Neural dynamics and interactions in the human ventral visual pathway

Yuanning Li

December 2018

Center for the Neural Basis of Cognition,
Dietrich College of Humanities and Social Sciences
and
Machine Learning Department,
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Avniel Singh Ghuman (Co-Chair),
Max G. G'Sell (Co-Chair),
Robert E. Kass,
Christopher I. Baker (National Institute of Mental Health)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

This work was supported by the National Institute on Drug Abuse Predoctoral Training Grants under R90DA023426 and R90DA023420, the National Institute of Mental Health under awards R01MH107797 and R21MH103592, and the National Science Foundation under award 1734907. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

Keywords: Cognitive neuroscience, computational neuroscience, machine learning, statistical methods, decoding, functional connectivity, neuroimaging, intracranial electroencephalography, visual perception

To all the beautiful souls that lit up my life

Abstract

The ventral visual pathway in the brain plays central role in visual object recognition. The classical model of the ventral visual pathway, which poses it as a hierarchical, distributed and feed-forward network, does not match the actual structure of the pathway, which is highly interconnected with reciprocal and non-hierarchical projections. Here we address three major consequences of this non-classical structure with regard to neural dynamics and interactions: (i) the model does not consider any extended information processing dynamics; (ii) the model does not allow for adaptive and recurrent interactions between areas; (iii) the model only characterizes evoked-response with no state-dependence from the neural context. To begin to address these gaps in the classical model, we focus on the categorical-selective regions in the ventral pathway and study the neural dynamics and interactions using intracranial electroencephalography (iEEG), which overcomes the limitations of spatiotemporal resolution in current non-invasive human neuroimaging techniques.

With respect to the first consequence, we applied multivariate pattern analysis (MVPA) methods to the iEEG signal to analyze the dynamic roles of the word and face sensitive areas. We found that both areas demonstrated a similar multi-stage information processing dynamic wherein the representation in category-selective fusiform gyrus evolves from a gist category-level and similarity-based representation to an invariant and highly detailed individual representation over the course of 500 ms. In addition, our results also suggest a dissociation between structural and motion in the face processing streams.

Regarding to the second consequence, we introduced a novel method termed Multi-Connection Pattern Analysis (MCPA) to extract the discriminant information about cognitive states solely from the shared activity between neural populations from the interacting brain areas. Our results on iEEG and fMRI data with MCPA support the hypothesis that individual-level exemplar information is not only encoded by the population activity within certain brain populations, but also represented through recurrent interactions between multiple distributed populations at the network level.

Finally, to address the third consequence, we designed a two-stage generalized linear model to study the relationship between category tuning and the ongoing neural activity in category selective cortical areas. We used this model to demonstrate that endogenous activity modulates the category selective tuning in the post-stimulus evoked response, and the same aspects of endogenous activity that modulate tuning also predict perceptual behavior.

Taken together, in this thesis we develop and apply statistical methods to assess the properties of the non-classical structure in the ventral visual stream, and highlight contributions of regions to multiple stages of processing through interactive and distributed computation that is influenced by ongoing neural context.

Acknowledgments

I would like to thank my advisor, Avniel Ghuman, for all his guidance and support over the past five years. This thesis would not be possible without him taking me into his lab five years ago, when I was struggling to find my way toward a Ph.D. in computational neuroscience. I was fortunate to join the lab as the first grad student, and to be involved the process of shaping and developing a research program in the exciting intersection between cognitive and computational neuroscience. From him I learned not only knowledge and experience in doing cognitive neuroscience, but also pure enthusiasm about science and nature. I still remember vividly the countless times of brief conversation at the end of a day turning into hours of deep discussions on sparking research ideas. I am always amazed by his insights and ability to identify the critical scientific questions behind the experiments and data.

I would like to thank my co-advisor, Max G'Sell, for teaching me how to think in statistics and how to applied statistics in science. From him I learned to always think about the validity of my model assumptions, and always be cautious about the interpretations. Working with Avniel and Max together is the best thing a graduate student could ask for. They would challenge me with both the scientific merit and methodological rigorousness. I also got to learn how to communicate neuroscience to statistics and machine learning community, and how to talk about statistics and machine learning with the neuroscience community.

I also would like to thank my committee members, Rob Kass and Chris Baker, for their help over the course of writing up this thesis. It is always a pleasure to interacting with Chris. Everyone in the PNC program owes a lot to Rob. Rob cared about my progress throughout my Ph.D. years and introduced me to Max. Every time we talked, Rob would always offer me his wise advice. I would like to thank my collaborators on the projects presented in this thesis: Mark Richardson, Julie Fiez, Michael Ward, Elizabeth Hirshorn, and Nicolas Brunet. None of these would be possible without their great intelligence and hard work. I want to thank Mark for always making time from his tight neurosurgeon's schedule to keep track of the progress of our projects and to make insightful feedbacks. And I want to thank Mike for helping to collect and maintain the tremendous amount of experimental data and records over the years, and his artistic suggestions on many of the graphics.

Many people in CMU and Pitt community have helped and influenced me in various ways over the years. I want to thank Byron Yu for introducing me into the realm of computational neuroscience, and for being an inspiring role model to me. I want to thank many amazing professors who have taught me neuroscience, statistics and machine learning, and have given me numerous suggestions and feedbacks on my research: Marlene Behrmann, Michael Tarr, David Plaut, Carl Olson, Ryan Tibshirani, Larry Wasserman, Tom Mitchell, Aarti Singh, Barnabás Póczos and many others. I want to thank my current and former lab mates: Matt Boring, Brett Bankson, Arish Alreja, Shahir Mowlaei, Maxwell Wang, Zachary Jessen, Roma Konecky, Ellyanna Kessler, Shawn Walls and Laura Morett, and thank my brilliant fellow graduate stu-

dents and postdocs, Shupeng Sun, Fa Wang, Pengcheng Zhou, Ying Yang, Josue Orellana, Mariya Toneva, Tong Liu, Elissa Aminoff, John Pyles, Witold Lipski and many others, for their help and support along the journey. I am also very grateful to the excellent staff members of the Center for the Neural Basis of Cognition and the Machine Learning Department, especially Melissa Stupka and Diane Stidle.

I am very lucky to have a few dear long time friends to share all the joys and sorrows in our lives. I want to thank Shi Gu, Chen Zhang and Zhan Su for the years we stuck together through the ups and downs in our lives towards doctorate from different corners of the planet. I am also thankful for my friends in the Internet communities, 67373 and 710676. Many of them I may have never met in person, but they have inspired me and lit up my life in many different ways.

Last but not least, I want to thank my family, especially my parents and grandparents for their love and support. I want to thank my girlfriend, Xiaochen Zhang, for the years we lived through and for the years to come.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Vision is extremely important for human beings in almost every aspect of daily life. A great amount of the information that human brain receives is through vision, and a great portion of the cortex is devoted to vision (DiCarlo et al., 2012; Felleman and Van Essen, 1991). It is crucial for humans to process visual information and recognize visual stimuli rapidly and reliably, and humans have tremendous ability to recognize objects. For example, we can effortlessly detect and identify objects from among tens of thousands of possibilities within less than a second, despite the huge amount of variation in the appearance of the objects and the environment. Understanding how visual information is represented and processed in the brain is one of the central questions in cognitive neuroscience. Throughout the years, researchers have studied the visual system and made significant breakthrough in understanding how vision works in the brain, from cellular to system level.

The prevalent view of visual recognition is that visual information is represented and processed in a hierarchical and distributed manner that involves multiple brain areas and circuits (Felleman and Van Essen, 1991; Haxby et al., 2001; Mishkin et al., 1983). This framework has led to many successful computational models of visual system, such as Neocognitron (Fukushima and Miyake, 1982), HMAX (Riesenhuber and Poggio, 1999), and convolutional neural network (LeCun et al., 1989). In such framework, the information processing proceeds along two distinct cortical pathways, the ventral stream and the dorsal stream, which are mainly involved in object and spatial vision respectively (Kravitz et al., 2013; Mishkin et al., 1983) (Figure 1.1A). With such framework, it is essential to address the following two questions:

(1) What is the role of each brain area that is involved in the network? In other words, what information is encoded and processed in each area of the network? This would lead to the encoding and decoding problems in neuroscience (Averbeck et al., 2006; Dayan and Abbott, 2001). In the encoding problem, we try to understand that given a specific visual stimulus, how would it be represented in the neural activity of the corresponding brain area? For the decoding problem, on the other hand, we try to understand that given a specific pattern of neural activity, what is the triggering visual input.

(2) How are different areas connected? For each area of interest, what other areas is it com-

municating to during the process of object perception? And how are these connections facilitating or contributing to object perception? This leads to the concept of functional connectivity (Friston et al., 1993; Gerstein and Perkel, 1969; Gerstein et al., 1989), which describes the temporal correlation between neural activity measured from distinct neural populations.

In this thesis, we mainly focus on the ventral stream and the process of visual object recognition, and elaborate on these two questions from novel aspects.

The ventral stream gets its input from the lateral geniculate nucleus (LGN) and proceeds along the occipitotemporal cortex (areas V1, V2, V3, V4, lateral occipital complex [LOC], etc.) into anterior inferior temporal cortex (aIT) (see Figure 1.1). Functionally, as we progress along the stream, the size of the neuronal receptive field becomes larger, and the tuning properties of the neurons become more complex and abstract. V1 neurons have small receptive field size and are tuned to directed lines, while IT neurons have larger receptive field size and are tuned to more complex categorical information, such as cats and flowers (Kravitz et al., 2013). Moreover, these category-selective neurons often clustered in different regions of the ventral temporal cortex, forming different category-selective cortical regions. A large body of neuroimaging studies have revealed multiple distinct regions, primarily located in the temporal cortex, that demonstrate functional selectivity to specific categories of visual stimuli, e.g. the fusiform face area (Kanwisher et al., 1997), the parahippocampal place area (Epstein and Kanwisher, 1998), the visual word form area (Cohen et al., 2000), etc. These category-selective brain regions play central role in the visual cognition of objects, and damage to these regions can lead to selective deficits with respect to the visual cognition of the specific category of objects (Farah, 2004). (Cohen et al., 2000; Downing et al., 2001; Epstein and Kanwisher, 1998; Kanwisher et al., 1997; Martin, 2007). The category-level functional specificity alone apparently does not constitute the full story of visual perception. Throughout the years, many studies are devoted to investigate different aspects of the perceptual process across different levels. For example, with regard to the processes of face and word recognition, multiple regions of interest have been identified, and their coding, temporal dynamics, as well as interactions between them, have been investigated (Nestor et al., 2011; Puce et al., 1999; Sugase et al., 1999; Wandell, 2011). In this thesis, we particularly focus on the dynamic properties and interactions within and between these category-selective regions,

The classical hierarchical framework of the ventral stream comes from the static view that the information flows in one direction from posterior to anterior (Figure 1.1B,C). Each node in the network acts as a passive filter that cumulates inputs from the outputs of the previous layers and feeds the information forward after performing local computations. This classical framework captures many of the key characteristics of the ventral pathway and has been employed in many visual cognition models with significant success (Riesenhuber and Poggio, 2000; Yamins et al., 2014). To facilitate such a static feedforward model, we would expect that the actual neural anatomical structure of the cortex is also largely feedforward. However, a close look at the actual neural anatomy reveals that the majority of the connections between areas in the ventral visual stream are reciprocal, and the only one-directional connections are actually feedback connections coming from top-down projections (Kravitz et al., 2013). This inconsistency between the computational model and the actual anatomical structure in the ventral visual stream would lead to several important caveats that the model cannot fully account for. In this thesis, we primarily

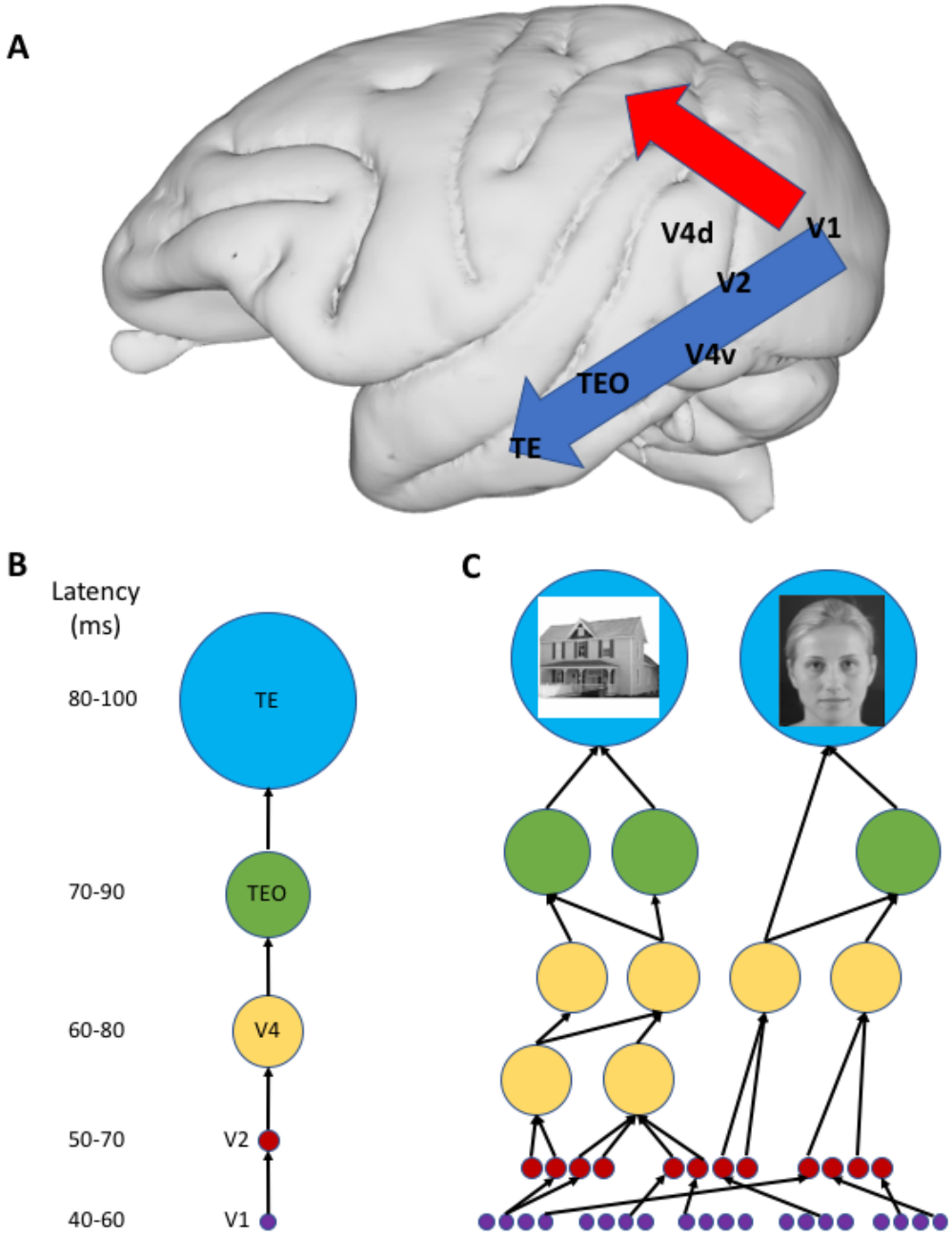


Figure 1.1: **Classical framework of visual object perception. A)** The ventral (blue) and dorsal (red) pathways in the macaque monkey brain. **B)** The classical hierarchical view of visual information processing in the ventral pathway. The approximate range of latencies of first response in each area is shown on the left. Adapted from (Kravitz et al., 2013). **C)** The classicial hierarchical view of the feedforward computations in ventral visual pathway as described in the HMAX model (Riesenhuber and Poggio, 2000; Serre et al., 2007a). Adapted from (Kravitz et al., 2013)

focus on three of the gaps in the classical framework:

(i) the model does not consider any extended information processing dynamics;

(ii) the model does not allow for adaptive and recurrent interactions between areas;

(iii) the model only characterizes evoked-response with no state-dependence from the neural context.

From a methodological point of view, two of the major methods that have been developed to address these two questions are multivariate pattern analysis (MVPA) and functional connectivity analysis.

MVPA uses classification techniques from statistical machine learning to decode the patterns of multivariate neural activity with respect to different stimuli or cognitive conditions, and to infer the underlying neural coding within certain neural populations (Haxby et al., 2001; Norman et al., 2006). The rationale behind MVPA is that if a region is contributed to the visual perception of a certain category, then it must be involved in the encoding of the information, and has discriminant neural representation for the categorical information. And this representational difference would lead to different visual perception. On the other hand, if we read out this discriminant neural representation, we can decode the visual information. Therefore, we can analyze the role of the area in the visual perception process by decoding the visual representation in a categorical-selective region, and see what kind of discriminant information is represented in that area. Therefore, two major factors with regard to MVPA should be noticed and will be addressed in this thesis: 1) the outcome of MVPA depends on the features fed into the model, and we can address different questions by manipulating the input features of the model 2) MVPA is a correlational methods, and no causal link can be directly established through MVPA.

Functional connectivity assumes that statistical dependence between neural signals from different areas implies information communication between regions (Friston et al., 1993). Functional connectivity has been applied to find the interaction between regions in the visual perceptual system, e.g. Ishai (2008); Nestor et al. (2011). However, the prevalent functional connectivity methods can only answer the "yes/no" question with regard to the interactions between brain areas, i.e. whether or not two areas are talking to each other, but are not able to probe the "how" question, i.e. how are the two areas talking to each other and what information is being communicated through the interactions.

1.2 Overview of Thesis Contributions and Structure

1.2.1 Information processing dynamics

The first gap is about the extended information processing dynamics beyond the simple hierarchical feed-forward sweep along the pathway. According to this feedforward framework, the whole process, as a serial information flow, would finish within $\sim$100 ms as a result of the synaptic delay from V1 to aIT (Figure 1.1B). However, as the anatomical evidence suggests, there are both feedforward and feedback connections between areas in the ventral pathway, and majority of the connections between these areas are indeed reciprocal (Kravitz et al., 2013). In addition, electrophysiological studies have also identified late time signatures that account for object

recognition, which suggests feedback and recurrent interactions (Puce et al., 1999; Sugase et al., 1999). Therefore, what is the timecourse of information processing in each area becomes an important question. Previous human cognitive studies often rely on imaging techniques that have low temporal resolution (Haxby et al., 2000; McCandliss et al., 2003; Price and Devlin, 2003), such as fMRI, which makes it difficult to identify multiple stages of processing within the time scale of several hundred milliseconds. On the other hand, previous human electrophysiology studies mainly relied on univariate statistics on the signature peaks in the event related potentials (ERPs) (Bentin et al., 1996; Puce et al., 1999), which often does not give a full picture about the dynamic timecourse. As a result, it remains unclear what dynamic roles the category-selective regions play during the process of visual stimuli, such as faces and words.

To investigate the spatiotemporal dynamics, we seek for a signal modality that satisfies the following requirements: (1) a neural signal modality with high temporal resolution (e.g. on the order of millisecond); (2) fine spatial resolution is also necessary in order to localize the category-selective region of interest with high accuracy; (3) probing the dynamic neural coding would also require high signal-to-noise ratio (SNR). Despite the magnificent breakthroughs over the past few decades, these requirements on spatiotemporal resolution and SNR have exceeded the limitations of non-invasive imaging modalities, such as fMRI, MEG and scalp EEG. On the contrary, intracranial EEG (iEEG) recording, as an alternative technique, fulfills these requirements and serves as the ideal signal modality for investigating the spatiotemporal dynamics of the neural activity in the category-selective visual areas. In this thesis, we collect iEEG data from a large cohort of human patients and study the recognition process of two representative categories of visual objects, words and faces.

