchical Cortical Organization of Human Speech Processing. *The Journal of Neuroscience*, 37(27):6539–6557, July 2017.
- [44] Floris P de Lange, Marjolein Spronk, Roel M Willems, Ivan Toni, and Harold Bekkering. Complementary Systems for Understanding Action Intentions. *Current Biology*, 18(6):454–457, March 2008.
- [45] Charles-Alban Deledalle, Samuel Vaiteer, Jalal Fadili, and Gabriel Peyré. Stein Unbiased Gradient estimator of the Risk (SUGAR) for Multiple Parameter Selection. *SIAM Journal on Imaging Sciences*, 7(4):2448–2487, January 2014.
- [46] Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15, October 2010.
- [47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv.org*, October 2018.
- [48] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *The 25th International Conference in Machine Learning*, pages 272–279, New York, New York, USA, 2008. ACM Press.
- [49] Nadiya El-Sourani, Moritz F Wurm, Ima Trempler, Gereon R Fink, and Ricarda I Schubotz. Making sense of objects lying around: How contextual objects shape brain activity during action observation. *NeuroImage*, 167: 429–437, February 2018.

- [50] Yaara Erez and John Duncan. Discrimination of Visual Categories Based on Behavioral Relevance in Widespread Regions of Frontoparietal Cortex. *The Journal of Neuroscience*, 35(36):12383–12393, September 2015.
- [51] Seyedeh-Rezvan Farahibozorg, Richard N Henson, Anna M Woollams, and Olaf Hauk. Distinct roles for the Anterior Temporal Lobe and Angular Gyrus in the spatio-temporal cortical semantic network. *bioRxiv*, February 2019.
- [52] Evelina Fedorenko, Edward Gibson, and Douglas Rohde. The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language*, 54(4):541–553, May 2006.
- [53] Evelina Fedorenko, Edward Gibson, and Douglas Rohde. The nature of working memory in linguistic, arithmetic and spatial integration processes. *Journal of Memory and Language*, 56(2):246–269, February 2007.
- [54] Gidon Felsen and Yang Dan. A natural approach to studying vision. *Nature Neuroscience*, 8(12):1643–1646, December 2005.
- [55] Stefania Ferri, Giacomo Rizzolatti, and Guy A Orban. The organization of the posterior parietal cortex devoted to upper limb actions: An fMRI study. *Human Brain Mapping*, 36(10):3845–3866, October 2015.
- [56] Karl J Friston, C D Frith, R Turner, and R S J Frackowiak. Characterizing evoked hemodynamics with fMRI. *NeuroImage*, 2(2):157–165, January 1995.
- [57] Karl J Friston, A P Holmes, J B Poline, P J Grasby, S C R Williams, R S J Frackowiak, and R Turner. Analysis of fMRI time-series revisited. *NeuroImage*, 2(1):45–53, January 1995.
- [58] Christopher S Furmanski, Denis Schluppeck, and Stephen A Engel. Learning strengthens the response of primary visual cortex to simple patterns. *Current Biology*, 14(7):573–578, April 2004.

- [59] James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9(22):162, September 2015.
- [60] Jeffrey P Gavornik and Mark F Bear. Learned spatiotemporal sequence recognition and prediction in primary visual cortex. *Nature Neuroscience*, 17(5):732–737, January 2014.
- [61] R Giryes, Michael Elad, and Yonina C Eldar. The projected GSURE for automatic parameter tuning in iterative shrinkage methods. *Applied and Computational Harmonic Analysis*, 30(3):407–422, January 2011.
- [62] Gene H Golub, Michael Heath, and Grace Wahba. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21(2):215–223, May 1979.
- [63] Enhao Gong, Feng Huang, Kui Ying, Wenchuan Wu, Shi Wang, and Chun Yuan. PROMISE: Parallel-imaging and compressed-sensing reconstruction of multicontrast imaging using Sharable information. *Magnetic Resonance in Medicine*, 73(2):523–535, February 2015.
- [64] Scott T Grafton and Antonia F de C Hamilton. Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science*, 26(4):590–616, August 2007.
- [65] Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 47(6):1202–1210, June 2002.
- [66] Marie H el ene Grosbras, Susan Beaton, and Simon B Eickhoff. Brain regions involved in human movement perception: A quantitative voxel-based meta-analysis. *Human Brain Mapping*, 33(2):431–454, February 2012.
- [67] Emily D Grossman and Randolph Blake. Brain areas active during visual perception of biological motion. *Neuron*, 35(6):1167–1175, September 2002.

- [68] Chunli Guo and Mike E Davies. Near Optimal Compressed Sensing Without Priors: Parametric SURE Approximate Message Passing. *IEEE Transactions on Signal Processing*, 63(8):2130–2141, April 2015.
- [69] Antonia F de C Hamilton and Scott T Grafton. Goal representation in human anterior intraparietal sulcus. *Journal of Neuroscience*, 26(4):1133–1137, January 2006.
- [70] Giacomo Handjaras, Giulio Bernardi, Francesca Benuzzi, Paolo F Nichelli, Pietro Pietrini, and Emiliano Ricciardi. A topographical organization for action representation in the human brain. *Human Brain Mapping*, 36(10):3832–3844, October 2015.
- [71] Per Christian Hansen. Analysis of Discrete Ill-Posed Problems by Means of the L-Curve. *SIAM Review*, 34(4):561–580, December 1992.
- [72] Per Christian Hansen and Dianne Prost O’Leary. The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems. *SIAM Journal on Scientific Computing*, 14(6):1487–1503, November 1993.
- [73] Assaf Harel, Dwight J Kravitz, and Chris I Baker. Task context impacts visual object processing differentially across the cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(10):E962–71, March 2014.
- [74] Stephenie A Harrison and Frank Tong. Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458(7238):632–635, April 2009.
- [75] U Hasson, E Yang, I Vallines, David J Heeger, and N Rubin. A Hierarchy of Temporal Receptive Windows in Human Cortex. *The Journal of Neuroscience*, 28(10):2539–2550, March 2008.
- [76] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject Synchronization of Cortical Activity during Natural Vision. *Science*, 303(5664):1634–1640, March 2004.

- [77] Uri Hasson, Rafael Malach, and David J Heeger. Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, 14(1): 40–48, January 2010.
- [78] Uri Hasson, Janice Chen, and Christopher J Honey. Hierarchical process memory: Memory as an integral component of information processing. *Trends in Cognitive Sciences*, 19(6):304–313, June 2015.
- [79] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, October 2011.
- [80] James V Haxby, Andrew C Connolly, and J Swaroop Guntupalli. *Decoding neural representational spaces using multivariate pattern analysis*, volume 37. Università degli Studi di Trento, Trento, Italy, January 2014.
- [81] John Herrington, Charlotte Nymberg, Susan Faja, Elinora Price, and Robert Schultz. The responsiveness of biological motion processing areas to selective attention towards goals. *NeuroImage*, 63(1):581–590, October 2012.
- [82] Tom Hilbert, Damien Nguyen, Jean Philippe Thiran, Gunnar Krueger, Tobias Kober, and Oliver Bieri. True constructive interference in the steady state (trueCISS). *Magnetic Resonance in Medicine*, 79(4):1901–1910, April 2018.
- [83] Kevin D Himberger, Hsiang Yun Chien, and Christopher J Honey. Principles of Temporal Processing Across the Cortical Hierarchy. *Neuroscience*, 389:161–174, October 2018.
- [84] Markus Hoeren, Christoph P Kaller, Volkmar Glauche, Magnus-Sebastian Vry, Michel Rijntjes, Farsin Hamzei, and Cornelius Weiller. Action semantics and movement characteristics engage distinct processing streams during the observation of tool use. *Experimental Brain Research*, 229(2):243–260, July 2013.

- [85] M B Holte, T B Moeslund, and P Fihl. View-invariant gesture recognition using 3D optical flow and harmonic motion context. *Computer Vision and Image Understanding*, 114(12):1353–1361, December 2010.
- [86] Christopher J Honey, Thomas Thesen, Tobias H Donner, Lauren J Silbert, Chad E Carlson, Orrin Devinsky, Werner K Doyle, Nava Rubin, David J Heeger, and Uri Hasson. Slow Cortical Dynamics and the Accumulation of Information over Long Timescales. *Neuron*, 76(2):423–434, October 2012.
- [87] Junzhou Huang, Chen Chen, and Leon Axel. Fast multi-contrast MRI reconstruction. *Magnetic Resonance Imaging*, 32(10):1344–1352, December 2014.
- [88] D H Hubel and T N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of Physiology*, 160(1):106–154, January 1962.
- [89] Matthew Hudson, Toby Nicholson, Rob Ellis, and Patric Bach. I see what you say: Prior knowledge of other’s goals automatically biases the perception of their actions. *Cognition*, 146:245–250, January 2016.
- [90] Matthew Hudson, Toby Nicholson, William A Simpson, Rob Ellis, and Patric Bach. One step ahead: The perceived kinematics of others’ actions are biased toward expected goals. *Journal of Experimental Psychology: General*, 145(1):1–7, January 2016.
- [91] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, 76(6):1210–1224, December 2012.
- [92] Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, April 2016.
- [93] Marco Iacoboni, Roger P Woods, Marcel Brass, Harold Bekkering, John C Mazziotta, and Giacomo Rizzolatti. Cortical mechanisms of human imitation. *Science*, 286(5449):2526–2528, December 1999.