The first part of this thesis evolves from the two basic questions mentioned earlier in this chapter, probing the spatiotemporal dynamics of the neural activity in cateogory-selective areas from the following aspects:

Apply multivariate pattern analysis (MVPA) methods to the iEEG signal to analyze the dynamic roles of the face and word sensitive areas during the face and word recognition processes.

- In Chapter 2, we apply pattern classification method to elucidate the dynamic role the left midfusiform gyrus (lmFG) plays with an early processing stage organized by orthographic similarity and a later stage supporting individuation of single words. Furthermore, we utilized direct cortical stimulation to demonstrate a causal role of lmFG in word naming. This study try to resolve a central issue in the neurobiology of reading, which is a debate regarding the visual representation of words, particularly in lmFG.
- In Chapter 3, we investigate the dynamic role of the face sensitive patches in the fusiform gyrus plays during the perception of face category and individual faces. We demonstrate a similar gist-to-fine temporal dynamics in the face sensitive fusiform gyrus.
- In Chapter 4, we use similar approaches to further study the role that fusiform plays in the perception of emotional facial expressions. We directly test the competing hypotheses about whether fusiform contributes to the processing of facial expressions. Our results suggest a dissociation between structural and motion in the face processing streams.

1.2.2 The representational structure of the neural interactions

The second gap comes from the fact that the classical framework is based upon the idea of passive filtering. Recent studies demonstrate that neural populations in perceptual areas alter their response properties based on context, task demands, etc. (Gilbert and Li, 2013). These modulations of response properties suggest that lateral and long-distance interactions are adaptive and dynamic processes responsive to the type of information being processed. However, not only the classical framework does not account for such adaptive process, we also lack an analytical tool to probe such information presentations. With respect to the first basic question of neural coding within a certain population, pattern classification methods from modern statistics and machine learning, such as MVPA (Haxby et al., 2001, 2014), have gained popularity in recent years for decoding the information content contained in neuroimaging data analysis. These methods allow one to go beyond examining the involvement of a population in a particular neural process and infer the representational content of the population activity. However, when we turn to the second basic question and focus on the information represented through interactions between areas, current MVPA methods do not allow one to assess the discriminant information encoded in the pattern of functional connections between different neural populations. Furthermore, traditional methods for assessing functional connectivity only allow one to examine differences in the degree of coupling across conditions and not the information carried by the pattern of interregional connections (Coutanche and Thompson-Schill, 2013; Finn et al., 2015; Kriegeskorte and Kievit, 2013; Richiardi et al., 2011; Rosenberg et al., 2016; Shirer et al., 2012; Wang et al., 2015). Therefore, we seek for a novel method to decode the representation structure of the neural interactions between populations.

- In Chapter 5, we introduce a novel method termed Multi-Connection Pattern Analysis (MCPA) to extract the discriminant information about cognitive states solely from the shared activity between neural populations from two brain areas. With this new tool, we can perform single-trial prediction and classification, and probe the represetntational structure of the interactions between areas of interest. Specifically, MCPA is applied to iEEG and fMRI data recorded from interacting regions in the visual cortex to evaluate the information representation in the pattern of interactions between areas.

1.2.3 State-dependence of neural coding

The third gap in the classical framework comes from the fact that brain does not act in a deterministic manner and neural activations are state-dependent. Even identical repetitions of the exact input stimulus would result in different neural activation, which would ultimately lead to variance in the behavioral domain, such as reaction time, sensory perception, etc. In the last part of the thesis, we study the space of the state-dependent dynamics in neural activity from the ventral visual pathway.

An important source of the variation in the neural dynamics is the spontaneous ongoing activity. Previous studies have demonstrated that both the post-stimulus evoked response (Arieli et al., 1996; Başar, 1980; Brandt and Jansen, 1991; Fox et al., 2006; Fries et al., 2001; Henriksson et al., 2015; Kisley and Gerstein, 1999; Luczak et al., 2009; Tsodyks et al., 1999) and the performance of sensory perception (Busch et al., 2009; Ergenoglu et al., 2004; Mathewson et al.,

2009; Ress et al., 2000; Thut et al., 2006; Van Dijk et al., 2008; VanRullen et al., 2011) depend on the internal brain state before stimulus presentation. However, most of the previous studies have explored these two relationships separately. To influence perceptual behavior, the variance in spontaneous ongoing activity should modulate the discriminant neural coding that directly relates to visual perception. It is yet unclear whether the endogenous activity modulates neural coding and category-selectivity in the ventral stream, which then provides a neural pathway for behavioral modulation. Therefore, we need to analyze how the pre-stimulus activity modulate the categorical sensitivity in the evoked response on a single-trial basis.

- In Chapter 6, we build a block-wise generalized linear model (GLM) to study the correlation between the categorical sensitivity and the pre-stimulus neural activity in the same brain area. We use this GLM to test the hypothesis that pre-stimulus spontaneous activity can modulate the categorical sensitivity in post-stimulus evoked response and that the same aspects of pre-stimulus activity that modulate tuning also correlate to variance in the perceptual behavior.

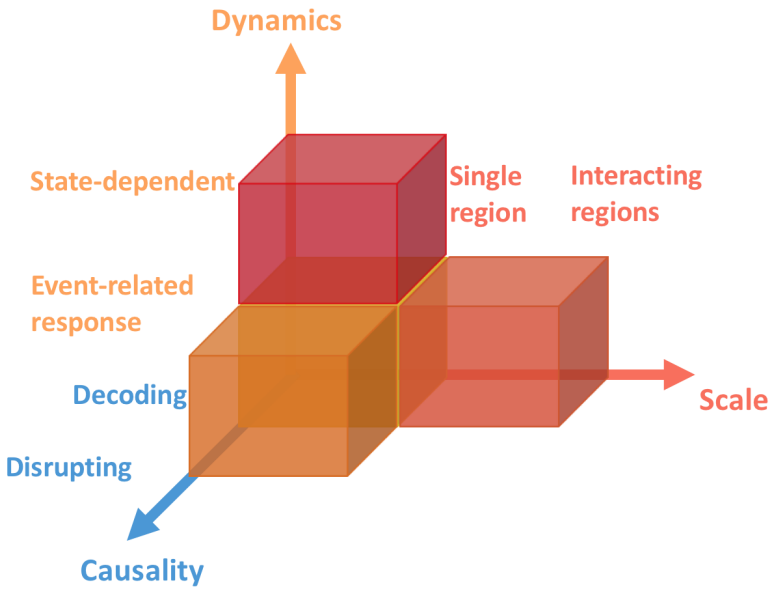


Figure 1.2: **Probing the dynamics and interactions in the visual system along different methodological dimensions.**

1.2.4 Methodological summary

From the methodological point of view, we elaborate around the multivariate representational space and explore along different methodological dimensions in order to attain a comprehensive understanding of the aforementioned questions with regard to the spatiotemporal dynamics and interactions underlying visual perception (Figure 1.2):

- The basic methodological building block that we employ is decoding the dynamics of neural representation using event-related response from a single region of interest (ROI) (Chapters 2, 3, 4).

- Moving along the dimension of causality, we use electrical stimulation to make causal inference about the neural computation taking place within certain ROI (Chapter 2, and see also (Aminoff et al., 2016)).
- Moving from a single region to interacting regions, we develop new algorithm to probe the representation of neural communications between interacting regions (Chapter 5).
- By involving the spontaneous activity into the decoding model, we evaluate the state-dependency in the neural dynamics of each ROI (Chapter 6).

We hope this thesis not only addresses the gaps in the classical model of the visual hierarchy and sheds light upon the underlying information processing dynamics of visual perception, but also demonstrates how novel applications of statistical machine learning techniques can allow cognitive neuroscientists to ask fine-grained questions about neural information processing and information flow at both the scale of local brain regions and the scale of broadly distributed networks.

Related publications

- Chapter 2 - Decoding and disrupting left mid-fusiform gyrus activity during word reading (*PNAS* 2016).
- Chapter 3 - Dynamic encoding of face information in the human fusiform gyrus (*Nature Communications* 2014).
- Chapter 4 - Posterior Fusiform and Midfusiform Contribute to Distinct Stages of Facial Expression Processing (*Cerebral Cortex* 2018).
- Chapter 5 - Multi-Connection Pattern Analysis: decoding the representational content of neural communication (*Neuroimage* 2017).
- Chapter 6 - Endogenous activity modulates category tuning in cortex and influence perceptual behavior (Work presented at VSS 2018, CCN 2018 and SfN 2018, manuscript in preparation for submission).

Chapter 2

Temporal dynamics in human fusiform underlying word individuation

In the first part of the thesis, we mainly explore the dynamics of the neural representation in two neighboring category-selective areas in the ventral pathway, the visual word form area and the fusiform face area. These two areas show strong selectivity with respect to two important visual categories in daily life, the visual words and the faces. We start with the analysis of visual word form area in this chapter.

The nature of the visual representation for words has been fiercely debated for over 150 years. In this chapter, we used direct brain stimulation, pre- and post-surgical behavioral measures, and intracranial electroencephalography to provide support for, and elaborate upon, the visual word form hypothesis. This hypothesis states that activity in the left mid-fusiform gyrus (lmFG) reflects visually organized information about words and word-parts. Using machine learning methods to analyze the temporal dynamics of electrophysiological data from four patients with electrodes placed directly in their lmFG, we found that information contained in early lmFG activity was consistent with an orthographic similarity space. Furthermore, disrupting lmFG activity through stimulation or surgical resection led to impaired perception of whole words and word-parts. Finally, the lmFG contributed to at least two distinguishable stages of word processing, an early stage that reflects gist-level visual representation sensitive to orthographic statistics, and a later stage that reflects more precise representation sufficient for the individuation of orthographic word forms. These results provide strong support for the visual word form hypothesis and demonstrate the dynamic role the lmFG plays in multiple stages of orthographic representation.

2.1 Introduction

A central debate in understanding how we read, documented at least as far back as Charcot, Dejerine, and Wernicke, has revolved around whether or not visual representations of words can be found in the brain. Specifically, Charcot and Dejerine posited the existence of a center for the visual memory of words (Bub et al., 1993), whereas Wernicke firmly rejected that notion, proposing that reading only necessitates representations of visual letters that feed forward into

the language system (Wernicke, 1977). Similarly, the modern debate revolves around whether or not there is a visual word form system that becomes specialized for the representation of orthographic knowledge (e.g. the visual forms of letter combinations, morphemes, and whole words; (Bub et al., 1993; Dehaene et al., 2002; Warrington and Shallice, 1980). One side of the debate is characterized by the view that the brain possesses a visual word form area that is a major, reproducible site of orthographic knowledge (Dehaene and Cohen, 2011), while the other side disavows any need for reading-specific visual specialization, arguing instead for neurons that are general purpose analyzers of visual forms (Price and Devlin, 2011).

The visual word form hypothesis has attracted great scrutiny because the historical novelty of reading makes it highly unlikely that evolution has created a brain system specialized for reading. This places the analysis of visual word forms in stark contrast to other processes that are thought to have specialized neural systems, such as social, verbal language, or emotional processes, which can be seen in our evolutionary ancestors. Thus, testing the word form hypothesis is critical not only for understanding the neural basis of reading, but also for understanding how the brain organizes information that must be learned through extensive experience and for which we have no evolutionary bias.

Advances in neuroimaging and lesion mapping have refocused the modern debate surrounding the visual word form hypothesis on the left mid-fusiform gyrus (lmFG). This focus reflects widespread agreement that the lmFG region plays a critical role in reading. Supporting evidence includes demonstrations that literacy shapes the functional specialization of the lmFG in children and adults (Ben-Shachar et al., 2011; Brem et al., 2010; Dehaene et al., 2010; Schlaggar and McCandliss, 2007), the lmFG is affected by orthographic training in adults (Glezer et al., 2015; Xue and Poldrack, 2007), and damage to the lmFG impairs visual word identification in literate adults (Behrmann and Shallice, 1995; Gaillard et al., 2006). However, debate remains about whether the lmFG constitutes a visual word form area (Binder et al., 2006; Cohen et al., 2002; Dehaene and Cohen, 2011; Glezer et al., 2009; McCandliss et al., 2003; Warrington and Shallice, 1980) or not (Farah and Wallace, 1991; Price and Devlin, 2003, 2011). That is: does it support the representation of orthographic knowledge about graphemes, their combinatorial statistics, orthographic similarities between words, and word identity (Vinckier et al., 2007), or does it have receptive properties tuned for general purpose visual analysis, with lexical knowledge emerging from the spoken language network (Price and Devlin, 2011)?

To test the limits of the modern visual word form hypothesis, we present results from four neurosurgical patients (P1-4) with electrodes implanted in their lmFG. We acquired pre and post surgery neuropsychological data in P1, performed direct cortical stimulation in P1 and P2, and recorded intracranial electroencephalography (iEEG) in all four participants to examine a number of indicators that have been proposed as tests for the visual word form hypothesis by both supporters and opponents of this hypothesis (Dehaene and Cohen, 2011; Price and Devlin, 2011). Pattern classification methods from machine learning were then used to measure whether neural coding in this region is sufficient to represent different aspects of orthographic knowledge, including the identity of a printed word. We separately evaluated the timecourse of lmFG sensitivity to different aspects of orthographic information to assess both early processing, which should exclusively or predominantly capture bottom-up visual processing, and later processing, which likely captures feedback and recurrent interactions with higher-level visual and non-visual regions. Consequently, we were able to assess the dynamic nature of orthographic representation

within the lmFG and thereby provide a novel perspective on the nature of visual word representation in the brain.

2.2 Methods

2.2.1 Subjects

Four patients (2 males, ages 25-45) undergoing surgical treatment for medicine-resistant epilepsy participated in the experiments. The patients gave written informed consent to participate in this study, under a protocol approved by the University of Pittsburgh Medical Center Institutional Review Board. See supplement for demographic and clinical information about each participant.

2.2.2 Experimental paradigm

The experiment paradigm and the data pre-processing method were similar to those described previously by Ghuman and colleagues (Ghuman et al., 2014). Paradigms were programmed in MATLAB using Psychtoolbox and custom written code. All stimuli for the Category Localizer, Covert Naming, Word Individuation and Stimulation were presented on a 22-inch LCD computer screen placed approximately 2 meters from participant's head at the center of the screen ($\sim 10^\circ \times 10^\circ$ of visual angle). All stimuli for P1-P3 were identical. Due to a considerable delay in testing, the covert naming and word individuation stimuli were modified and updated for P4 in order to address additional questions beyond the scope of the current study. However, the critical characteristics of the stimuli and contrasts in the analyses remain consistent across all four patients. The category localizer was identical for all patients.

2.2.3 Category localizer

Stimuli

In the localizer experiment, 90 different images from 3 categories were used, with 30 images of bodies (50% male), 30 images of words, and 30 phase scrambled images. Phase scrambled images were created in MATLABTM by taking the 2-dimensional Fourier transform of the image, extracting the phase, adding random phases, recombining the phase and amplitude, and taking the 2-dimensional inverse Fourier transform.

Design and procedure

In the category localizer, each image was presented for 900 ms with 900 ms inter-trial interval, during which a fixation cross was presented at the center of the screen. There were two consecutive blocks in a session. Each block consisted of all the 180 images with a random presenting order. At random, 1/3 of the time an image would be repeated, which yielded a total of 480 trials in one recording session. The participant was instructed to press a button on a button box when an image was repeated (1-back task).

2.2.4 Electrical brain stimulation

Stimuli

The stimuli used during electrode stimulation for P1 included 60 7-letter words with 11.35 (10.60-13.67) mean log frequency, determined by the HAL Study used in the English Lexicon project (`http://elexicon.wustl.edu/`); single letters; and 13 famous faces that were familiar and nameable by P1. Stimuli were presented repeatedly during the session, starting with low stimulation trials. Thus, stimuli presented during high stimulation trials were likely to have been seen previously. The stimuli used during electrode stimulation for P2 included 46 7-letter words with 10.93 (10.02-13.13) mean log frequency, and black and white pictures of common objects and animals. The 46 words that were presented during stimulation trials were out of a set of 155 words total that did not repeat.

Design and procedure

Electrical current during stimulation passed between adjacent electrode pairs (e.g., 1 & 2, 3 & 4, etc.). During the stimulation session pre-surgery, stimulation (1-10 mA, peak-to-peak amplitude, which is the distance between the negative and positive square waves delivered to the two contacts, i.e. this is 2 times the amplitude of the square waves) was alternatingly applied with sham stimulation while P1 and P2 overtly named words (P1 and P2), letters (P1), famous faces (P1), and pictures (P2). Each stimulus trial began with a beep, followed by 750 ms of fixation and then the stimulus. The stimulus remained on the screen until it was named, after which an experimenter manually advanced to the next item. Naming times were computed by calculating the time between the beep and the response (minus 750 ms). Only trials in which the electrode stimulation overlapped with the first 500ms of stimulus presentation were included in further statistical analyses. T-tests comparing high and low stimulation trials were computed assuming unequal variances and df adjusted based on Levene's test for equality of variances.

2.2.5 Covert Naming: Sensitivity to Bigram Frequency

Stimuli

In the covert word-naming experiment, words with non-overlapping high and low bigram frequency (70 each for P1, 40 each for P4), controlled for lexical frequency, were used as visual stimuli.

Design and Procedure

In the covert word-naming experiment, each word was presented once, in a random order, for 3000 ms with 1000 ms inter-trial interval during which a fixation cross was presented at the center of the screen. The patient was instructed to press a button the moment when he began to covertly name the word to himself in order to ensure phonological encoding of each word and to avoid potential movement artifacts that could result from overt articulation.

2.2.6 Word Individuation

Stimuli

In the word individuation experiment, 20 different English words, with word length ranging from 2 to 5, were used as visual stimuli. Similar word pairs differed by one letter and different word pairs did not share any letters. All comparisons were made within the same word length.

Design and Procedure

In the word individuation experiment, each image was presented for 900 ms with 900 ms inter-trial interval, during which time a fixation cross was presented at the center of the screen. There were 24 consecutive blocks within a session. Each block consisted of all the 20 words with a random order. At random, 1/6 of the time an image would be repeated, which yielded a total of 560 trials in one session. The patient was instructed to press a button on a button box when an image was repeated.

2.2.7 Data Preprocessing

Local field potential (LFP) Data for the Category Localizer, Covert Naming, and Word Individuation tasks were collected at 1000 Hz using a Grapevine neural interface system (Ripple, LLC). They were subsequently band-pass filtered offline from 1-115 Hz and notch filtered from 59-61 Hz, both using fifth order Butterworth filters in MATLAB, to remove slow and linear drift, the line noise, and high frequency noise. Raw data was inspected for ictal events and none were found during experimental recordings. To further reduce potential artifacts in the data, trials with peak amplitude 5 standard deviations away from the mean across the rest of the trials or with absolute peak amplitude larger than 350 μV were eliminated. In addition, trials with a difference larger than 25 μV between consecutive sampling points were eliminated. These criteria resulted in the elimination of less than 1% of trials in each session.

2.2.8 Electrode Selection

Word sensitive electrodes were chosen based on anatomical and functional considerations. Electrodes of interest were restricted to those that were located on the fusiform gyrus. In addition, electrodes were selected such that their peak 3-way classification d' score (see below for how this was calculated) exceeded 1 ($p < 0.001$ based on a permutation test, as described below) and the event related potential (ERP) for words was larger than the ERP for the other non-words object categories, namely bodies and phase scrambled images.

P1, P2 and P3 each had 8 electrode contacts on a single strip on the ventral temporal lobe. P4 had 28 electrode contacts on two high-density strips on the ventral temporal lobe. For P1, out of the 8 electrode contacts, only the first three channels satisfied the criteria described above and all analyses included data from all 3 of these electrode channels. The remaining five channels failed to satisfy either of the two criteria. For P2, 3 out of the 8 electrode channels (channels 1, 3 and 4) satisfied the criteria. Only channels 3 and 4 were used for all analyses because channel 1 was non-contiguous with the other channels and more medial than would be expected for

word sensitive lmFG. For P3, out of the 8 electrode channels, only one channel (the third electrode channel on the strip) satisfied the criteria. Hence we used the data from this one electrode channel for the multivariate classification analysis. For P4, 3 out of the 28 high-density ventral temporal electrode channels satisfied the two criteria (channels 8, 9 and 22) and all analyses included data from all 3 of these channels. The precise locations varied slightly, which is a typical characteristic of word-selective cortex described in the literature (Glezer and Riesenhuber, 2013). All patients' postoperative structural MRIs were normalized to Talairach space using AFNI's auto_tlrc program to confirm the location of the word-selective contacts in the fusiform gyrus (Figure 2.1).

2.2.9 Multivariate classification analysis

Considering that the size of the training set was smaller than the data dimensionality, a low-variance classifier (specifically, Gaussian naïve Bayes) was used. Principle component analysis (PCA) and linear discriminant analysis (LDA) were used to lower the dimensions in the case of multi-way categorical classifications. However, we found the dimensionality reduction method was not plausible in the pair-wise words classification case, because the smaller number of trials made the estimation of covariance unreliable. For all classification analyses, the Gaussian naïve Bayes classifier was trained based on the data from each time point of 100 ms windows from single trials in the training set (the time course pattern from 100 ms of single trial potentials) and was used to label the condition of the corresponding data from that time window from the testing trial. The classification accuracy was estimated by counting the correctly labeled trials. This procedure was then repeated for all time windows slid with 10 ms steps between $-100 \sim 600$ ms relative to the presentation of the stimuli. A recent study showed that combining single trial potentials with the broadband signal improves classification accuracy in iEEG. In P1 and P4, we found results consistent with this report showing increased classification accuracy in the range of $\sim 1\%$ at the peak timepoints (no statistical changes were seen in that all effects reported as significant remain significant and all effects reported as not significant remain not significant). In P3 however, the broadband signal was flat despite clear single trial potential effects. Thus, for the sake of consistency across subjects, and because the results did not substantively change in P1 or P4, we use only single trial potentials throughout.

For the multi-way categorical classifications with K categories (here $K = 2, 3$), the classification accuracy was estimated through nested leave-p-out cross-validation. In the first level of cross-validation, single-trial potentials were first split into training (80% of the trials) and testing set (20% of the trials) randomly. For each random split, PCA was trained based on the training set to lower the dimensionality down to P. Then LDA was used to project the data into K-1 dimensional space. Finally a Gaussian naïve Bayes classifier was trained based on the projected training set. The selection of the model parameter P was achieved by finding the P that gave greatest d' for Bayes classification based on an additional level of random sub-sampling validation with 50 repeats using only the training set. After training, true positive and false alarm rates of the target condition were calculated across all of the test trials. d' was calculated as $d' = \Phi^{-1}(\text{true positive rate}) - \Phi^{-1}(\text{false alarm rate})$, where $\Phi^{-1}(x)$ is the inverse of the Gaussian cumulative distribution function. The random split was repeated for 200 times and the classification accuracy was estimated by averaging across results from these 200 random splits.

For the pair-wise classification in the word individuation task, the pairwise classification accuracy was estimated through leave-one-out cross-validation. Specifically, for each pair of words, each trial was left out in turn as the testing trial, with the remaining trials used for the training set. Finally, the overall pairwise classification accuracy was estimated through averaging across all 190 word-pairs. The classification accuracy for each specifically controlled condition was estimated by averaging the corresponding word-pairs.

2.2.10 Permutation test

Permutation testing was used to assess the statistical significance of classification accuracy and the corresponding d' value against the chance level for all the classification analyses described above. Specifically, the null hypothesis could be stated as that the peak classification accuracy was at chance level, using a global null hypothesis over the entire time course. This results in significance values corrected for multiple time comparisons (Maris and Oostenveld, 2007). For each permutation, the condition labels of all the trials were randomly permuted, and the same classification procedure as described above was performed on the data with permuted labels. The maximum classification accuracy across the 100 - 600 ms time window was then extracted as the test statistic. The permutation procedure was repeated for N times ($N = 200$ or 500, N is chosen heuristically based on the computational complexity of the problem and the accuracy of estimation that is needed). The estimated p-value of the classification accuracy, corrected for multiple comparisons, was then determined based on the distribution that results from the permutation procedure.

Notably, the classification accuracy reported is generally greater than what is found using non-invasive measures of neural activity, such as fMRI (Nestor et al., 2011). Nonetheless, the fact that iEEG pools over the activity of hundreds of thousands of neurons likely means that finer scale recordings, such as recording simultaneously from many single neurons, may have improved classification accuracy.

It is also notable that a recent study showed that combining both the single trial potentials, as we did here, and the broadband signal results in higher classification accuracy than either of those signals alone (Miller et al., 2016). In our case, P3 showed clean single trial potential data, but poor quality broadband data. For that reason, we chose to use the single trial potential data for all of our analyses. That said, in P1 and P4, the classification accuracy for single words did improve when combining single trial potentials and the broadband signal, as predicted by Miller et al. (2016). However, the classification accuracy improvement was quantitative and none of hypothesis testing (e.g. what was and was not significant at the $p < 0.05$ level), between time-course comparisons, or indeed, none of the conclusions from the results change when combining broadband and single trial potential data.

2.3 Results

2.3.1 Verification of orthographic selectivity at lmFG electrode sites

In order to identify their seizure foci, four patients with medically intractable epilepsy, underwent iEEG, which included insertion of multi-contact electrodes into or on their ventral temporal cortex (VT) (Figure 2.1). To assess the word sensitivity and specificity of lmFG, we used a Gaussian naïve Bayes classifier to decode the neural activity (single trial potentials) while participants viewed three different categories of visual stimuli: words, bodies, and phase-scrambled objects (30 images per category, each repeated once). In each patient in electrode contacts in lmFG, we observed a strong early sensitivity to words at 100 ms 400 ms (Figure 2.2a, 2.2b), which was verified using a classifier model (Figure 2.2c; averaged peak $d' = 1.26$, at 245 ms after stimulus onset, $p < 0.001$; see Supplement Figure S2-5 for each individual contact on the electrodes from each participant). The position of the lmFG electrode contacts in the anterior end of the posterior fusiform sulcus is consistent with the putative "visual word form area" described in the functional neuroimaging literature (Baeck et al., 2015; Wandell, 2011; Whaley et al., 2016). Further, the timing of the category selective response is consistent with evoked potential findings obtained from scalp electrodes (Maurer et al., 2005) and previous iEEG studies (Hamamé et al., 2013, 2014; Nobre et al., 1994; Whaley et al., 2016), which have described orthographic-specific effects approximately 200 ms after stimulus onset.

After completion of the iEEG study, in P1 a focal resection in the posterior basal temporal lobe was performed. This included removal of tissue at the location of the implanted VT electrode (Supplemental Figure 6), leading us to predict that P1 would exhibit post-surgical changes in visual word recognition consistent with acquired alexia (Gaillard et al., 2006). Neuropsychological assessments of naming times were conducted pre- and post-surgery at 1.5-weeks (acute), 6-weeks, and 3-months to assess the impact of the resection on his perception of visual stimuli. P1 was asked to name words (3, 5, or 7 letters (Behrmann and Shallice, 1995)) and a mixed set of stimuli (words, letters, single digits, 3-digit numbers, famous faces, objects, music notes, and guitar tabs) aloud as rapidly and accurately as possible. After removal of the area surrounding the VT electrode, P1 showed the characteristics of acquired alexia, specifically letter-by-letter reading (Figure 2.3c) and longer naming times particularly for letters and words (Figure 3d) as predicted based on the role of this area in orthographic processing (Behrmann and Shallice, 1995; Gaillard et al., 2006). Additionally, orthographic processes were impacted to a greater degree than phonological processes by the resection (Supplemental Figure 1). See Supplemental Results for further description and elaboration on P1's post-resection reading deficits.

The anatomical locus and category-specificity of the recorded iEEG response in P1-P4, and the post-resection alexia in P1, were highly consistent with our localization of lmFG electrodes to tissue that is central to the visual word form debate. We then tested specific putative indicators of the visual word form hypothesis using data obtained from cortical stimulation (P1 & P2) and iEEG (P1, P3 & P4) from these electrode sites.

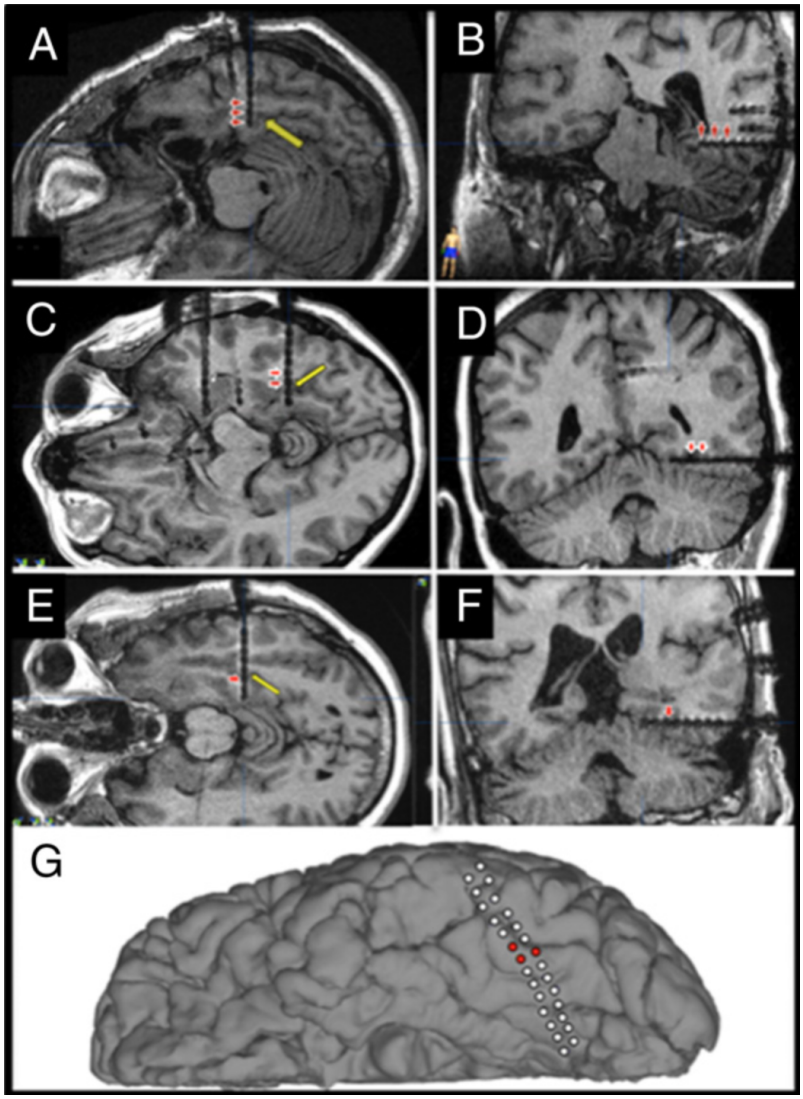


Figure 2.1: **Location of implanted electrodes. Individual electrode contacts are visible on axial (A, C, E, G) and coronal (B, D, F) views of the post-implantation MRI** (P1: A-B; P2: C-D; P3: E-F; P4: G). The VT depth electrodes were placed at the anterior end of the mid-fusiform sulcus in P1-P3 (yellow arrow), and P4 was implanted with a left temporal subdural grid crossing the lmFG. Red arrowheads (A-F) and red filled circles (G) indicate the word-selective contacts identified in the category localizer, which were used in subsequent electrophysiological and/or stimulation experiments. Talairach coordinates corresponding to the word-selective contacts were located in post-operative MRI structural images, and were all identified in the left fusiform gyrus, BA 37 (P1 electrodes: -31, -36, -13; -35, -37, -13; -39, -38, -12; P2 electrodes: -30, -46, -11; -34, -46, -12; P3 electrode: -31, -35, -14; P4 electrodes: -38, -51, -21 ; -41, -50, -22; -41, -54, -20).

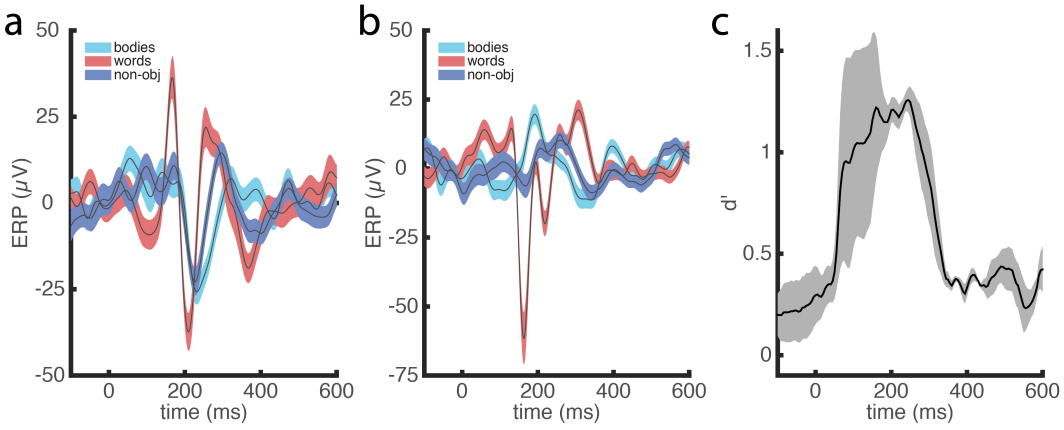


Figure 2.2: **Verification of orthographic selectivity at lmFG electrode site. a)** Example of averaged event related potential (ERP) across lmFG electrodes in one of the participants (P1) for three different stimulus categories (bodies, words and non-objects). The colored areas indicate standard errors. **b)** Averaged event related potential (ERP) across all lmFG electrodes and across all of the participants for three different stimulus categories (bodies, words and non-objects). The colored areas indicate standard errors. **c)** Time course of word categorical sensitivity in lmFG electrodes measured by sensitivity index d' (mean d' plotted against the beginning of the 100 ms sliding window), averaged across three participants. The MTPA classifier uses time-windowed single-trial potential signal from the electrodes from each subject (window length = 100 ms) with each time point in the window from each electrode as multivariate input features (see Methods for details). Across-participant standard errors are shaded grey. See Supplemental Figure 2-5 for single electrode word categorical sensitivity.

2.3.2 Disrupting lmFG Activity Impairs Both Lexical and Sublexical Orthographic Processing

One indicator of whether the lmFG functions as a specialized visual word form system is whether disrupting its activity using electrical stimulation impairs the normal perception of both printed words and sublexical orthographic components (Hamamé et al., 2013; Nobre et al., 1994), but not other kinds of visual stimuli. As part of pre-surgical language mapping, P1 and P2 underwent an electrical stimulation session where they named two kinds of orthographic stimuli (words [P1 & P2] and letters [P1]), as well non-orthographic objects (faces [P1] and pictures [P2]). We hypothesized that high stimulation (6-10 mA) to the lmFG electrodes would cause greater disruption to reading orthographic stimuli than low stimulation (1-5 mA) due to the observed category-specificity of the iEEG response, but no disruption would be seen for stimulation during object (face or picture) naming. Indeed, P1 and P2 were significantly slower at reading words at high stimulation than low stimulation (Figure 2.3a,2.3b; P1: Mean $\text{RT}_{\text{low stim}} = 967$ms, Mean $\text{RT}_{\text{high stim}} = 1860$ ms, $t(18) = 2.42$, Cohen's $d = 1.14$, $p = 0.026$; P2: Mean $\text{RT}_{\text{low stim}} = 1586$ms, Mean $\text{RT}_{\text{high stim}} = 8700$ms, $t(7) = 11.28$, Cohen's $d = 5.15$, $p < .001$). P1 also misidentified 5% of words (naming 'number' as 'nature') under high stimulation on the lmFG electrodes. P2 did not misidentify any words, but was generally unable to name words until the

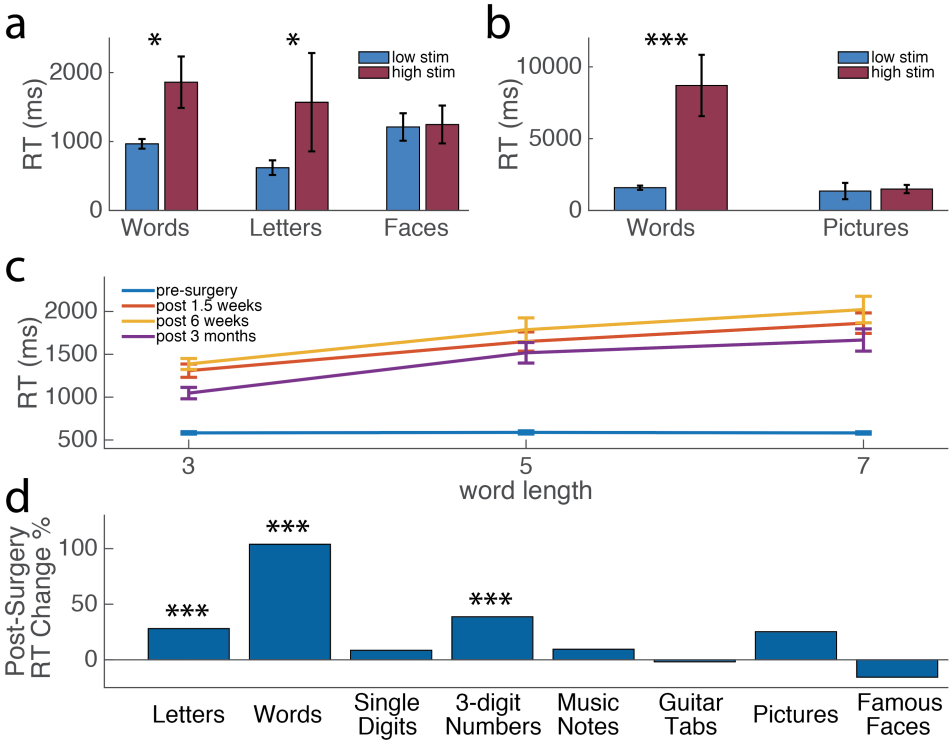


Figure 2.3: **The effect of stimulation on naming times in lmFG and pre and post-surgery neuropsychological naming task performance.** **a)** The average naming reaction time for words, letters, and faces under low stimulation (1-5 mA) and high stimulation (6-10 mA) to lmFG electrodes in P1. Error bars correspond to standard error, * $p < 0.05$. **b)** The average naming reaction time for words and pictures under low stimulation (1-5 mA) and high stimulation (6-10 mA) to lmFG electrodes in P2. Error bars correspond to standard error, *** $p < 0.001$. **c)** Word length effect pre- and post-surgery in P1. **d)** Average percent change in reaction time in the Mixed Naming Task Pre vs. Post-surgery in P1, *** $p < 0.001$.

stimulation had ceased. Her self-report suggested an orthographic disruption rather than speech arrest. Specifically, for the word 'illegal' she reported thinking two different words at the same time, and trying to combine them. For the word 'message', she reported thinking that there was an 'n' in the word (see SI Video 2). P1 was also asked to name single letters during stimulation in lmFG electrodes. With limited letter trials during stimulation (two low stimulation and five high stimulation), there was no significant difference in reaction time in letter naming between high and low stimulation. However, P1 responded incorrectly to two letter stimuli, initially responding 'A' for 'X', and responding 'F' and then 'H' to the visual stimulus 'C', both of which he had previously named accurately during the stimulation session (see SI Video 1). Importantly, naming times for non-orthographic stimuli were not significantly affected by stimulation in lmFG electrodes (P1, faces: Mean $\text{RT}_{\text{low stim}} = 1211$ ms, Mean $\text{RT}_{\text{high stim}} = 1246$ ms, $t(12) = 0.11$, Cohen's $d = 0.05$, $p = 0.92$; P2, pictures: Mean $\text{RT}_{\text{low stim}} = 1350$ ms, Mean $\text{RT}_{\text{high stim}} = 1490$ ms, $t(10) = 0.18$, Cohen's $d = 0.13$, $p = 0.86$). These results are consistent with previous

reports of selective impairments due to stimulation in the lmFG for reading orthographic stimuli (Mani et al., 2008). Notably, the category specific perceptual alteration seen in P1 and P2 shows similar feature-level distortions of identity as has been reported faces when stimulating right mFG (Parvizi et al., 2012). These stimulation results indicate that disruption of lmFG function impairs both the skilled identification of visual words and sublexical components of word forms (i.e., letters), supportive of the visual word form hypothesis.