- [94] Marco Iacoboni, Istvan Molnar-Szakacs, Vittorio Gallese, G Buccino, and John C Mazziotta. Grasping the intentions of others with one’s own mirror neuron system. In *PLoS Biology*, pages 0529–0535. University of California, Los Angeles, Los Angeles, United States, March 2005.
- [95] Uwe J Ilg, Stefan Schumann, and Peter Thier. Posterior parietal cortex neurons encode target motion in world-centered coordinates. *Neuron*, 43(1):145–151, July 2004.
- [96] Efe Ilicak and Tolga Çukur. Parameter-Free Profile Encoding Reconstruction for Multiple-Acquisition bSSFP Imaging. In *International Society for MR in Medicine (ISMRM)*, October 2017.
- [97] Efe Ilicak, Lutfi Kerem Senel, Erdem Biyik, and Tolga Çukur. Profile-encoding reconstruction for multiple-acquisition balanced steady-state free precession imaging. *Magnetic Resonance in Medicine*, 78(4):1316–1329, October 2016.
- [98] Shailee Jain and Alexander G Huth. Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems*, pages 6628–6637. University of Texas at Austin, Austin, United States, January 2018.
- [99] L Jäncke, T W Buchanan, K Lutz, and N J Shah. Focused and nonfocused attention in verbal and emotional dichotic listening: An FMRI study. *Brain and Language*, 78(3):349–363, January 2001.
- [100] J Jastorff and G A Orban. Human Functional Magnetic Resonance Imaging Reveals Separation and Integration of Shape and Motion Cues in Biological Motion Processing. *The Journal of Neuroscience*, 29(22):7315–7329, June 2009.
- [101] Jan Jastorff, Chiara Begliomini, Maddalena Fabbri-Destro, Giacomo Rizzolatti, and Guy A Orban. Coding observed motor acts: Different organizational principles in the parietal and premotor cortex of humans. *Journal of Neurophysiology*, 104(1):128–140, July 2010.

- [102] Jan Jastorff, Rouhollah O Abdollahi, Fabrizio Fasano, and Guy A Orban. Seeing biological actions in 3D: An fMRI study. *Human Brain Mapping*, 37(1):203–219, January 2016.
- [103] Xiaoming Jiang. Prefrontal Cortex: Role in Language Communication during Social Interaction. In *Prefrontal Cortex*. InTech, October 2018.
- [104] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, June 1973.
- [105] Sabine Kastner, P De Weerd, M A Pinsk, M I Elizondo, Robert Desimone, and Leslie G Ungerleider. Modulation of sensory suppression: implications for receptive field sizes in the human visual cortex. *Journal of Neurophysiology*, 86(3):1398–1411, September 2001.
- [106] A Khare, U S Tiwary, W Pedrycz, and Moongu Jeon. Multilevel adaptive thresholding and shrinkage technique for denoising using daubechies complex wavelet transform. *Imaging Science Journal*, 58(6):340–358, December 2010.
- [107] Kedar Khare, Christopher J Hardy, Kevin F King, Patrick A Turski, and Luca Marinelli. Accelerated MR imaging using compressive sensing with no free parameters. *Magnetic Resonance in Medicine*, 68(5):1450–1457, January 2012.
- [108] Florian Knoll, Kristian Bredies, Thomas Pock, and Rudolf Stollberger. Second order total generalized variation (TGV) for MRI. *Magnetic Resonance in Medicine*, 65(2):480–491, December 2010.
- [109] Lingqiang Kong, Samantha W Michalka, Maya L Rosen, Summer L Shermata, Jascha D Swisher, Barbara G Shinn-Cunningham, and David C Somers. Auditory spatial attention representations in the human cerebral cortex. *Cerebral Cortex*, 24(3):773–784, March 2014.
- [110] Bryan Kressler, Ludovic De Rochefort, Tian Liu, Pascal Spincemaille, Quan Jiang, and Yi Wang. Nonlinear regularization for per voxel estimation of magnetic susceptibility distributions from MRI field maps. *IEEE Transactions on Medical Imaging*, 29(2):273–281, February 2010.

- [111] Nikolaus Kriegeskorte and Rogier A Kievit. Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412, August 2013.
- [112] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(NOV):4, November 2008.
- [113] Rachel J Kyd and David K Bilkey. Prefrontal cortex lesions modify the spatial properties of hippocampal place cells. *Cerebral Cortex*, 13(5):444–451, May 2003.
- [114] Ivan Laptev. On space-time interest points. In *International Journal of Computer Vision*, pages 107–123. Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes, France, Kluwer Academic Publishers, September 2005.
- [115] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes, France, January 2008.
- [116] A M Leaver and J P Rauschecker. Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object Category. *Journal of Neuroscience*, 30(22):7604–7612, June 2010.
- [117] Tatia M C Lee, Mei-kei Leung, Tiffany M Y Lee, Adrian Raine, and Chetwyn C H Chan. I want to lie about not knowing you, but my precuneus refuses to cooperate. *Scientific Reports*, 3:1636, April 2013.
- [118] Yulia Lerner, Christopher J Honey, Lauren J Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience*, 31(8):2906–2915, February 2011.
- [119] Mark D Lescroart and Jack L Gallant. Human Scene-Selective Areas Represent 3D Configurations of Surfaces. *Neuron*, 101(1):178–192.e7, January 2019.