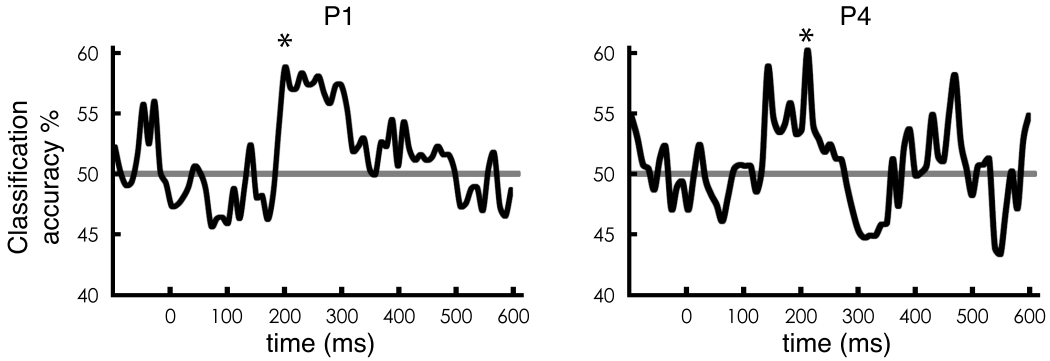


Figure 2.4: **Dynamics of sensitivity to sublexical orthographic statistics (bigram frequency) in the lmFG.** Classification accuracy timecourse for comparison between low bigram frequency real words (low BG) vs. high bigram frequency real words (high BG) in lmFG electrodes for P1 and P4 respectively, plotted against the beginning of the 100 ms sliding window. The classifier uses time-windowed single-trial potential signal from the electrodes from each subject (window length = 100 ms) with each time point in the window from each electrode as multivariate input features (See Methods for details). The * corresponds to the peak of the windows in which $p < 0.05$ corrected for multiple comparisons. The $p = 0.05$ significance threshold corresponds to accuracy = 58.2% (P1), and 59.3% (P4). The horizontal grey line at 50% indicates chance level.

2.3.3 Electrophysiological Evidence for a Visual Word Form Representation in the lmFG

We next used techniques from machine learning in iEEG data from P1 and P4 to assess the sensitivity of lmFG to sublexical, orthographic statistics (bigram frequency) that has been hypothesized as an indicator for a visual word form system (Binder et al., 2006; Vinckier et al., 2007). To examine the dynamics of orthographic statistic sensitivity, we used a multivariate temporal pattern analysis (MTPA) classification procedure to test how the lmFG represents aspects of orthographic knowledge critical to the word form hypothesis at different stages of the timecourse.

In order to measure sublexical sensitivity as a test of the word form hypothesis, P1 and P4 performed a covert naming task with high and low bigram frequency words, controlled for lexical frequency. The MTPA classifier was sensitive to differences between high and low bigram frequency during a relatively early time window in both participants (Figure 2.4; P1: peak accuracy

= 58.6%, $p < 0.05$ at 200-330 ms after stimulus onset; P4: peak accuracy = 60.2%, $p < 0.05$ at 210-310 ms after stimulus onset; all classification analyses were tested using permutation tests to correct for multiple comparisons). This finding is consistent with early discrimination between words and pseudowords in Kanji, which differ in sublexical statistical properties (i.e., likelihood of a particular character preceding another) in the basal temporal cortex (Tanji et al., 2005). It has been noted that testing the visual word form hypothesis requires examining the representation in lmFG that results primarily from feed-forward input from earlier parts of the ventral visual processing stream (Dehaene and Cohen, 2011). Thus, the result that sublexical aspects of orthographic information began at a relatively early time point in processing is supportive of the word form hypothesis (Binder et al., 2006; Dehaene and Cohen, 2011; Duncan et al., 2010; Price and Devlin, 2011; Vinckier et al., 2007).

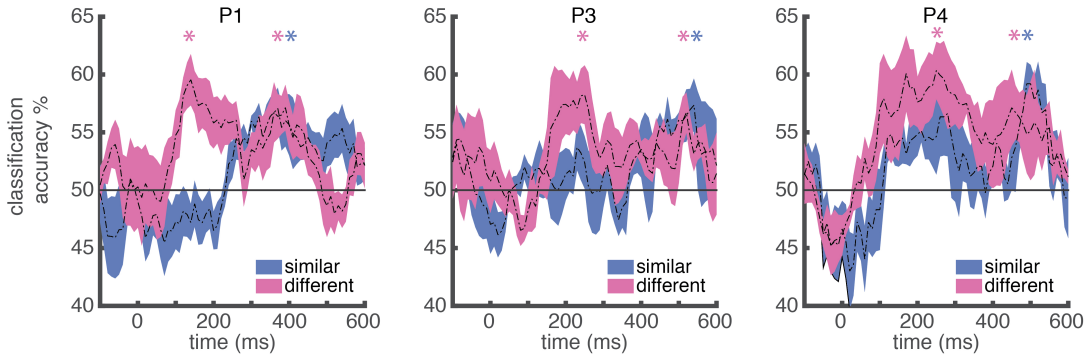


Figure 2.5: **Dynamics of word individuation selectivity in the lmFG.** Dynamics of averaged pair-wise word individuation accuracy for different conditions in lmFG electrodes for P1, P3, and P4 respectively, plotted against the beginning of the 100 ms sliding window. The classifier uses time-windowed single-trial potential signal from the electrodes from each subject (window length = 100 ms) with each time point in the window from each electrode as multivariate input features (See Methods for details). The time course of the accuracy is averaged across all word-pairs of the corresponding conditions. The colored areas indicate standard errors. Similar pair: a pair of words that have the same length and are only different in one letter, e.g. 'lint' and 'hint'. Different pair: a pair of words that have the same length and are different in all letters, e.g. 'lint' and 'dome'. Horizontal grey line indicates chance level (accuracy = 50%). Colored * corresponds to the peak of the windows in which $p < 0.05$ corrected for multiple comparisons. The $p = 0.05$ significance threshold corresponds to accuracy = 56.5% (P1), 56.0% (P3), and 57.1% (P4).

2.3.4 Temporal Dynamics of Word Individuation in lmFG

To further elucidate the dynamic nature of orthographic representation, we next looked at the sensitivity of lmFG to different aspects of individual words in P1, P3, and P4. Using words that varied in their degree of visual similarity (e.g., words that differed by one letter vs. all letters), we determined at what similarity level an MTPA classifier could discriminate between any two items. We found that at an early time window after stimulus onset, an MTPA classifier could

significantly discriminate between words that did not share any letters (e.g., 'lint' vs. 'dome'; P1: peak classification accuracy = 59.6%, $p < 0.05$ from 120-250ms; P3: peak classification accuracy = 58.3%, $p < 0.05$ from 180-360ms; P4: peak classification accuracy = 60.3%, $p < 0.05$ from 100-430ms, all p-values were corrected for multiple time comparisons; Figure 2.5), but could not discriminate between words that only differed by one letter (e.g., 'lint' vs. 'hint'; P1: peak classification accuracy = 52.7%, $p > 0.1$; P3: peak classification accuracy = 53.7%, $p > 0.1$; P4: peak classification accuracy = 56.6%, $p > 0.05$; Figure 2.5). This result demonstrates an organization governed by an orthographic similarity space at the sublexical level, a finding consistent with our observation of bigram frequency effects in a relatively early time window. However, within a later time window, an MTPA classifier could discriminate between any two words (Figure 2.5). Notably, this includes word pairs with only one letter difference (P1: peak classification accuracy = 57.1%, $p < 0.05$ from 360-470 ms; P3: peak classification accuracy = 57.3%, $p < 0.05$ from 470-640 ms; P4: peak classification accuracy = 59.2%, $p < 0.05$ from 490-620 ms).

2.4 Discussion

Our findings, which indicate that orthographic representation within the lmFG qualitatively shifts over time, provide a novel advancement on the debate about the visual word form hypothesis (Bub et al., 1993; Wernicke, 1977). Specifically, we demonstrated that lmFG meets all of the proposed criteria for a visual word form system: early activity in lmFG coded for orthographic information at the sublexical level, disrupting lmFG activity impaired both lexical and sublexical perception, and early activity reflected an orthographic similarity space. Early activity in lmFG is sufficient to support a gist-level representation of words that differentiates between words with different visual statistics (e.g., orthographic bigram frequency).

Notably, the results in the late time window suggest that orthographic representation in lmFG shifts from gist-level representations to more precise representations sufficient for the individuation of visual words. In this late window, the lmFG became nearly insensitive to orthographic similarity as shown by similar classification accuracy for word pairs that differed by one letter compared to word pairs that were completely orthographically different. This kind of unique encoding of words is required to permit the individuation of visual words, a necessary step in word recognition (see Table 2.4 for summary). The time window in which this individuation signal is seen suggests that interactions with other brain regions transform the orthographic representation within the lmFG in support of word recognition. Such interactivity could function to integrate the orthographic, phonological, and semantic knowledge that together uniquely identifies a written word (Whaley et al., 2016). Lack of spatiotemporal resolution to detect dynamic changes in lmFG coding of orthographic stimuli using fMRI may help to explain competing evidence for and against the visual word form hypothesis in the literature (Dehaene and Cohen, 2011; Price and Devlin, 2011).

The dynamic shift in the specificity of orthographic representation in the lmFG has a very similar time course as the coarse-to-fine processing shown in face sensitive regions of the human fusiform (Ghuman et al., 2014). Considering that only an gist-level representation is available until approximately 250 ms, and that saccade planning and execution generally occur within

Table 2.1: Summary of positive (✓) results in early and late time windows

	Early	Late
Word category > Objects	✓	✓
Sensitivity to bigram frequency	✓	✓
Sensitivity to lexical status		✓
Orthographic similarity space: 'lint' vs 'dome' > 'lint' vs 'hint'	✓	

200-250 ms during natural reading (Reichle et al., 1998), the gist-to-individuated word processing dynamic has important implications for neurobiological theories of reading. It suggests that when visual word form knowledge first makes contact with the language system, it is in the form of gist-level information that is insufficient to distinguish between visually similar alternatives. This is consistent with evidence that readers are vulnerable to making errors in word individuation during natural reading, but contextual constraints are normally sufficient to avoid misinterpretations (Levy et al., 2009). In other words, in most cases, accurate individuation is achieved through continued processing that likely involves mutually constraining orthographic, phonological, semantic, and contextual information resulting in a more precise individuated word representation.

Another notable pattern in the gist-to-individuation temporal dynamic is that during the later time window when individuation is significant ($\sim 300 - 500$ ms, see Figure 2.5), the power to detect category-level word selectivity (i.e., words vs. bodies and scrambled images; see Figure 2), which arguably only requires gist-level discrimination, weakens and the ERP response has waned. This is also consistent with a temporal selectivity pattern described for faces (Ghuman et al., 2014). One potential explanation for this selectivity and power shift could be that individuation is achieved by relatively few neurons (sparse coding) (Young and Yamane, 1992). Sparse coding would imply that relatively few word sensitive neurons were active and that the summed approximate word-related activity in this time period therefore would be weak. However, the neurons that were active encode for more precise word information, which would explain the significant word individuation reported here.

The mechanism underlying the representational shift from gist-to-individuation could have implications for models of reading disorders, like dyslexia, where visual word identification is impaired (Bruck, 1990). Indeed, the effects of lmFG stimulation, especially slower reading times, are suggestive of acquired (Behrmann and Shallice, 1995) and developmental reading pathologies (Bowers and Wolf, 1993), which have been linked to dysfunction of lmFG (Martin et al., 2016). The extent to which individual word reading may be impaired by excess noise in the visual word form system, or the inadequate ability to contextually constrain noisy input into the language system, is for future research to untangle.

In summary, our results provide strong evidence that the lmFG is involved in at least two temporally distinguishable processing stages: an early stage that allows for category level word decoding and gist-level representation organized by orthographic similarity and a later stage supporting precise word individuation. An unanswered question is how the representation in the lmFG transitions between stages in these local neural populations and how interactions between areas involved in reading may govern these transitions. Taken together, the current results sug-

gest a model in which lmFG contributes to multiple levels of orthographic representation, via a dynamic shift in the computational analysis of different aspects of word information.

2.5 Appendix: Supplement Methods and Results

Patient Medical History

Patient P1 was a 25-year-old right-handed man with medically intractable epilepsy since the age of 7. The clinical onset of his partial complex, secondary generalizing seizures was characterized by behavioral arrest and inability to speak. His 3-Tesla MRI was negative for any visible lesions. The patient had undergone two surgeries prior to the current one, first a partial anterior temporal lobectomy (7 years prior), and then a second surgery to complete the resection of residual mesial structures (1 year prior). The patient exhibited baseline low average to moderately impaired skills on IQ measures, but exhibited stable performance after his first and second surgeries, with the exception of a decline on a verbal memory task after the first surgery. Following further evaluation that included repeat video-EEG and magnetoencephalography, the multidisciplinary epilepsy team recommended that the patient undergo intracranial monitoring via stereo-electroencephalography (SEEG) to attempt to definitively delineate the seizure focus. Ictal intracranial EEG suggested a seizure focus in the posterior left inferior temporal gyrus. Language mapping was performed for surgical planning purposes. The epilepsy board recommended resection of the presumed seizure onset zone that included portions of the middle, inferior and fusiform gyri, to which the patient consented, after discussion of the potential risks and benefits of surgery, including the eventuality of a reading impairment. The patient had a period of seizure freedom for 10 weeks following surgery, but subsequently has continued to experience seizures.

Patient P2 was a 31-year-old female with a 4-year duration of medically intractable epilepsy, with seizures occurring several times per week. Seizures began as alteration of awareness, progressing to generalized convulsions. The patient's highest level of education is college coursework, and neuropsychological testing revealed impairment in verbal greater than visual memory, with an estimate of overall intelligence in the low-average range. A 3T MRI was normal. She underwent left frontotemporal SEEG, which revealed seizure onset in the left anterior mesial temporal lobe. She underwent a left anterior temporal lobectomy and has remained seizure free.

Patient P3 was a 41-yar-old-man with a 3-year duration of medically intractable epilepsy, with seizures occurring several times per week. Seizures began with automatisms and alteration of awareness, progressing to alterations in speech. The patient's highest level of education is a high school degree, and neuropsychological testing revealed deficits in verbal learning and memory skills, with overall intelligence estimated in the average range. A 3T MRI was normal. He underwent bilateral frontotemporal SEEG, which predominately revealed a left anterior temporal neocortical onset zone, with a possible smaller, independent focus in the right frontal operculum. He underwent a left anterior temporal lobectomy and has remained seizure free.

Patient P4 was a 45-year-old female with an 11-year duration of medically intractable epilepsy, with seizures occurring several times per month. Seizures began with an aura of anxiety, progressing to automatisms, alteration of awareness and tongue biting. The patients highest level of education is a Master of Science, and neuropsychological testing revealed no lateralizing find-

ings, with an estimate of overall intelligence in the high-average range. A 3T MRI revealed a lesion in the left inferior temporal gyrus consistent with low-grade glioma. She underwent intracranial monitoring via implantation of left temporal subdural grids and depth electrodes. Findings were consistent with peri-lesional seizure onset. She underwent resection of the lesion, found to be a ganglioglioma, and has remained seizure free.

Neuropsychological Tests

Stimuli

Neuropsychological assessments with P1 were conducted for research purposes before and 1.5-weeks, 6-weeks, and 3-months post-surgery. Tasks included the word naming test reported by Behrmann and colleagues (Behrmann and Shallice, 1995) that manipulates word-length (40 of each three, five, and seven-letter words) with frequency and concreteness matched, and a mixed-naming test that included 10 of each type of the following stimuli: letters, six-letter words, single digits, three-digit numbers, famous faces known by the patient, pictures, and single musical notes and guitar tab chords (due to the patients interest in reading music). A broad array of standardized neuropsychological tests were administered pre- and post-surgery, but we only report the two most relevant to his reading ability: TOWRE (Sight Word Efficiency and Phonemic Decoding Efficiency tests (Torgesen et al., 1999)) and CTOPP Phonological Awareness (Elision and Blending Words (Wagner et al., 1999)), which were administered before, and 6-weeks and 3-months post-surgery (Figure S1).

Design and Procedure

For both the word-length effect task and the mixed naming task, stimuli were presented in the center of the screen until they were named without a time limit. The patient pressed the spacebar upon naming the item and all responses were recorded using a digital recorder. A fixation cross was displayed between each stimulus and the patient had to press the spacebar again to display the next stimulus. A single tone was played simultaneously as the stimulus was presented, and precise naming times were later extracted from the digital auditory files using the Audacity program (audacity.sourceforge.net). Standard procedures were followed for the TOWRE (Torgesen et al., 1999) and CTOPP (Wagner et al., 1999).

Post-Resection Neuropsychological Assessment

P1 showed no difference in reading times based on word length pre-surgery (mean reading times for 3, 5, and 7 letter words were 583 ms, 589 ms, and 582 ms respectively; Figure 3b). At 1.5-weeks, 6-weeks and 3-months post-surgery, P1 showed a linear increase in reading times as a function of word length after surgery and lmFG resection (mean latency: 1583 ms), and there was a consistent letter-by-letter reading pattern in each session, with longer latencies for longer words (slopes of 277 ms, 317 ms, and 310 ms per letter in each session, see supplement for more details). A 2×3 ANOVA was conducted with session (pre and post- surgery mean latencies) and word length (3, 5, 7) as repeated measures. There was a significant main effect of session,

$F(1, 39) = 557.75$, $\eta^2 = 0.94$, $p < 0.001$, with longer latencies in the post-surgery session, and a significant main effect of word length, $F(2, 78) = 21.47$, $\eta^2 = 0.36$, $p < 0.001$, with longer words having greater latencies. Importantly, there was a significant interaction between session and word length, $F(2, 78) = 23.33$, $\eta^2 = 0.37$, $p < 0.001$, such that the word length effect was greater post-surgery than pre-surgery.

Significant increases in naming times from pre-surgery to post-surgery (average of all three post-sessions) were observed for words $t(28) = 4.63$, $d = 1.74$, $p < 0.001$, and letters, $t(32) = 3.87$, $d = 1.35$, $p < 0.001$, in addition to 3-digit numbers, $t(35) = 3.49$, $d = 1.18$, $p < 0.001$ (t-tests assuming unequal variances and *df* adjusted based on Levene's test for equality of variances for all three conditions; Figure 2.3c). The largest magnitude increase in naming times was observed with words (103%). The finding of slower numeral naming after removal of the lmFG is consistent with a weaker left-hemisphere 'visual number form area' that is also sensitive to letters and words (Shum et al., 2013). Significant changes were not found for any other categories. The selectivity of P1's deficits confirms that the resected tissue was an integral component of a symbolic orthographic processing network that operates at both the sublexical and lexical levels.

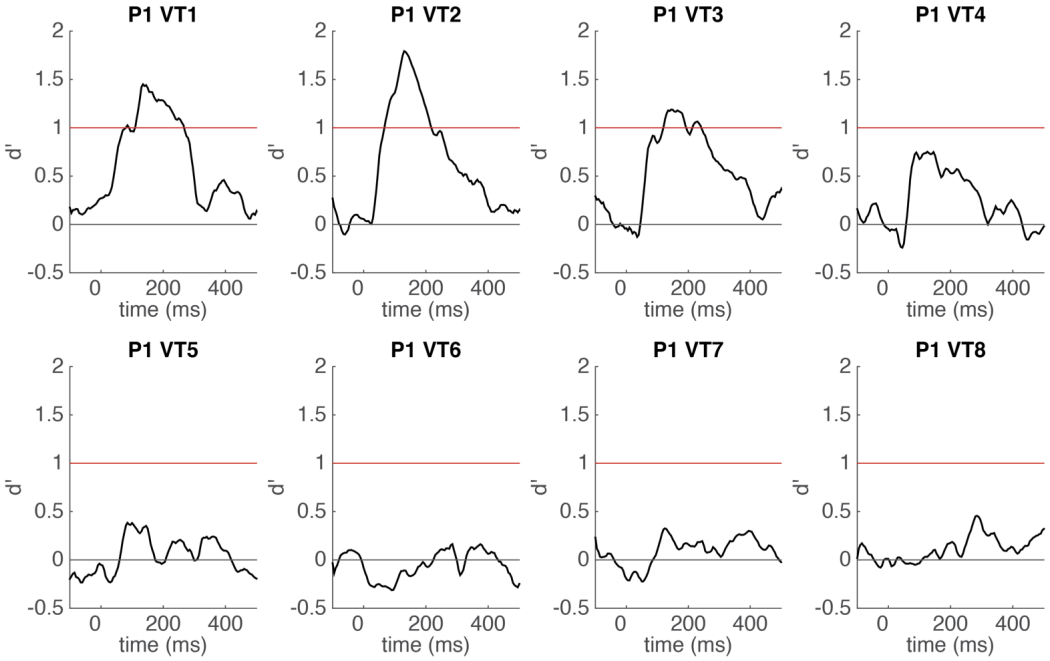


Figure 2.6: **Time course of word categorical sensitivity in each single electrode of P1.** Time course of word categorical sensitivity in each single electrode of P1 measured by sensitivity index d'(mean d' plotted against the beginning of the 100 ms sliding window). The classifier uses time-windowed ERP signal from a single electrode (window length = 100 ms) as input features (See Methods for details). Standard errors of cross-validations are shaded grey. Horizontal red line indicates significance threshold. Horizontal grey line indicates chance level ($d' = 0$). Electrodes 1-3 were the contact of interest for further analysis in P1.

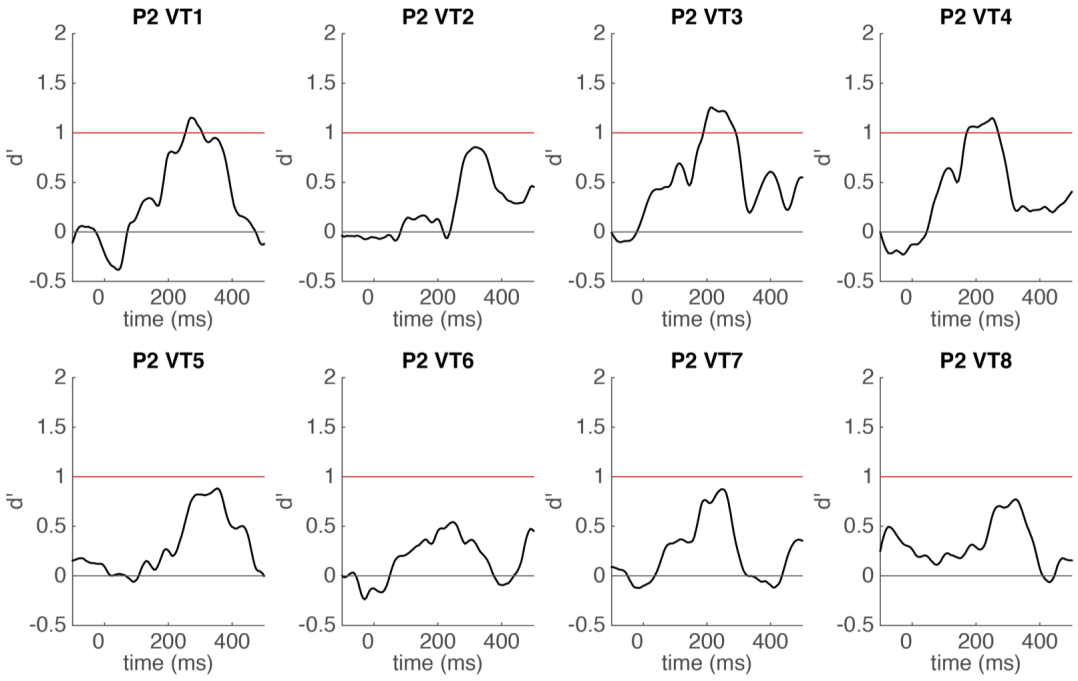


Figure 2.7: **Time course of word categorical sensitivity in each single electrode of P2.** Time course of word categorical sensitivity in each single electrode of P2 measured by sensitivity index d'(mean d' plotted against the beginning of the 100 ms sliding window). The classifier uses time-windowed ERP signal from a single electrode (window length = 100 ms) as input features (See Methods for details). Standard errors of cross-validations are shaded grey. Horizontal red line indicates significance threshold. Horizontal grey line indicates chance level ($d' = 0$). Electrodes 3 and 4 were the contact of interest for further analysis in P2 as electrode 1 was non-contiguous with the other word sensitive electrodes and medial to the fusiform.

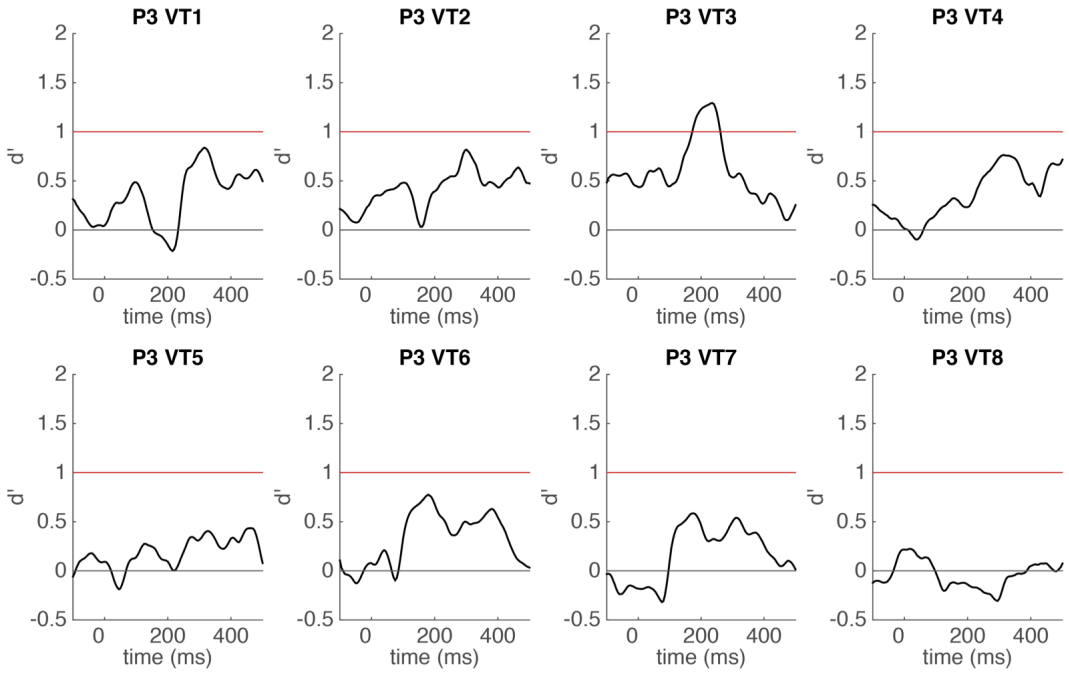


Figure 2.8: **Time course of word categorical sensitivity in each single electrode of P3.** Time course of word categorical sensitivity in each single electrode of P2 measured by sensitivity index d'(mean d' plotted against the beginning of the 100 ms sliding window). The classifier uses time-windowed ERP signal from a single electrode (window length = 100 ms) as input features (See Methods for details). Standard errors of cross-validations are shaded grey. Horizontal red line indicates significance threshold. Horizontal grey line indicates chance level ($d' = 0$). Electrode 3 was the contact of interest for further analysis in P3.

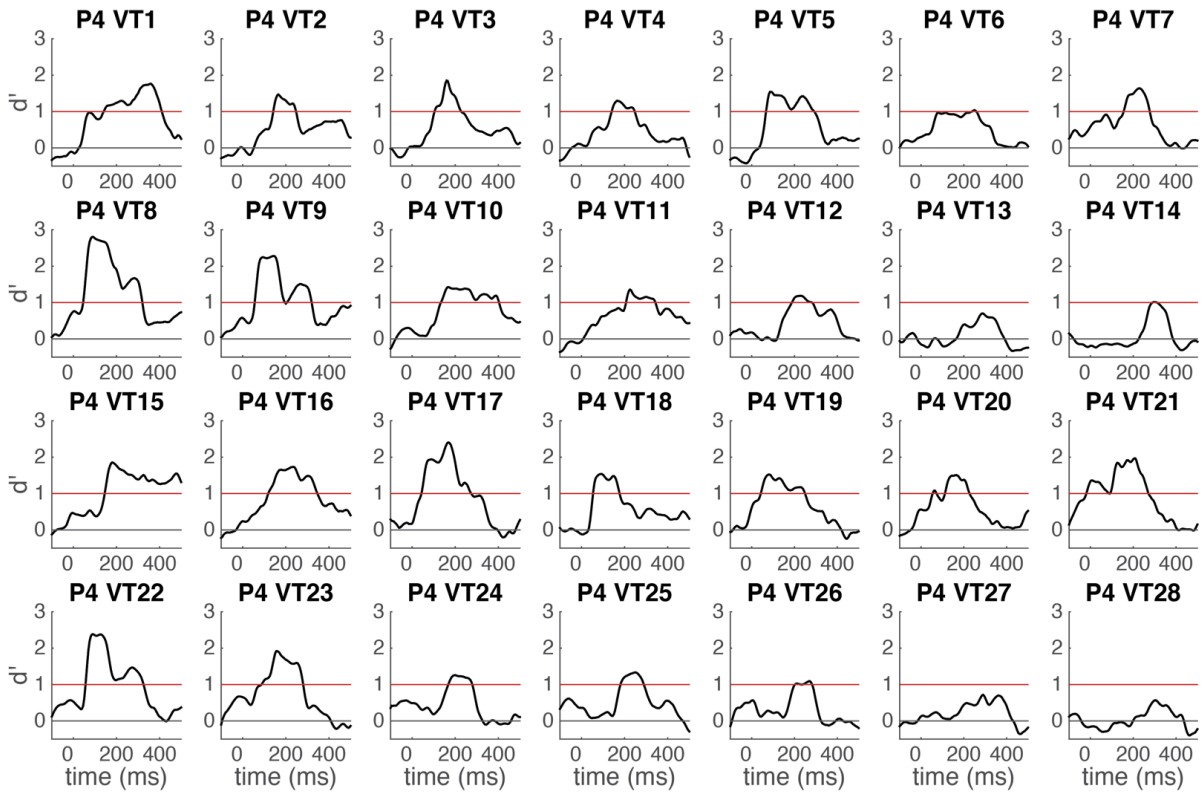


Figure 2.9: **Time course of word categorical sensitivity in each single electrode of P4.** Time course of word categorical sensitivity in each single electrode of P2 measured by sensitivity index d'(mean d' plotted against the beginning of the 100 ms sliding window). The classifier uses time-windowed ERP signal from a single electrode (window length = 100 ms) as input features (See Methods for details). Standard errors of cross-validations are shaded grey. Horizontal red line indicates significance threshold. Horizontal grey line indicates chance level ($d' = 0$). A high-density electrode strip was used in P4. This strip contained 2 rows of 14 electrode contacts and thus the electrodes of interest, 8, 9, and 22, were next to each other. Other electrodes were either substantially medial to the fusiform (1-5, 15-19) or the classification accuracy was a result of stronger activity for the non-word control stimuli (7,10-12, 20, 21, 23-25).

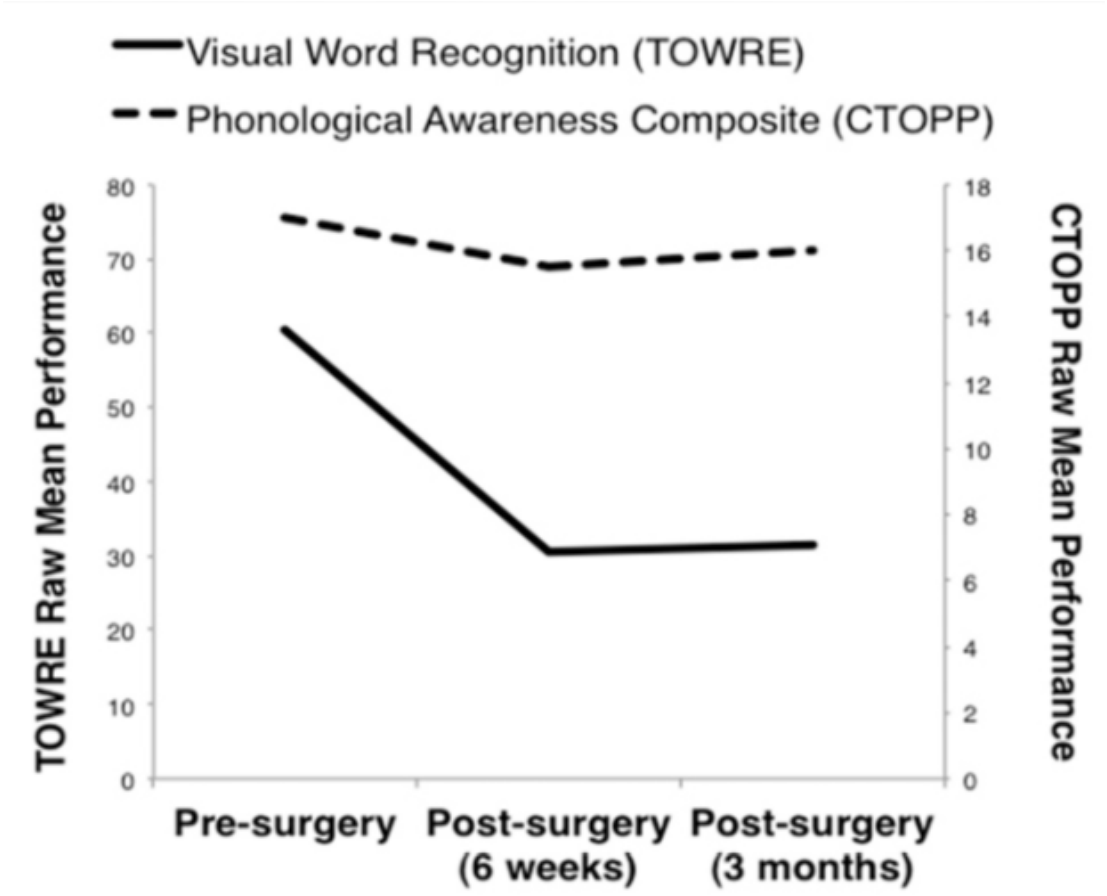


Figure 2.10: **P1 performance on neuropsychological tests.** P1 performance on neuropsychological tests. Visual Word Recognition (mean number of correctly named words in TOWRE Sight Word and Phonemic Decoding Efficiency; pre-surgery: Form A; post-surgery: Form B) and Phonological Awareness (mean correct trials in CTOPP Elision and Blending Words) were measured pre- and post-surgery (6-weeks and 3-months). P1's post-surgery performance on a standardized test of visual word recognition (Torgesen et al., 1999) decreased to a greater extent compared to a test of phonological awareness, which remained stable after surgery (Wagner et al., 1999)(see Figure 2.11). This suggests that P1's resection disrupted orthographic, but not phonological processes.

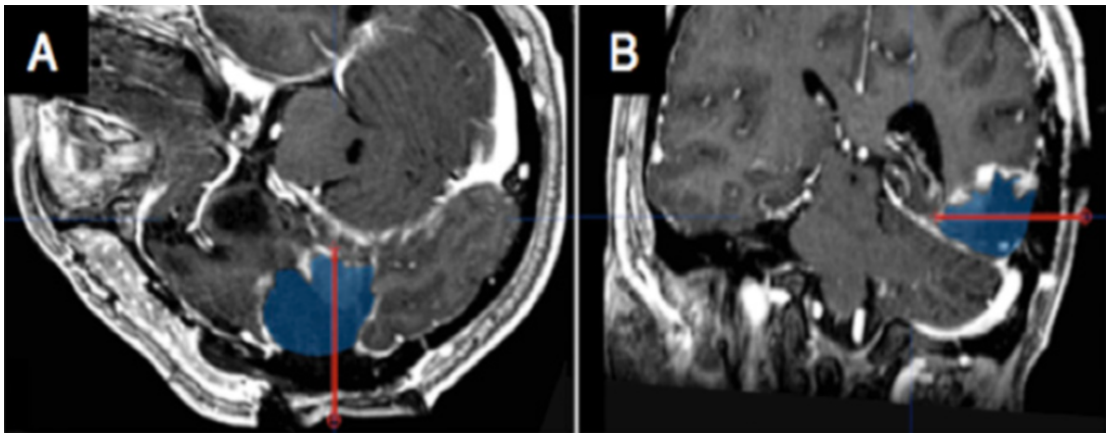


Figure 2.11: **P1 resection location. .**The former location of the same depth electrode (red line) is indicated on co-registered views of the postoperative MRI (A, B), in relation to the cortical regions that were resected (blue region).