- [120] Gwyneth A Lewis, David Poeppel, and Gregory L Murphy. Contrasting Semantic versus Inhibitory Processing in the Angular Gyrus: An fMRI Study. *Cerebral Cortex*, 29(6):2470–2481, June 2019.
- [121] Dong Liang, Bo Liu, JiunJie Wang, and Leslie Ying. Accelerating SENSE using compressed sensing. *Magnetic Resonance in Medicine*, 62(6):1574–1584, December 2009.
- [122] M Liljeström, A Tarkiainen, T Parviainen, J Kujala, J Numminen, J Hiltunen, M Laine, and R Salmelin. Perceiving and naming actions and objects. *NeuroImage*, 41(3):1132–1141, July 2008.
- [123] Fa-Hsuan Lin, Fu-Nien Wang, Seppo P Ahlfors, Matti S Hämäläinen, and John W Belliveau. Parallel MRI reconstruction using variance partitioning regularization. *Magnetic Resonance in Medicine*, 58(4):735–744, 2007.
- [124] Fausta Lui, G Buccino, Davide Duzzi, Francesca Benuzzi, Girolamo Crisi, Patrizia Baraldi, Paolo Nichelli, Carlo Adolfo Porro, and Giacomo Rizzolatti. Neural substrates for observing and imagining non-object-directed actions. *Social Neuroscience*, 3(3-4):261–275, November 2008.
- [125] Florian Luisier, Thierry Blu, and Michael Unser. A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding. *IEEE Transactions on Image Processing*, 16(3):593–606, March 2007.
- [126] Michael Lustig and John M Pauly. SPIRiT: Iterative self-consistent parallel imaging reconstruction from arbitrary k-space. *Magnetic Resonance in Medicine*, 64(2):457–471, August 2010.
- [127] Michael Lustig, David L Donoho, and John M Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.
- [128] Rogier B Mars and Meike J Grol. Dorsolateral prefrontal cortex, working memory, and prospective coding for action. *Journal of Neuroscience*, 27(8):1801–1802, February 2007.

- [129] William Matchin, Christopher Hammerly, and Ellen Lau. The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *Cortex*, 88:106–123, March 2017.
- [130] Massimo Matelli and Giuseppe Luppino. Parietofrontal circuits for action and space perception in the macaque monkey. In *NeuroImage*, pages S27–32. Università degli Studi di Parma, Parma, Italy, January 2001.
- [131] J H Maunsell and W T Newsome. Visual processing in monkey extrastriate cortex. *Annual Review of Neuroscience*, 10:363–401, 1987.
- [132] Samantha W Michalka, Maya L Rosen, Lingqiang Kong, Barbara G Shinn-Cunningham, and David C Somers. Auditory Spatial Coding Flexibly Recruits Anterior, but Not Posterior, Visuotopic Parietal Cortex. *Cerebral Cortex*, 26(3):1302–1308, March 2016.
- [133] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, November 1995.
- [134] Suresh D Muthukumaraswamy and Krish D Singh. Modulation of the human mirror neuron system during cognitive activity. *Psychophysiology*, 45(6):896–905, November 2008.
- [135] Emese Nagy, Mario Liotti, Steven Brown, Gordon Waiter, Andrew Bromiley, Colwyn Trevarthen, and Gyorgy Bardos. The neural mechanisms of reciprocal communication. *Brain Research*, 1353:159–167, September 2010.
- [136] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, May 2011.
- [137] Samuel A Nastase, Andrew C Connolly, Nikolaas N Oosterhof, Yaroslav O Halchenko, J Swaroo Guntupalli, Matteo Visconti Di Oleggio Castello, Jason Gors, M Id Gobbini, and James V Haxby. Attention selectively reshapes the geometry of distributed semantic representation. *Cerebral Cortex*, 27(8):4277–4291, August 2017.
- [138] Samuel A Nastase, Yaroslav O Halchenko, Andrew C Connolly, M Ida Gobbini, and James V Haxby. Neural responses to naturalistic clips of behaving

- animals in two different task contexts. *Frontiers in Neuroscience*, 12(MAY):316, May 2018.
- [139] Koen Nelissen, Giuseppe Luppino, Wim Vanduffel, Giacomo Rizzolatti, and Guy A Orban. Observing others: Multiple action representation in the frontal lobe. *Science*, 310(5746):332–336, October 2005.
- [140] Koen Nelissen, Elena Borra, Marzio Gerbella, Stefano Rozzi, Giuseppe Luppino, Wim Vanduffel, Giacomo Rizzolatti, and Guy A Orban. Action observation circuits in the macaque monkey cortex. *Journal of Neuroscience*, 31(10):3743–3756, March 2011.
- [141] Roger Newman-Norlund, Hein T van Schie, Marline E C van Hoek, Raymond H Cuijpers, and Harold Bekkering. The role of inferior frontal and parietal areas in differentiating meaningful and meaningless object-directed actions. *Brain Research*, 1315:63–74, February 2010.
- [142] Toby Nicholson, Matt Roser, and Patric Bach. Understanding the Goals of Everyday Instrumental Actions Is Primarily Linked to Object, Not Motor-Kinematic, Information: Evidence from fMRI. *PLoS ONE*, 12(1):e0169700, 2017.
- [143] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, 21(19):1641–1646, October 2011.
- [144] Tömme Noesselt, Nadim Jon Shah, and Lutz Jäncke. Top-down and bottom-up modulation of language related areas - An fMRI study. *BMC Neuroscience*, 4(1):13–12, June 2003.
- [145] Matthijs L Noordzij, Sarah E Newman-Norlund, Jan Peter De Ruyter, Peter Hagoort, Stephen C Levinson, and Ivan Toni. Neural correlates of intentional communication. *Frontiers in Neuroscience*, 4(DEC):188, December 2010.

- [146] Uta Noppeney. The neural systems of tool and action semantics: A perspective from functional imaging. *Journal of Physiology Paris*, 102(1-3): 40–49, January 2008.
- [147] Lauri Nurminen, Sam Merlin, Maryam Bijanzadeh, Frederick Federer, and Alessandra Angelucci. Top-down feedback controls spatial summation and response amplitude in primate visual cortex. *Nature Communications*, 9(1):2281, December 2018.
- [148] Lindsay M Oberman, Jaime A Pineda, and Vilayanur S Ramachandran. The human mirror neuron system: A link between action observation and social skills. *Social Cognitive and Affective Neuroscience*, 2(1):62–66, March 2007.
- [149] Jonas Obleser, Richard J S Wise, M Alex Dresner, and Sophie K Scott. Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience*, 27(9):2283–2289, February 2007.
- [150] Guy A Orban. Action observation as a visual process: Different classes of actions engage distinct regions of human PPC. In *Cultural Patterns and Neurocognitive Circuits II: East-West Connections*, pages 1–32. December 2017.
- [151] James O’Sullivan, Jose Herrero, Elliot Smith, Catherine Schevon, Guy M McKhann, Sameer A Sheth, Ashesh D Mehta, and Nima Mesgarani. Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. *Neuron*, October 2019.
- [152] Sofia Paneri and Georgia G Gregoriou. Top-down control of visual attention by the prefrontal cortex. Functional specialization and long-range interactions. *Frontiers in Neuroscience*, 11(SEP):545, September 2017.
- [153] Neal Parikh. Proximal Algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

- [154] Gloria G Parras, Javier Nieto-Diego, Guillermo V Carbajal, Catalina Valdés-Baizabal, Carles Escera, and Manuel S Malmierca. Neurons along the auditory pathway exhibit a hierarchical organization of prediction error. *Nature Communications*, 8(1):2148, December 2017.
- [155] Marius V Peelen, Li Fei-Fei, and Sabine Kastner. Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, 460(7251):94–97, July 2009.
- [156] L Pizzamiglio, T Aprile, G Spitoni, S Pitzalis, E Bates, S D’Amico, and F Di Russo. Separate neural systems for processing action- or non-action-related sounds. *NeuroImage*, 24(3):852–861, February 2005.
- [157] Daniel A Pollen, Andrzej W Przybyszewski, Mark A Rubin, and Warren Foote. Spatial receptive field organization of macaque V4 neurons. *Cerebral Cortex*, 12(6):601–616, June 2002.
- [158] Bradley R Postle. Working memory as an emergent property of the mind and brain. *Neuroscience*, 139(1):23–38, April 2006.
- [159] K P Pruessmann, M Weiger, M B Scheidegger, and Peter Boesiger. SENSE: Sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine*, 42(5):952–962, November 1999.
- [160] Radek Ptak, Armin Schnider, and Julia Fellrath. The Dorsal Frontoparietal Network: A Core System for Emulated Action. *Trends in Cognitive Sciences*, 21(8):589–599, August 2017.
- [161] Sathish Ramani, Zhihao Liu, Jeffrey Rosen, Jon Fredrik Nielsen, and Jeffrey A Fessler. Regularization parameter selection for nonlinear iterative image restoration and mri reconstruction using GCV and SURE-based methods. *IEEE Transactions on Image Processing*, 21(8):3659–3672, July 2012.
- [162] Richard Ramsey and Antonia F D C Hamilton. Understanding actors and object-goals in the human brain. *NeuroImage*, 50(3):1142–1147, April 2010.