Chapter 3

Temporal dynamics in human fusiform underlying face individuation

In addition to reading visual word forms, the other important category of stimuli that requires extensive amount of recognition processing in daily life is the faces. Humans' ability to rapidly and accurately detect, identify, and classify faces under variable conditions derives from a network of brain regions highly tuned to face information. The fusiform face area (FFA) is thought to be a computational hub for face processing, however temporal dynamics of face information processing in FFA remains unclear. In this chapter we use multivariate pattern classification to decode the temporal dynamics of expression-invariant face information processing using electrodes placed directly upon FFA in humans. Early FFA activity (50-75 ms) contained information regarding whether participants were viewing a face. Activity between 200-500 ms contained expression-invariant information about which of 70 faces participants were viewing along with the individual differences in facial features and their configurations. Long-lasting (500+ ms) broadband gamma frequency activity predicted task performance. These results elucidate the dynamic computational role FFA plays in multiple face processing stages and indicate what information is used in performing these visual analyses.

3.1 Introduction

Face perception relies on a distributed network of interconnected and interactive regions that are strongly tuned to face information (Haxby et al., 2000). One of the most face selective regions in the brain is located in fusiform gyrus (the fusiform face area, FFA). Damage to FFA results in profound impairments in face recognition (Barton et al., 2002) and the FFA is thought to be a processing hub for face perception (Nestor et al., 2011). Recent studies have demonstrated that the FFA activity contains information about individual faces invariant across facial expression (Nestor et al., 2011) and gaze/viewpoint (Anzellotti et al., 2013) and have started to describe some of the organizing principles of individual-level face representations (Cowen et al., 2014; Davidesco et al., 2013; Goesaert and de Beeck, 2013). However, due to the use of low temporal resolution analyses or imaging modalities, little is known regarding the relative timing of when FFA becomes sensitive to different aspects of face-related information. Specifically, face pro-

cessing is thought to occur through a set of partially distinct stages (Bruce and Young, 1986) and it remains unclear in which of these stages FFA participates and, more generally, when they occur in the brain.

Evidence from FFA in humans and the putative analog to FFA in non-human primates has demonstrated that FFA shows strong selectivity for faces versus non-face objects (Allison et al., 1999; Kanwisher et al., 1997; McCarthy et al., 1997; Perrett et al., 1982; Sugase et al., 1999; Tsao et al., 2006). There is disagreement about when exactly the FFA, and the human brain in general, first responds selectively to faces (Itier and Taylor, 2004; Pitcher et al., 2012; Rossion and Caharel, 2011). In particular, it is unknown when FFA becomes face selective relative to areas in lateral occipital cortex (Itier and Taylor, 2004; Pitcher et al., 2009, 2012), relative to single neurons in the cortex of non-human primates (Baylis et al., 1985; Perrett et al., 1982; Sugase et al., 1999; Tsao et al., 2006), and relative to rapid behavioral face detection (Crouzet et al., 2010). A recent study using intracranial electrocorticography (ECoG) showed that fusiform becomes sensitive to the category of a visual object around 100 ms after stimulus onset (Liu et al., 2009). However, the brain network highly tuned to face information (Haxby et al., 2000) may allow faces to be processed more rapidly than other categories of objects. Therefore it remains unclear how early FFA becomes face selective and whether it contributes to face detection.

Regarding face individuation, ensembles of single neurons responsive to individual faces have been identified in face sensitive cortical regions of the non-human primate brain (Freiwald et al., 2009; Leopold et al., 2006; Sugase et al., 1999; Tsao et al., 2006). Studies with humans also show that FFA encodes information about individual faces (Davidesco et al., 2013; Nestor et al., 2011). However, little is known regarding the temporal dynamics of individual face processing in FFA, particularly relative to other processing stages.

Furthermore, it remains unknown whether FFA is sensitive to the key facial features used for face recognition, particularly the eyes, mouth, and configural face information. Single neurons of middle face patch in the non-human primate (a putative homolog of FFA) show sensitivity to external facial features (face aspect ratio, direction, hair length, etc.) and properties of the eyes (Freiwald et al., 2009). A recent ECoG study showed that FFA is sensitive to global and external features of the face and head (face area, hair area, etc.) (Davidesco et al., 2013). Behavioral studies have shown that the eyes are the most important facial feature used for face recognition, followed by the mouth (Haig, 1986) and that configural and holistic processing of faces is correlated with face recognition ability (DeGutis et al., 2013). It remains unknown whether FFA is sensitive to individual differences in these featural and configural properties critical to face recognition, particularly when changeable aspects of the face (e.g. expression) are taken into account.

Finally, how FFA contributes to task-related stages of face processing is undetermined. Specifically, previous studies have described a late, long-lasting (lasting many hundreds of milliseconds) face specific broadband gamma frequency (40+ Hz) activity (Davidesco et al., 2013; Engell and McCarthy, 2010; Kawasaki et al., 2012). Broadband gamma activity is closely related to the underlying population firing rates (Manning et al., 2009; Ray and Maunsell, 2011), both of which are face selective for many hundreds of milliseconds after seeing a face (Engell and McCarthy, 2010; Kawasaki et al., 2012; Tsao et al., 2006), extending well beyond the timeframe of face individuation seen in non-human primates (Tsao et al., 2006). It is unknown what role this long-lasting activity plays in face processing. Here we examine whether this long-lasting gamma band

activity reflects the maintenance of face information in support of perceptual decision-making and working memory processes (Freedman et al., 2003; Shadlen and Newsome, 2001).

We used intracranial ECoG in humans and multivariate pattern classification methods to document the temporal dynamics of face information processing in the FFA from the moment a face is first viewed through response-related processing. Multivariate pattern classification was used to decode the contents and timecourse of information processing in FFA in order to elucidate the dynamics and computational role of this area in face perception. Electrophysiological activity (specifically the timecourse of the single-trial voltage potentials and broadband gamma frequency power) from the epileptically unaffected FFA was assessed while each of four patients (P1-4) participated in two face processing experiments (see Figure 3.1 for electrode locations; all face sensitive electrodes appear to be in mid-fusiform, lateral to the mid-fusiform sulcus, see Weiner et al. (2014) for a detailed description regarding the face sensitive regions of the fusiform). Experiment 1 was adopted to examine the temporal dynamics of face sensitivity and specificity in FFA (e.g. face detection) and experiment 2 was employed to examine the temporal dynamics of face individuation and categorization invariant with respect to facial expression. The results of these experiments demonstrate that within 75 ms of presentation, FFA activity encodes the presence of a face (face detection), between 200-450 ms FFA activity encodes which face it is (face individuation), and late (500+ ms) broadband gamma FFA activity encodes task-related information about faces. These results demonstrate the dynamic contribution of FFA to multiple, temporally distinct face processing stages.

3.2 Methods

3.2.1 Subjects

The experimental protocols were approved by the Institutional Review Board of the University of Pittsburgh. Written informed consent was obtained from all participants.

Four human subjects underwent surgical placement of subdural electrode grids and ventral temporal electrode strips as standard of care for surgical epilepsy localization. P1 was male, age 26, and had seizure onset in the hippocampus. P2 was female, age 30, and had seizure onset in the frontal lobe. P3 was female, age 30, and had seizure onset in premotor cortex. P4 was male, age 65, and had seizure onset in the hippocampus. None of the participants showed evidence of epileptic activity on the FG electrode used in this study. The order of the participants (P1-P4) is chronological based on their recording dates.

3.2.2 Stimuli

In experiment 1, 30 images of faces (50% male), 30 images of bodies (50% male), 30 images of shoes (50% men's shoes), 30 images of hammers, 30 images of houses, and 30 images of phase scrambled faces were used. Phase scrambled images were created in Matlab by taking the 2-dimensional spatial Fourier spectrum of the image, extracting the phase, adding random phases, recombining the phase and amplitude, and taking the inverse 2-dimensional spatial Fourier spectrum. Each image was presented in pseudorandom order and repeated once in each session.

Faces in experiment 2 were taken from the Karolinska Directed Emotional Faces stimulus set (Lundqvist et al., 1998). Frontal views and 5 different facial expressions (happy, sad, angry, fearful, and neutral) from all 70 faces (50% male) in the database were used for a total of 350 face images, each presented once in random order during a session. Due to time and clinical considerations, P3 was shown 40 faces (50% male) from the database for a total of 200 faces each presented once in random order during a session.

All stimuli were presented on an LCD computer screen placed approximately 2 meters from participants' heads.

3.2.3 Experimental paradigms

In experiment 1, each image was presented for 900 ms with 900 ms inter-trial interval during which a fixation cross was presented at the center of the screen ($\sim 10^\circ \times 10^\circ$ of visual angle). At random, 20% of the time an image would be repeated. Participants were instructed to press a button on a button box when an image was repeated (1-back). Only the first presentations of repeated images were used in the analysis.

In experiment 2, each face was presented for 1500 ms with 500 ms inter-trial interval during which a fixation cross was presented at the center of the screen. Subjects were instructed to report whether the face was male or female via button press on a button box. Each individual participated in two sessions of experiment 2 on different days.

Paradigms were programmed in MATLAB™ using Psychtoolbox and custom written code.

3.2.4 Data preprocessing

Data were collected at 2000 Hz. They were subsequently bandpass filtered offline from 1-115 Hz using a second order Butterworth filter to remove slow and linear drift, the 120 Hz harmonic of the line noise, and high frequency noise. Data were also notch filtered from 55-65 Hz using a second order Butterworth filter to remove line noise. To reduce potential artifacts in the data, trials with maximum amplitude 5 standard deviations above the mean across the rest of the trials were eliminated. In addition, trials with a change of more than 25 μV between consecutive sampling points were eliminated. These criteria resulted in the elimination of less than 6% of trials in each subject.

3.2.5 Electrode localization

Coregistration of iEEG electrodes used the method of Hermes et al. (2010). High resolution CT scans of patients with implanted electrodes are combined with anatomical MRI scans before neurosurgery and electrode implantation. The Hermes method accounts for shifts in electrode location due to the deformation of the cortex by utilizing reconstructions of the cortical surface with FreeSurfer™ software and co-registering these reconstructions with a high-resolution post-operative CT scan. It should be noted that electrodes on the ventral surface typically suffer minimal shift as compared to those located near the craniotomy. A cortical surface reconstruction

was not possible in P4 due to the lack of a high-resolution MRI. Instead the high-resolution post-operative CT scan was transformed into MNI space using a low resolution T1 MRI and the electrode locations manually determined.

3.2.6 Electrode selection

Electrodes were chosen based on anatomical and functional considerations. Electrodes of interest were restricted to those that were located on the fusiform gyrus. In addition, electrodes were selected such that their peak 6-way face classification d' score (see below for how this was calculated) exceeded 1.5 and the ERP for faces was larger than the ERP for other the other object categories. To avoid concerns about circularity with regards to electrode selection, only the data from the training set (odd trials, see below) for the classification results reported were used for electrode selection. Thus, all statistical values and classification accuracies reported for 6-way face classification are derived from data independent of those used for electrode selection and classifier training.

This procedure yielded 1 electrode per participant, except for P1 where it yielded 3 nearby electrodes (see Supplementary Fig. 1). In the case of P1, we averaged the signal from the three face sensitive electrodes (all three electrodes are shown in Figure 3.1). For P2 the third electrode displayed a peak d' greater than 1.5, however, in examining the ERP it was evident that face classification accuracy in the third electrode on the strip was due to lesser face activity relative to the other conditions (see Supplementary Fig. 4). Face classification on the fourth electrode for P2 was also above threshold and the activity in this electrode followed the pattern from other subjects (e.g. greater face activity relative to other conditions), thus we chose this electrode. It should be noted that even if the anatomical restriction was lifted and all electrodes were used, no additional electrodes would have been chosen in any participant.

In addition to the 4 participants included in the study, 6 other individuals participated in the experimental paradigm during the study period. None of these individuals had any electrodes that met the selection criteria and thus were not included in the analysis. In 2 of these individuals, there were no electrodes on ventral temporal cortex. The electrode locations from the 4 excluded participants with ventral temporal cortex electrodes are shown in Supplementary Fig. 2. In 1 of these individuals, data quality was poor (excessive noise) for unknown reasons (EP2, none of the electrodes showed any visual response and were anterior to FFA). In 3 of these individuals, data quality was reasonable and there were electrodes on ventral temporal cortex, yet none met the selection criteria (see Supplementary Fig. 3). In one of the non-included participants one electrode exceeded the d' threshold (see Supplementary Fig. 3), but this was due to lesser face activity relative to the other conditions (see Supplementary Fig. 4). Considering the ventral electrode strips are placed without functional or anatomical/visual guidance, a yield of 4/7 individuals with ventral strip electrodes having electrodes placed over highly face selective regions is a substantial yield.

3.2.7 Experiment 1 classification analysis and statistics

For classification, single-trial potentials were first split into odd trials used as the training set and even trials used as the test set. The Euclidean distance between the time windowed data from

each of the test and each of the training trials was then calculated. The single-trial potentials from the test trial were assigned to the stimulus condition with k-nearest neighbors classifier. Alternatively, using the correlation (instead of Euclidean distance) between the test and training sets and the results did not yield substantively different results. The selection of k was determined by finding the greatest d' for k-nearest neighbors classification based on random subsampling validation with 50 repeats using only the training set. True positive and false alarm rates were calculated across all of the test trials. d' was calculated as $d' = \Psi^{-1}(\text{true positive rate}) - \Psi^{-1}(\text{false alarm rate})$ where $\Psi^{-1}(x)$ is the inverse of the Gaussian cumulative distribution function.

Because training and test data were separated (rather than cross validation) and not reversed (e.g. the training and test sets were not switched), there is no statistical dependence between the training and test sets and classification accuracy follows the binomial distribution. The null hypothesis for statistical testing was that the true positive rate was equal to the false positive rate under the binomial distribution (this justifies the use of a one tailed t-test).

3.2.8 Experiment 2 classification analysis and statistics

To determine if information regarding individual faces was present in the timecourse of the single-trial potentials, we used across sessions binary nearest neighbors classification (e.g. $k = 1$). Specifically, the neural responses for the five presentations (each with a different facial expression) of two faces in the second session were used as the training set. The test set was the average signal across the five presentations of one of those faces in the first session. The Euclidean distance between the single-trial potentials from the test face and each training face in a 100 ms window was calculated. The test neural activity was classified as belonging to the face that corresponded to the neural activity in the training set that was closest to the neural activity from the test trial. This procedure was then repeated for all possible pairs of faces and all time windows slid with 5 ms steps between 0-500 ms after the presentation of the face. It should be noted that single trial classification was also examined and while classification accuracy was lower, it was still as statistically significant in each participant as when using the average activity across expressions for the 70 face identities (statistical significance was higher due to the use of 350 individual trials instead of 70 averaged trials, which increased statistical power, 40 faces and 200 trials in P3).

In addition, cross-expression classification was also calculated using the same classifier and time windows as above. In this case the neural response for the eight presentations of four of the expressions (4 expressions $\times$ 2 sessions) of two faces were used as the training set. The test set was the average signal across the two presentations of the remaining expressions for one of those faces in the first session. This procedure was repeated for each pair of faces and with each expression left out as the test set (e.g. leave-one-expression-out cross-validation). Note that using cross-validation, instead of holdout validation as in the cross-session classification, and analyzing the 5 expressions separately, lowered the statistical threshold for this analysis.

Permutation testing was used for statistical testing of classification accuracy in experiment 2. Specifically, the labels of the faces in each session were randomly permuted. The same procedure as above was performed on these permuted trials. The maximum classification accuracy across the 0-500 ms time window was then extracted. Using the maximum classification accuracy across

the time window implies a global null hypothesis over the entire window, which corrects for multiple time comparisons (Maris and Oostenveld, 2007). The labels were randomly permuted again and this procedure was repeated 500 times. Using this procedure, $p = 0.05$, corrected for multiple comparisons, corresponded to a classification accuracy of approximately 57% ($\pm 0.2\%$ across the 4 individuals).

Classification of the 5 facial expressions (Supplementary Fig. 5) was done using k-nearest neighbors as in experiment 1.

Classification accuracy when the two training faces were the same gender or when they were different gender was also compared in Supplementary Fig. 6. This was done because participants' task was gender classification and we wanted to address the potential concern that neural classification for individual faces could have been driven by task demands.

3.2.9 Facial feature analysis

Facial features were determined based on anatomical landmarks found by IntraFace (Xiong and De la Torre, 2013). This toolbox marks 49 points on the face along the eyebrows, down the bridge of the nose, along the base of the nose, and outlining the eyes and mouth. Based on these landmarks we calculated the first 12 facial feature dimensions listed in Figure 3.3B. Red, green, and blue intensities were calculated by taking the average intensity for these colors in two 20×20 pixel squares, one on each cheek, the bottom of which was defined to align with the bottom of the nose and the middle to horizontally align with the middle of the eye. High, middle and low spatial frequencies were determined by calculating the mean power at different levels of a Laplacian pyramid (Burt and Adelson, 1983). The image was iteratively low-pass filtered and subtracted from the original image to generate a 6 level Laplacian pyramid (from level 0 to level 5), similar to 2-dimensional wavelet decomposition. The level with smaller index contained higher frequency components. By adding up in pairs, e.g. level 0&1, level 2&3, level 4&5, we get 3 images that corresponding to the high, mid and low frequency components of the original image (note that if we add all 6 levels together we will get the original image). We then performed a 2-dimensional Fast Fourier Transform for these three images to calculate the mean power for each of them.

The values for these 18 feature dimensions were averaged across the five facial expressions for each of the 70 faces (40 for P3). Finally, the values for each variable were normalized by subtract the mean and dividing by the standard deviation across the 70 faces so that none would unduly influence the canonical correlation analysis.

3.2.10 Canonical correlation analysis

Canonical correlation analysis (CCA) finds the maximally correlated linear combinations of two multidimensional variables (Hotelling, 1936), in this case variable one was the 18 facial feature dimensions and variable two was the single-trial potentials between 200 and 500 ms after stimulus onset. Briefly, the first canonical coefficients of the face and neural variables ($x_1, x_2, , x_m$ and $y_1, y_2, , y_n$) respectively are found by maximizing the correlation between the canonical variables

(W_1 and V_1) defined by:

$$\underset{W_1 \in \mathbb{R}^m, V_1 \in \mathbb{R}^n}{\text{maximize}} \quad (W_1^T x)^T V_1^T y \tag{3.1}$$

$$\text{subject to} \quad \|W_1\| = \|V_1\| = 1 \tag{3.2}$$

This procedure is then repeated for W_2 and V_2 to W_p and V_p where $p = \min(m, n)$ and all Ws are uncorrelated to one another and all Vs are uncorrelated to find subsequent canonical coefficients and functions. Significance of Wilks' λ (the multivariate generalization of the inverse of R^2) was based on the chi-squared statistic.

In the presence of noise, CCA is prone to overfit the data unless the number of samples substantially exceeds the dimensionality of the data. To reduce the dimensionality of the neural data, we performed a principal components analysis (PCA) on the faces x timepoints data (70 faces $\times$ 300 time points) and used the first N eigenvalues as the neural dimensions in the CCA. The number of eigenvalues (N) was chosen such that they accounted for 90% of the variance in the neural data. This yielded 9 eigenvalues for P1, 8 for P2, 9 for P3, and 8 for P4.