- [163] Leila Reddy and Nancy G Kanwisher. Category Selectivity in the Ventral Visual Pathway Confers Robustness to Clutter and Diverted Attention. *Current Biology*, 17(23):2067–2072, December 2007.
- [164] Leila Reddy, Nancy G Kanwisher, and Rufin VanRullen. Attention and biased competition in multi-voxel object representations. *Proceedings of the National Academy of Sciences of the United States of America*, 106(50):21447–21452, December 2009.
- [165] Mor Regev, Erez Simony, Katherine Lee, Kean Ming Tan, Janice Chen, and Uri Hasson. Propagation of information along the cortical hierarchy as a function of attention while reading and listening to stories. *Cerebral Cortex*, 3:e784, April 2018.
- [166] Teresa Regińska. A Regularization Parameter in Discrete Ill-Posed Problems. *SIAM Journal on Scientific Computing*, 17(3):740–749, May 1996.
- [167] Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418, July 2012.
- [168] Mark Rijpkema, Sandra I van Aalderen, Jens V Schwarzbach, and Frans A J Verstraten. Activation patterns in visual cortex reveal receptive field size-dependent attentional modulation. *Brain Research*, 1189:90–96, January 2008.
- [169] Dario L Ringach. Mapping receptive fields in primary visual cortex. *Journal of Physiology*, 558(3):717–728, August 2004.
- [170] G Rizzolatti, L Fogassi, and V Gallese. Parietal cortex: from sight to action. *Current Opinion in Neurobiology*, 7(4):562–567, August 1997.
- [171] Giacomo Rizzolatti and Massimo Matelli. Two different streams form the dorsal visual system: Anatomy and functions. In *Experimental Brain Research*, pages 146–157. Università degli Studi di Parma, Parma, Italy, November 2003.

- [172] Jennifer M Rodd, Matthew H Davis, and Ingrid S Johnsrude. The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, 15(8):1261–1269, August 2005.
- [173] Corianne Rogalsky and Gregory Hickok. Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cerebral Cortex*, 19(4):786–796, April 2009.
- [174] Ashley S Safford, Elizabeth A Hussey, Raja Parasuraman, and James C Thompson. Object-based attentional modulation of biological motion processing: spatiotemporal dynamics using functional magnetic resonance imaging and electroencephalography. *The Journal of Neuroscience*, 30(27):9064–9073, July 2010.
- [175] Stefanie Schuch, Andrew P Bayliss, Christoph Klein, and Steven P Tipper. Attention modulates motor system activation during action observation: Evidence for inhibitory rebound. *Experimental Brain Research*, 205(2):235–249, August 2010.
- [176] Mohammad Shahdloo, Efe Ilicak, Mohammad Tofghi, Emine Ulku Saritas, A Enis Cetin, and Tolga Çukur. Adaptive Wavelet Thresholding for Profile-Encoding Reconstruction of Balanced Steady-State Free Precession Acquisitions. In *European Society for MR and Medicine (ESMRMB)*, page 391, Barcelona, 2017.
- [177] Mohammad Shahdloo, Efe Ilicak, Mohammad Tofghi, Emine Ulku Saritas, A Enis Cetin, and Tolga Çukur. Rapid Self-Tuning Compressed-Sensing MRI Using Projection onto Epigraph Sets. In *International Society for MR in Medicine (ISMRM)*, page 0251, Paris, 2018.
- [178] Mohammad Shahdloo, Mert Acar, and Tolga Çukur. Attention During Story Listening Modulates Temporal Receptive Windows Across Human Cortex. In *Conference on Cognitive Computational Neuroscience (CCN)*, page PS1A.52, Berlin, 2019.
- [179] Mohammad Shahdloo, Emin Çelik, and Tolga Çukur. Biased Competition

- in Semantic Representation During Natural Visual Search. *NeuroImage*, page 116383, November 2019.
- [180] Mohammad Shahdloo, Efe Ilicak, Mohammad Tofghi, Emine Ulku Saritas, A Enis Cetin, and Tolga Çukur. Projection onto Epigraph Sets for Rapid Self-Tuning Compressed Sensing MRI. *IEEE Transactions on Medical Imaging*, 38(7):1677–1689, July 2019.
- [181] Mohammad Shahdloo, Burcu A Urgan, and Tolga Çukur. Attention to Action Categories Shifts Semantic Tuning Toward Targets Across the Brain. In *Organization for Human Brain Mapping (OHBM)*, page T661, Rome, 2019.
- [182] Bennett A Shaywitz, Sally E Shaywitz, Kenneth R Pugh, Robert K Fulbright, Pawel Skudlarski, W Einar Mencl, Patrick R Constable, Karen E Marchione, Jack M Fletcher, Rafael Klorman, Cheryl Lacadie, and John C Gore. The functional neural architecture of components of attention in language-processing tasks. *NeuroImage*, 13(4):601–612, January 2001.
- [183] D L Sheinberg and N K Logothetis. Noticing familiar objects in real world scenes: The role of temporal cortical neurons in natural vision. *Journal of Neuroscience*, 21(4):1340–1350, 2001.
- [184] Peter J Shin, Peder E Z Larson, Michael A Ohliger, Michael Elad, John M Pauly, Daniel B Vigneron, and Michael Lustig. Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion. *Magnetic Resonance in Medicine*, 72(4):959–970, November 2013.
- [185] Yuri Shtyrov. Automaticity and attentional control in spoken language processing: neurophysiological evidence. *The Mental Lexicon*, 5(2):255–276, 2010.
- [186] Jeremy I Skipper, Susan Goldin-Meadow, Howard C Nusbaum, and Steven L Small. Speech-associated gestures, Broca’s area, and the human mirror system. *Brain and Language*, 101(3):260–277, June 2007.
- [187] Stephen M Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, November 2002.