3.2.11 Gamma band analysis and statistics

Time-frequency power spectra were calculated using a Fourier transform with a Hanning window taper calculated with a 200 ms sliding window and 2 Hz frequency step for each trial. The peak frequency in the gamma range for all trials in experiment 1 collapsed across conditions and subjects was found to be 65 Hz and a window of $\pm$ 25 Hz around this peak was used as the frequency window of interest. Trials in experiment 2 were ranked by reaction time (RT) and split into fastest, middle, and slowest thirds according to RT. In addition, Spearman's ρ between RT and gamma power across trials was calculated. Spearman's ρ was used to minimize the potential for outliers skewing the correlation, though it should be noted that Pearsons correlation and Spearman's ρ did not substantially differ in any participants and both were significant in all runs and participants.

3.3 Results

3.3.1 Timecourse and magnitude of face sensitivity in FFA

To assess the face sensitivity and specificity of FFA (experiment 1), we used a k-nearest neighbors algorithm to decode the neural activity while participants viewed 6 different categories of visual images: faces, human bodies, houses, hammers, shoes, and phase-scrambled faces (30 images per category, each repeated once, presented in random order; faces, bodies, and shoes were balanced for gender; see Figure 3.2A for examples). Participants pressed a button if an image was repeated in consecutive trials (20% of trials, repeated images were excluded from analysis). Each individual participated in two sessions of experiment 1; one session from P4 was not used due to evidence of an ictal event during the recording (a total of 7 sessions across 4 participants). We classified single trial voltage potentials between 100-250 ms after stimulus presentation into one of the six categories described above and examined the decoding accuracy using the signal

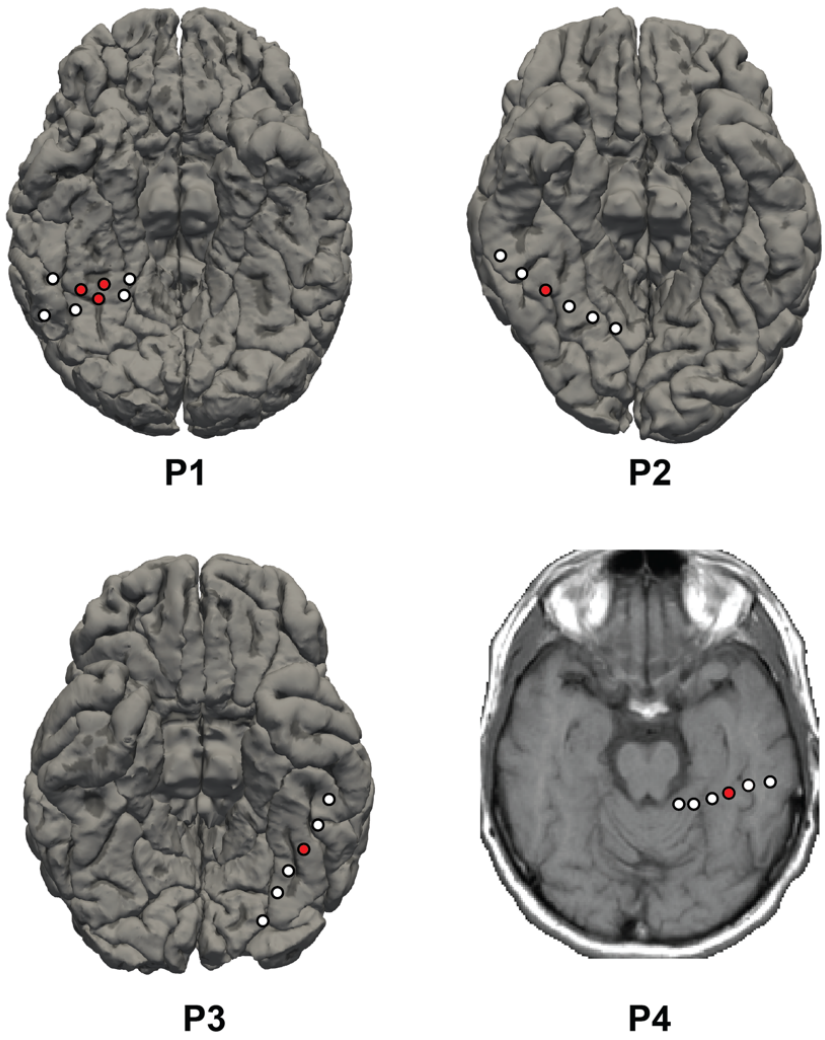


Figure 3.1: **Locations of electrodes used in the study and their neighboring electrodes on subjects' native pial surface reconstruction.** Electrodes in red denote the ones used in the experiment and electrodes in white denote the other contacts on the same electrode strip. A high resolution MRI was not available for pial surface reconstruction of P4 and thus the electrode is visualized on a low resolution T1 MRI slice. MNI coordinates of electrodes are as follows: P1 - (35, -59, -22), (33, -53, -22), (42, -56, -26); P2 - (40, -57, -23); P3 - (-33, -44, -31); P4 - (-38, -36, -30). All electrodes are over the fusiform gyrus.

recorded from face sensitive electrodes (see methods for details on electrode selection and Figure 3.1 for locations). This time range was selected for the initial analysis because it includes most of the previously described face sensitive electrophysiological responses (Allison et al., 1999; Engell and McCarthy, 2010; Itier and Taylor, 2004)(also see Figure 3.2A & B). We were able to identify the category of a stimulus presented on a given trial with 54 - 93% accuracy across the 7 sessions if the stimulus was a face (6-way classification, chance = 16.7%). Neural activity for non-face images was misclassified as a face in 0-8% across the sessions (P1 = 93%/0%,

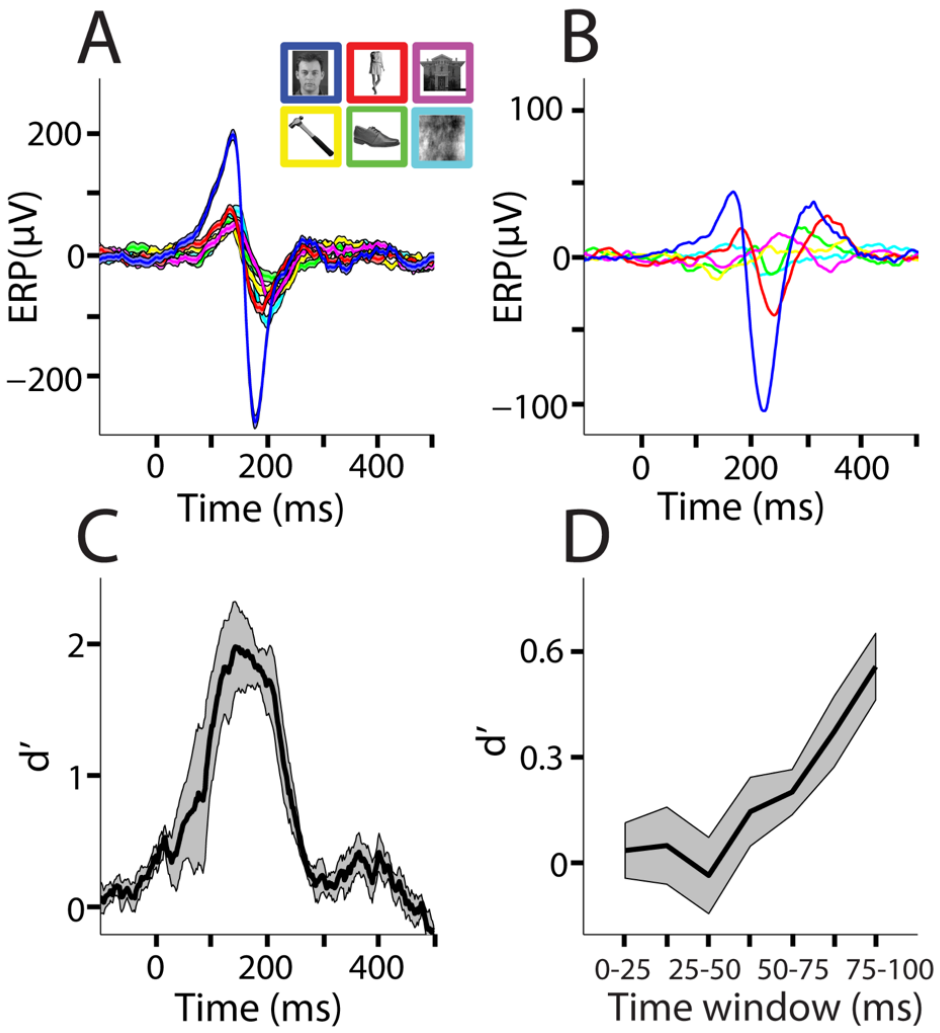


Figure 3.2: **Dynamics of face selectivity in human FFA. (A)** Example of stimuli from each condition and event related potential (ERP) waveforms from session 1 of P1. Across trial means are plotted and standard errors are shaded in light colors. **(B)** Average ERP waveforms across the four participants. In each participant a positive going face sensitive peak between 100-140 ms and a negative going face sensitive peak between 160-200 ms could be identified.**(C)** Face classification accuracy over time as measured by d' ($n = 4$, mean d' plotted against the beginning of the 100 ms sliding window), which takes into account both the true and false positive rate. Classification is based on single trial voltage potentials. See Supplementary Fig. 1 for individual subject d' time courses for these electrodes and neighboring electrodes. Standard deviations are shaded grey. **(D)** Face classification accuracy in the first 100 ms after stimulus onset with 25 ms windows. Classification is based on single trial voltage potentials. d' scores in panels B and C differ due to the different window sizes used for the respective analyses. Standard deviations are shaded grey.

82%/1%; P2= 88%/8%, 54%/8%; P3= 73%/6%, 77%/1%; P4= 67%/8%; true positive rate/false positive rate; chance = 16.7%/16.7%; $p < 10^{-5}$ in each of the eight sessions). Little consistency in classification accuracy was seen across sessions and participants for the five other object categories (Supplementary Table 1). In addition, in all participants electrodes 1-2 cm away from the electrodes of interest showed little face-sensitive (peak sensitivity index $d' < 1$, Figure 3.1 and Supplementary Fig. 1), suggesting that face sensitivity was constrained within 1-2 cm. The high sensitivity and specificity for face classification reported here demonstrates that human FFA regions are strongly face selective (Baylis et al., 1985; Tsao et al., 2006).

Figure 3.2C shows the temporal dynamics of single trial face classification averaged across participants in FFA using the sensitivity index (d'), which takes into account both the true and false positive rate for face detection. Face sensitivity was seen in FFA between approximately 50-350 ms after stimulus onset. To determine the onset of face selective activity in FFA, we examined the d' for face classification from 0-100 ms in 25 ms moving windows shifted by 12.5 ms. All windows between 50-100ms showed significant face sensitivity (Figure 3.2D, 50-75 ms: mean $d' = 0.200$, $t(3) = 3.13$, $p = 0.0260$; 62.5-87.5 ms: mean $d' = 0.368$, $t(3) = 3.72$, $p = 0.0169$; 75-100 ms: mean $d' = 0.551$, $t(3) = 5.91$, $p = 0.0048$), earlier time windows did not reach statistical significance. None of the other five categories, including phase scrambled faces, showed significant classification in these time windows. This suggests that this rapid face processing was not driven by spatial frequency information (Rossion and Caharel, 2011) as phase scrambled faces contain the same spatial frequency content as intact faces. The 50-75 ms time window is earlier than human fusiform becomes sensitive to other visual object categories (Liu et al., 2009). However, this time window is consistent with the reports of the earliest face sensitivity in single cortical neurons in non-human primates (Baylis et al., 1985; Perrett et al., 1982; Sugase et al., 1999; Tsao et al., 2006) and rapid behavioral face detection (Crouzet et al., 2010) suggesting that FFA is involved in face detection.

3.3.2 Timecourse of individual-level face processing in FFA

In each of two sessions recorded on separate days, P1-P4 were shown 70 different faces, each repeated 5 times with different facial expressions each time (happy, sad, angry, fearful, and neutral expressions) for a total of 350 unique images. The participants' task was to report the gender of each face they saw (50% male, 50% female faces). We used a nearest neighbor classification algorithm to determine how accurately we could predict which face (given two drawn from the set of faces) a participant was viewing at a particular moment in session 1 based on a model trained on the timecourse of the single-trial voltage potentials from session 2. Session 2 was used as the training set and session 1 as the test set for this analysis to test classification on previously unseen faces. In each of the four participants in experiment 2, above chance intra-session classification of the neural response to individual faces was observed (Figure 3.3A, $p < 0.05$ using a permutation test, corrected for multiple time comparisons). Classification accuracy peaked in P1 at 65% and was significant in the 210-390 ms time window, in P2 at 59% and was significant in the 280-460 ms time window, in P3 at 63% and was significant in the 270-490 ms time window, and in P4 at 60% and was significant in the 350-540 ms time window (chance = 50%; 57% corresponds to $p = 0.05$ corrected for multiple comparisons). In addition, we examined whether individual-level face classification was invariant over expression by training the classifier on four

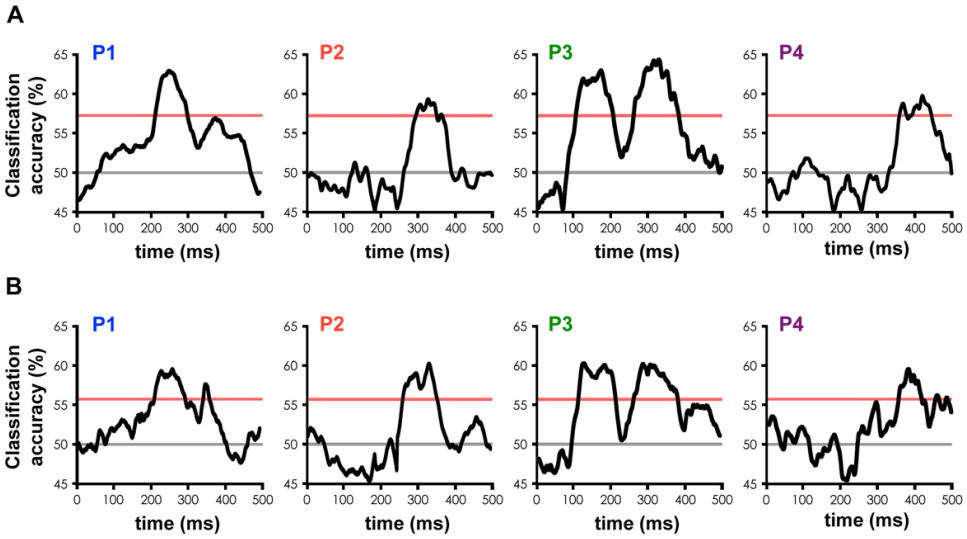


Figure 3.3: **Face individuation in human FFA. (A)** Time course of individual level face classification accuracy based on single trial voltage potentials in each participant. This shows, given two faces, how accurately we could predict which one the participant was viewing based on the neural data, plotted against the beginning of the 100 ms sliding window. Red line at 57% indicates $p = 0.05$, corrected for multiple time comparisons based on the permutation test, grey line indicates chance accuracy (50%). **(B)** Across-expression, individual level face classification accuracy. This shows, given two faces with a particular expression, how accurately we could predict which one the participant was viewing based on the neural data from the other four expressions used in the study. Red line at 55.5% indicates $p = 0.05$, corrected for multiple time comparisons based on the permutation test, grey line indicates chance accuracy (50%).

of the five expressions and testing the other, then repeating this with different expressions in the training and test set until each expression (leave-one-expression-out cross-validation). In each participant, above chance across-expression classification of the neural response to individual faces was observed (Figure 3.3B, $p < 0.05$ using a permutation test, corrected for multiple time comparisons). This across-expression classification had a similar timecourse as the across-session classification in Figure 3.3A suggesting that the coding for individual faces in FFA is not driven by low-level differences between images and is at least partially invariant over expression. Indeed, classification of expression failed to reach statistical significance at any point between 0 and 500 ms (Supplementary Fig. 5). In addition, classification accuracy across face genders was similar to classification within face gender (Supplementary Fig. 6), suggesting that classification of individual faces in FFA was not driven by task demands. Also, training with the data from session 1 and classifying the data from session 2 changed the peak classification accuracy by less than 0.5%, the peak time by less than 15 ms, and the significant time window by less than 25 ms. Furthermore, individual faces could not be classified above chance in the adjacent or nearby electrodes (Supplementary Fig. 7). These results suggest that the 200-500 ms time window is critical for expression-invariant face individuation in FFA.

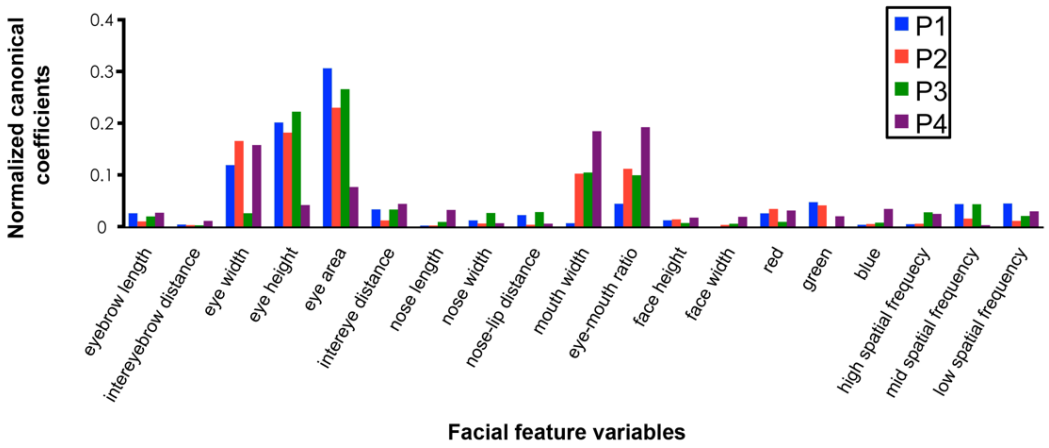


Figure 3.4: **Facial feature sensitivity of FFA electrodes.** Multivariate canonical correlation coefficients between the single trial voltage potentials and facial features for the individual faces. Canonical coefficients have a similar interpretation as beta coefficients in multiple regression. Coefficients were normalized by first taking the absolute value and then dividing by the sum of all coefficients across the 18 facial feature variables.