- [188] S Sourbron, R Luypaert, P Van Schuerbeek, M Dujardin, and T Stadnik. Choice of the regularization parameter for perfusion quantification with MRI. *Physics in Medicine and Biology*, 49(14):3307–3324, July 2004.
- [189] Steven Sourbron, Rob Luypaert, Peter Van Schuerbeek, Martine Dujardin, Tadeusz Stadnik, and Michel Osteaux. Deconvolution of dynamic contrast-enhanced MRI data by linear inversion: Choice of the regularization parameter. *Magnetic Resonance in Medicine*, 52(1):209–213, July 2004.
- [190] Kartik K Sreenivasan, Clayton E Curtis, and Mark D’Esposito. Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, 18(2):82–89, January 2014.
- [191] C M STEIN. Estimation of the Mean of a Multivariate Normal-Distribution. *Annals of Statistics*, 9(6):1135–1151, 1981.
- [192] Cornelia Stoeckel, Patricia M Gough, Kate E Watkins, and Joseph T Devlin. Supramarginal gyrus involvement in visual word recognition. *Cortex*, 45(9):1091–1096, October 2009.
- [193] Leyla Tarhan and Talia Konkle. Sociality and Interaction Envelope Organize Visual Action Representations. *bioRxiv*, (5):1443, 2019.
- [194] M Tervaniemi, S Kruck, W De Baene, E Schröger, K Alter, and A D Friederici. Top-down modulation of auditory processing: Effects of sound context, musical expertise and attentional focus. *European Journal of Neuroscience*, 30(8):1636–1642, October 2009.
- [195] James Thompson and Raja Parasuraman. Attention, biological motion, and action recognition. *NeuroImage*, 59(1):4–13, January 2012.
- [196] Steven P Tipper, Matthew A Paul, and Amy E Hayes. Vision-for-action: The effects of object property discrimination and action state on affordance compatibility effects. *Psychonomic Bulletin & Review*, 13(3):493–498, January 2006.
- [197] M Toepper, H Gebhardt, T Beblo, C Thomas, M Driessen, M Bischoff, C R Blecker, D Vaitl, and G Sammer. Functional correlates of distractor

- suppression during spatial working memory encoding. *Neuroscience*, 165 (4):1244–1253, February 2010.
- [198] Mohammad Tofghi, Kivanc Kose, and A Enis Cetin. Denoising images corrupted by impulsive noise using projections onto the epigraph set of the total variation function (PES-TV). *Signal, Image and Video Processing*, 9 (S1):41–48, October 2015.
- [199] Joshua D Trzasko and Armando Manduca. Calibrationless parallel MRI using CLEAR. In *Conference Record - Asilomar Conference on Signals, Systems and Computers*, pages 75–79. Mayo Clinic, Rochester, United States, IEEE, December 2011.
- [200] Martin Uecker, Peng Lai, Mark J Murphy, Patrick Virtue, Michael Elad, John M Pauly, Shreyas S Vasanaawala, and Michael Lustig. ESPIRiT-an eigenvalue approach to autocalibrating parallel MRI: Where SENSE meets GRAPPA. *Magnetic Resonance in Medicine*, 71(3):990–1001, May 2013.
- [201] Burcu A Urgan, Selen Pehlivan, and Ayse P Saygin. Distinct representations in occipito-temporal, parietal, and premotor cortex during action perception revealed by fMRI and computational modeling. *Neuropsychologia*, 127:35–47, April 2019.
- [202] Wessel O van Dam, Shirley-Ann Rueschemeyer, and Harold Bekkering. How specifically are action verbs represented in the neural motor system: An fMRI study. *NeuroImage*, 53(4):1318–1325, December 2010.
- [203] Dimitri Van de Ville and Michel Kocher. Nonlocal means with dimensionality reduction and SURE-based parameter selection. *IEEE Transactions on Image Processing*, 20(9):2683–2690, September 2011.
- [204] Frank Van Overwalle. Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, 30(3):829–858, March 2009.
- [205] Wim Vanduffel, Qi Zhu, and Guy A Orban. Monkey Cortex through fMRI Glasses. *Neuron*, 83(3):533–550, August 2014.

- [206] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv.org*, June 2017.
- [207] Julia V Velikina, Andrew L Alexander, and Alexey Samsonov. Accelerating MR parameter mapping using sparsity-promoting regularization in parametric dimension. *Magnetic Resonance in Medicine*, 70(5):1263–1273, November 2013.
- [208] Timothy D Verstynen and Vibhas Deshpande. Using pulse oximetry to account for high and low frequency physiological artifacts in the BOLD signal. *NeuroImage*, 55(4):1633–1644, April 2011.
- [209] M Visser, E Jefferies, and M A Lambon Ralph. Semantic processing in the anterior temporal lobes: A meta-analysis of the functional neuroimaging literature. *Journal of Cognitive Neuroscience*, 22(6):1083–1094, June 2010.
- [210] C R Vogel. Non-convergence of the L-curve regularization parameter selection method. *Inverse Problems*, 12(4):535–547, August 1996.
- [211] Katharina Von Kriegstein, Evelyn Eger, Andreas Kleinschmidt, and Anne Lise Giraud. Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, 17(1): 48–55, January 2003.
- [212] Catherine Wacongne, Etienne Labyt, Virginie van Wassenhove, Tristan A Bekinschtein, Lionel Naccache, and Stanislas Dehaene. Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 108(51):20754–20759, December 2011.
- [213] Pascal Wallisch and J Anthony Movshon. Structure and Function Come Unglued in the Visual Cortex. *Neuron*, 60(2):195–197, October 2008.
- [214] Peter H Weiss, Nuh N Rahbari, Silke Lux, Uwe Pietrzyk, Johannes Noth, and Gereon R Fink. Processing the spatial configuration of complex actions involves right posterior parietal cortex: An fMRI study with clinical implications. *Human Brain Mapping*, 27(12):1004–1014, December 2006.

- [215] D H Weissman. Dorsal Anterior Cingulate Cortex Resolves Conflict from Distracting Stimuli by Boosting Attention toward Relevant Events. *Cerebral Cortex*, 15(2):229–237, July 2004.
- [216] D H Weissman, B Giesbrecht, A W Song, G R Mangun, and M G Woldorff. Conflict monitoring in the human anterior cingulate cortex during selective attention to global and local object features. *NeuroImage*, 19(4):1361–1368, August 2003.
- [217] Christine D Wilson-Mendenhall, W Kyle Simmons, Alex Martin, and Lawrence W Barsalou. Contextual processing of abstract concepts reveals neural representations of nonlinguistic semantic content. *Journal of Cognitive Neuroscience*, 25(6):920–935, January 2013.
- [218] Moritz F Wurm and Alfonso Caramazza. Lateral occipitotemporal cortex encodes perceptual components of social actions rather than abstract representations of sociality. *NeuroImage*, 202:116153, November 2019.
- [219] Moritz F Wurm and Ricarda I Schubotz. Squeezing lemons in the bathroom: Contextual information modulates action recognition. *NeuroImage*, 59(2):1551–1559, January 2012.
- [220] Moritz F Wurm, Alfonso Caramazza, and Angelika Lingnau. Action Categories in Lateral Occipitotemporal Cortex Are Organized Along Sociality and Transitivity. *The Journal of Neuroscience*, 37(3):562–575, January 2017.
- [221] Leslie Ying, Dan Xu, and Zhi-Pei Liang. On Tikhonov regularization for image reconstruction in parallel MRI. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2:1056–1059, 2004.
- [222] Tao Zhang, John M Pauly, and Ives R Levesque. Accelerating parameter mapping with a locally low rank constraint. *Magnetic Resonance in Medicine*, 73(2):655–661, February 2015.

- [223] Xiao Ping Zhang and Mita D Desai. Adaptive denoising based on SURE risk. *IEEE Signal Processing Letters*, 5(10):265–267, December 1998.
- [224] Bo Zhao, Wenmiao Lu, T Kevin Hitchens, Fan Lam, Chien Ho, and Zhi-Pei Liang. Accelerated MR parameter mapping with low-rank and sparsity constraints. *Magnetic Resonance in Medicine*, 74(2):489–498, August 2015.
- [225] Xin Zhou, Fumi Katsuki, Xue-Lian Qi, and Christos Constantinidis. Neurons with inverted tuning during the delay periods of working memory tasks in the dorsal prefrontal and posterior parietal cortex. *Journal of Neurophysiology*, 108(1):31–38, July 2012.
- [226] Lin L Zhu and Michael S Beauchamp. Mouth and voice: A relationship between visual and auditory preference in the human superior temporal sulcus. *Journal of Neuroscience*, 37(10):2697–2708, March 2017.
- [227] Elana M Zion Golumbic, Nai Ding, Stephan Bickel, Peter Lakatos, Catherine A Schevon, Guy M McKhann, Robert R Goodman, Ronald Emerson, Ashesh D Mehta, Jonathan Z Simon, David Poeppel, and Charles E Schroeder. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, 77(5):980–991, January 2013.