3.3.3 Facial information used in service of face individuation

To investigate what specific face information FFA encodes in the service of face individuation we mapped anatomical landmarks on each of the faces presented in experiment 2 and projected each face into an 18-dimensional "feature space" that applied to all faces (e.g. eye area, nose length, mouth width, skin tone, etc.; see Figure 3.4 for a full list of the features used) (Xiong and De la Torre, 2013). The multivariate canonical correlation between these facial feature dimensions and the voltage potentials between 200-500 ms post-stimulus onset was then calculated to evaluate the shared relationship between these variable sets. The full canonical model between the neural activity and the face feature space was significant in P1, P3, and P4 and approached significance in P2 (P1: $\chi^2(171) = 211.33$, Wilks' $\lambda = 0.021$, $p = 0.019$; P2: $\chi^2(152) = 181.21$, Wilks' $\lambda = 0.045$, $p = 0.053$; P3: $\chi^2(171) = 230.93$, Wilks' $\lambda < 0.001$, $p < 0.001$; P4: $\chi^2(152) = 194.06$, Wilks' $\lambda = 0.03$, $p = 0.012$) demonstrating that FFA activity is sensitive to individual differences in these facial feature dimensions. Only the full model was significant as none of the other hierarchical statistical tests reached significance. Figure 3.4 presents the normalized function weights for the full canonical model demonstrating that the most relevant facial variables were related to the eyes, the mouth, and the ratio between eye and mouth dimensions. There are also notable differences across participants, with P1 showing strong sensitivity to eye information and almost no sensitivity to mouths and P4 showing strong sensitivity to mouth information and less to eyes. It is unclear if these differences are due to different electrode locations (see Figure 3.1), random variation (as we do not have the power with only 4 participants to statistically quantify these individual differences), or different face processing strategies among participants. More generally, we did not track eye movements and therefore cannot relate our results to particular face processing strategies or preclude FFA sensitivity to other internal or external facial features (Davidesco et al., 2013; Freiwald et al., 2009). Rather our results show that, under free viewing

conditions, FFA is tuned to natural variations in eye and mouth feature dimensions and configural information relating the eyes to the mouth in service of face individuation.

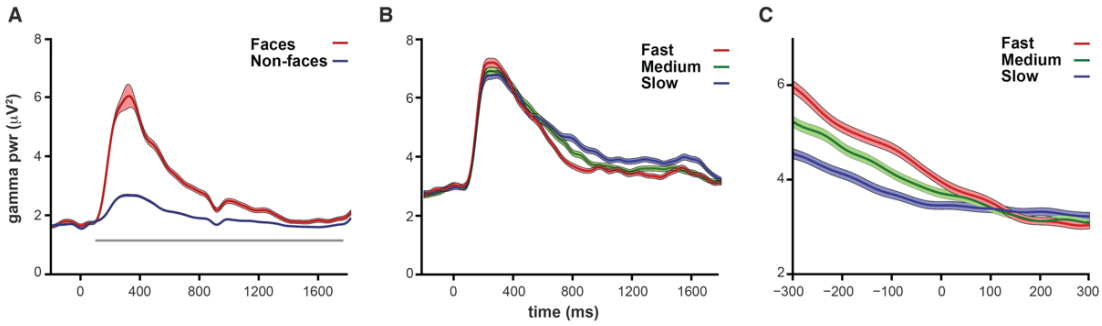


Figure 3.5: **Long-lasting task-related broadband gamma activity. (A)** Mean and standard error of gamma band (40 - 90 Hz) power for face and non-face trials across all participants in experiment 1 ($n = 4$). Grey bar indicates $p < 0.05$ for face versus non-face objects based on the Wilcoxon rank-sum test. See Supplementary Fig. 8 for face and non-face gamma band power for each individual participant. **(B)** Mean and standard error of gamma band power split into thirds by reaction time for gender discrimination in experiment 2 ($n = 4$; mean reaction time = P1: 788 ms; s.d. = 269 ms; P2: 870 ms; s.d. = 221 ms; P3: 1065 ms; s.d. = 299 ms; P4: 872 ms; s.d. = 216 ms). Significant correlation was seen in each individual participant between 500-1000 ms gamma band power and reaction time (Figure 3.5A). **(C)** Same as for (B) but with trials aligned to the behavioral response (time 0 = response onset) for the 4 participants in experiment 2. A significant correlation between pre-response gamma band power (-300 ms to -100 ms) and reaction time was seen in each individual participant (Figure 3.5B).

3.3.4 Broadband gamma activity predicts task performance

Finally, we examined the role of the slowly decaying broadband gamma power (40-90 Hz) activity that has been shown to be face sensitive (Davidesco et al., 2013; Engell and McCarthy, 2010; Kawasaki et al., 2012). The results from experiment 1 confirm that this gamma activity shows strong selectivity for faces and also showed that it lasts for the entire trial (Figure 3.5A and Supplementary Fig. 8). Experiment 1 was a working memory task and one possible role for face-specific activity that persists for the entire trial is task-related maintenance of face information that is manipulated by frontal and/or parietal regions involved in working memory and decision making (Freedman et al., 2003; Lara and Wallis, 2014; Shadlen and Newsome, 2001). In support of this hypothesis, in repeated trials face activity decayed more rapidly than in first presentations, potentially due to the release of task demands once detection was accomplished. However, the relative paucity of repeated face trials and decreased face activity due to repetition suppression makes interpreting these results difficult. Thus, to test the hypothesis that broadband gamma frequency activity was related to maintaining the face representation in support of task-related processing, we examined the relationship between long-lasting gamma activity and behavioral reaction time in experiment 2. In support of a role in task-related processing, the decay time of the gamma activity from 500-1000 ms after stimulus presentation predicted reaction

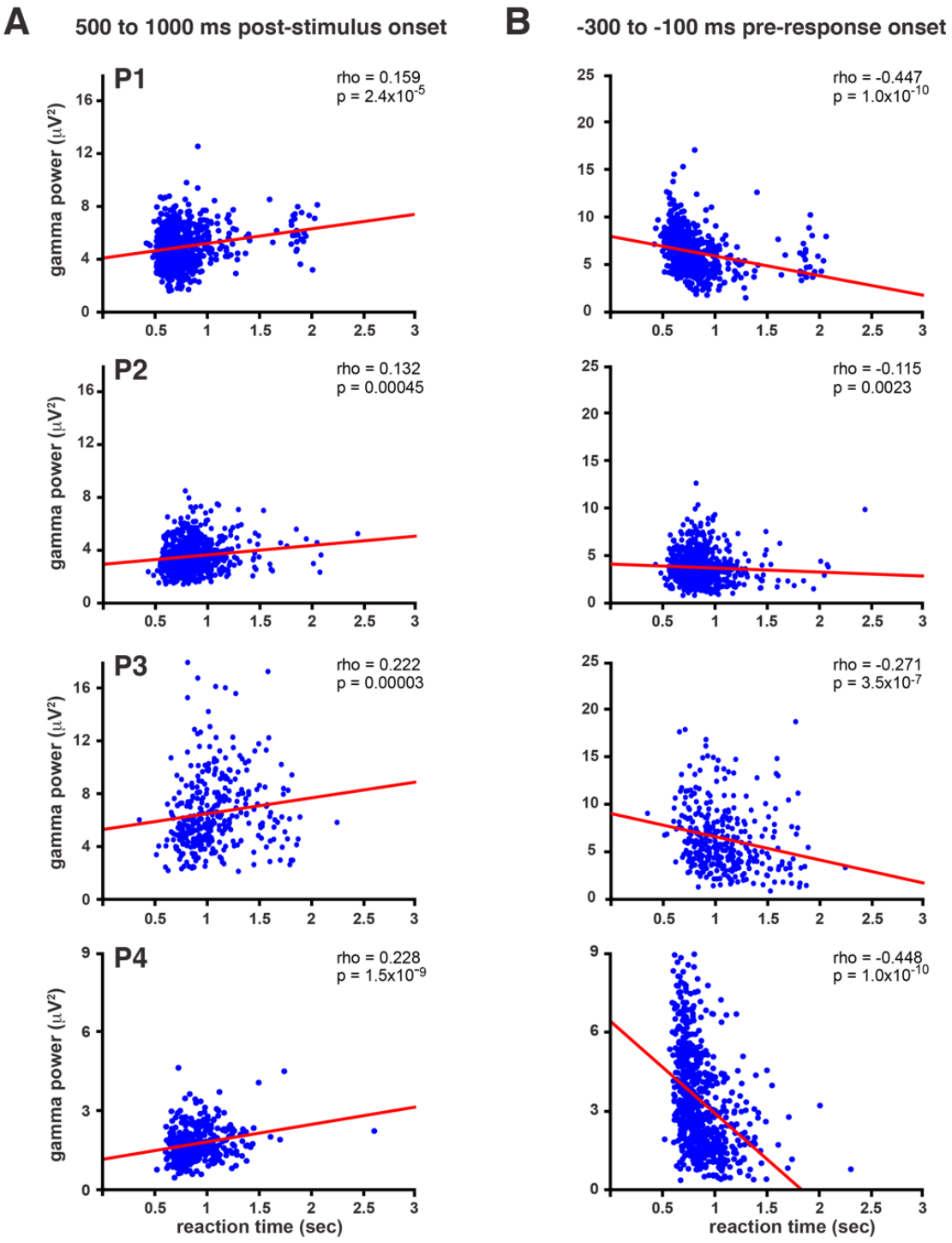


Figure 3.6: **Gamma power predicts reaction time in each participant.** **(A)** Scatter plots of mean gamma power over the 500 to 1000 ms post stimulus window versus reaction time. **(B)** Scatter plots of mean gamma power over the -100 to -300 ms pre response window versus reaction time.

time in experiment 2. Specifically, longer lasting gamma activity was significantly correlated with slower response times ($p < 0.05$) in the gender identification task for each participant (Figure 3.5B and 3.6A). The amplitude of this gamma band activity 100-300 ms prior to the response significantly predicted reaction time for each participant and this activity returned to baseline only once the participants had responded and task demands had waned (Figure 3.5C and 3.6B). While this gamma frequency power significantly predicted reaction time, we were unable to decode the gender decision of the participant from this activity. In summary, greater than baseline, face-specific broadband gamma power was seen until the point of behavioral response, and a larger gamma peak and more rapid decay predicted more rapid decisions of face gender, but this gamma activity did not predict behavioral responses (i.e. "male" or "female").

3.4 Discussion

Our results establish the timecourse of information processing in human FFA and elucidate the specific computations FFA performs on faces from the moment a face is first viewed through decision-related processing. These results demonstrate that FFA activity first contains face specific information approximately 50-75 ms after subjects viewed a face. FFA displays sharp face sensitivity between 100-250 ms, with little evidence for selectivity for four other categories of non-face objects or phase scrambled faces. Individual-level face information invariant over facial expression could be decoded for previously unseen faces between 200-500 ms. During this same time window, the neural activity from FFA contained information about individual differences in eye and mouth features and the relative size of eyes versus mouths, suggesting that the FFA uses this information to individuate faces. Finally, late, long-lasting (500+ ms) gamma frequency band activity (40-90 Hz) predicted participants trial-by-trial reaction times in a gender categorization task. Taken together, these results reveal the highly dynamic role that FFA plays in multiple distinct stages of face processing.

One caveat of the current work is that the input to all of our analyses was the timecourse of FFA activity recorded from single electrodes in each participant. The significant decoding demonstrated in this analysis suggests that FFA displays at least a degree of temporal encoding of face information (Richmond et al., 1987). However, the data we report are only weakly sensitive to information that is primarily coded spatially. Specifically, the data are differentially sensitive to neural populations with different proximity or different orientations relative to the electrodes. That said information primarily encoded spatially is far less likely to be detected by our analyses than information encoded temporally. Thus, lack of significant classification (for example for expression or the gender decision) does not necessarily imply that FFA is not sensitive to this information, but rather it is not coded temporally.

FFA is face sensitive in the 50-75 ms time window. This time window is as early (or earlier) as face sensitivity in lateral occipital face sensitive regions (Itier and Taylor, 2004; Pitcher et al., 2009) and is consistent with onset of face sensitivity reported for single cortical neurons in non-human primates (Baylis et al., 1985; Perrett et al., 1982; Sugase et al., 1999; Tsao et al., 2006). Behaviorally, it has been shown that humans can saccade towards a face within 100-150 ms (Crouzet et al., 2010). The decoding of face information in the 50-75 ms time window reported here is consistent with FFA playing a role in this rapid face detection. The early face sensitivity

of FFA reported here provides strong evidence that this area is involved in face detection.

A recent human ECoG study showed that category selective activity is first observed in temporal cortex around 100 ms after stimulus onset (Liu et al., 2009). Our results show that human FFA becomes face sensitive in the 50-75 ms window, suggesting that faces are processed more rapidly in temporal cortex than other object categories. Indeed, studies of single neuron firing latencies in non-human primates have reported that face sensitivity first arises around the 50-75 ms window (Baylis et al., 1985; Perrett et al., 1982; Sugase et al., 1999; Tsao et al., 2006). This more rapid processing of face information may be a result of the network of areas highly tuned to face properties (Haxby et al., 2000). Future studies will be required to determine if non-face categories with highly tuned perceptual networks (e.g. words (Behrmann and Plaut, 2013) and bodies (Peelen and Downing, 2007)) are also processed more rapidly than other categories of objects. One caveat is that the ECoG study by Liu et al. (2009) reported that the 100 ms object category response in temporal cortex shows invariance to viewpoint and scale changes and future studies will be required to determine if the 50-75 ms FFA face sensitive response is invariant over these transformations as well.

The time window critical for individual level classification occurred between 200-500 ms, after face sensitivity observed in experiment 1 had mostly waned. One potential explanation why face individuation occurred during a period where face-specific activity is relatively weak is that individual level face information may be represented by relatively few neurons (sparse coding) (Young and Yamane, 1992). Sparse coding would imply that relatively few face sensitive neurons were active and that the summed face-related activity in this time period therefore would be weak. However, the neurons that were active encode for individual level face information, which would explain the significant decoding of identity we report here. One point to note is that while face-specific voltage potentials had waned in this time period, significant face-specific broadband gamma activity was observed in the same time period as individual level face classification, though it too was declining. To the extent that this broadband gamma activity reflects single neuron firing (Manning et al., 2009; Ray and Maunsell, 2011), the decrease in this activity potentially also supports a sparse coding hypothesis. One caveat being that further studies are required to determine if the decrease in broadband gamma is due less neurons being active in this time period (sparse coding) or a decrease in the firing rate.

Neuroimaging studies and lesion studies in patients have implicated parts of anterior temporal cortex strongly connected to the FFA (Pyles et al., 2013; Thomas et al., 2009) as important to face individuation (Collins and Olson, 2014; Kriegeskorte et al., 2007; Nestor et al., 2011). Furthermore, a recent study suggested that FFA might act as a hub of face identity processing and act in concert with these anterior temporal face sensitive regions (Nestor et al., 2011). The timewindow in which we found individual-level face coding (200-500 ms) is generally consistent with the idea that recurrent top-down and bottom-up interactions are likely to be critical to face individuation. Note that in P3 and perhaps in P1 there are two peaks of individual-level face classification. More data will be required to statistically substantiate these two peaks, however the dual peaks suggest the possibility of a feedback loop involved in face individuation.

Neural activity in FFA during the same time window when significant individual-face decoding was observed (200-500 ms) displayed significant multivariate correlation to variation in the eyes, mouth, and eye-mouth ratio. Behavioral studies have shown that the eyes are the most important facial feature used for face recognition, followed by the mouth (Haig, 1986) and that

holistic and configural face processing ability is correlated with face recognition (DeGutis et al., 2013). A recent study revealed that electrical stimulation of FFA distorts the perception of facial features (Parvizi et al., 2012). Furthermore, previous studies have demonstrated the importance of the presence of the eyes for face perception in general, and FFA activity in particular (McCarthy et al., 1999). Our results lend strong evidence to the hypothesis that FFA uses individual differences in these facial features in service of face individuation and recognition.

We show that FFA shows face specific gamma frequency power that lasts until task demands wane and that the amplitude of this power predicts participants' reaction times. Recent studies demonstrate that long-lasting FFA gamma activity is modulated by task-related attention to faces and facial expression (Engell and McCarthy, 2010; Kawasaki et al., 2012), in support of the hypothesis that this activity is integral to task-performance. While this activity did predict reaction time, it did not predict the gender decision. This suggests that FFA supports task-related processing, potentially by keeping face information on-line, but decision-specific processing occurs elsewhere, likely in frontal and parietal regions using the information from FFA (Freedman et al., 2003; Shadlen and Newsome, 2001). Indeed, a recent study challenged the view that frontal areas store working-memory and task-relevant information and suggested that these areas instead control and manipulate information that is stored elsewhere (Lara and Wallis, 2014). In the case of faces our results suggest that at least some of this information is stored in FFA.

In summary, our results provide strong evidence that the FFA is involved in three temporally distinct, but partially overlapping processing stages: face detection, expression-independent individuation using facial features and their configuration, and task-related gender classification. Information about these processing stages was present in the recordings from electrodes within a 1 cm radius in each participant suggesting that the same, or at least very nearby, neural populations are involved in these multiple information processing stages. A key open question is how processing transitions between stages in these local neural populations. One hypothesis is that the dynamics of these processing stages are governed by interactions between multiple regions of the face processing network. Taken together with previous findings, the current results suggest a model in which FFA contributes to the entire face processing sequence through computational analysis of multiple aspects of face information at different temporal intervals.

3.5 Appendix: Supplement Results

Electrode selection

Figure 3.7 shows face classification accuracy in electrodes used in the study and their neighbors for all 4 subjects. For P1 there were 4 electrodes on two neighboring strips on the ventral temporal lobes, for P2-P4 there were 6 electrodes on a single ventral temporal strip (Figure 3.1). Electrodes were chosen based on the criteria that peak d' be above 1.5 ($p < .001$, shown in the center column, other than for P1). In P1, three electrodes exceeded this threshold (the middle and fourth electrodes in the first row and the middle electrode in the second row) and for all analyses the signal from these three electrodes was averaged. Independently loading these three electrodes into the analyses does not substantially alter the results and indeed each electrode showed similar d' timecourse in experiment 1 and each showed above chance classification and

similar classification timecourses for individual faces in experiment 2. In P2, two neighboring electrodes exceeded this threshold. However, the signal recorded from the second electrode on the strip (shown in the second column of the third row) was excluded because, unlike the responses from other face sensitive electrodes selected for this study, faces evoked substantially less activity than the other stimulus categories used in experiment 1 in this electrode (see Figure 3.12A for ERP from this channel and exclusion criteria in methods section). The signal recorded from the third electrode on the strip (shown in the center column) displayed the more typical pattern of greater activity for faces relative to the other conditions and thus for P2 this electrode was chosen in the study. Electrodes neighboring the face electrodes chosen for the study had significantly smaller d' in each participant (with the exception of the second electrode in P2, as mentioned above). In addition, in most participants the electrodes neighboring the electrode of interest did not show significant face sensitivity ($p > .05$ corrected for multiple comparisons, this corresponds to a peak d' of .97) and in all participants the electrodes 2 cm away did not show significant face sensitivity.

Table 3.1: **Classification accuracy in the 100-250 ms time window for non-face objects.** Cells contain the true positive rate/the false positive rate for each condition. Bold cells indicate $p < .01$ classification accuracy. Face classification accuracy was significant at $p < 10^{-5}$ in all sessions based on the binomial test.

	P1	P1	P2	P2	P3	P3	P4
Category	Session 1	Session 2	Session 1	Session 2	Session 1	Session 2	Session 1
Faces	**93 / 0**	**82 / 1**	**88 / 8**	**54 / 8**	**73 / 6**	**77 / 1**	**67 / 8**
Bodies	29 / 22	33 / 23	31 / 15	35 / 24	**59 / 14**	17 / 5	30 / 21
Hammers	11 / 15	32 / 30	23 / 11	7 / 18	28 / 23	17 / 9	27 / 23
Houses	22 / 11	37 / 17	15 / 5	31 / 15	10 / 10	23 / 4	33 / 16
Shoes	37 / 26	48 / 30	**44 / 17**	32 / 14	53 / 36	**57 / 24**	23 / 26
Phase scrambled faces	7 / 22	12 / 10	32 / 19	0/8	17 / 8	10 / 12	20 / 9

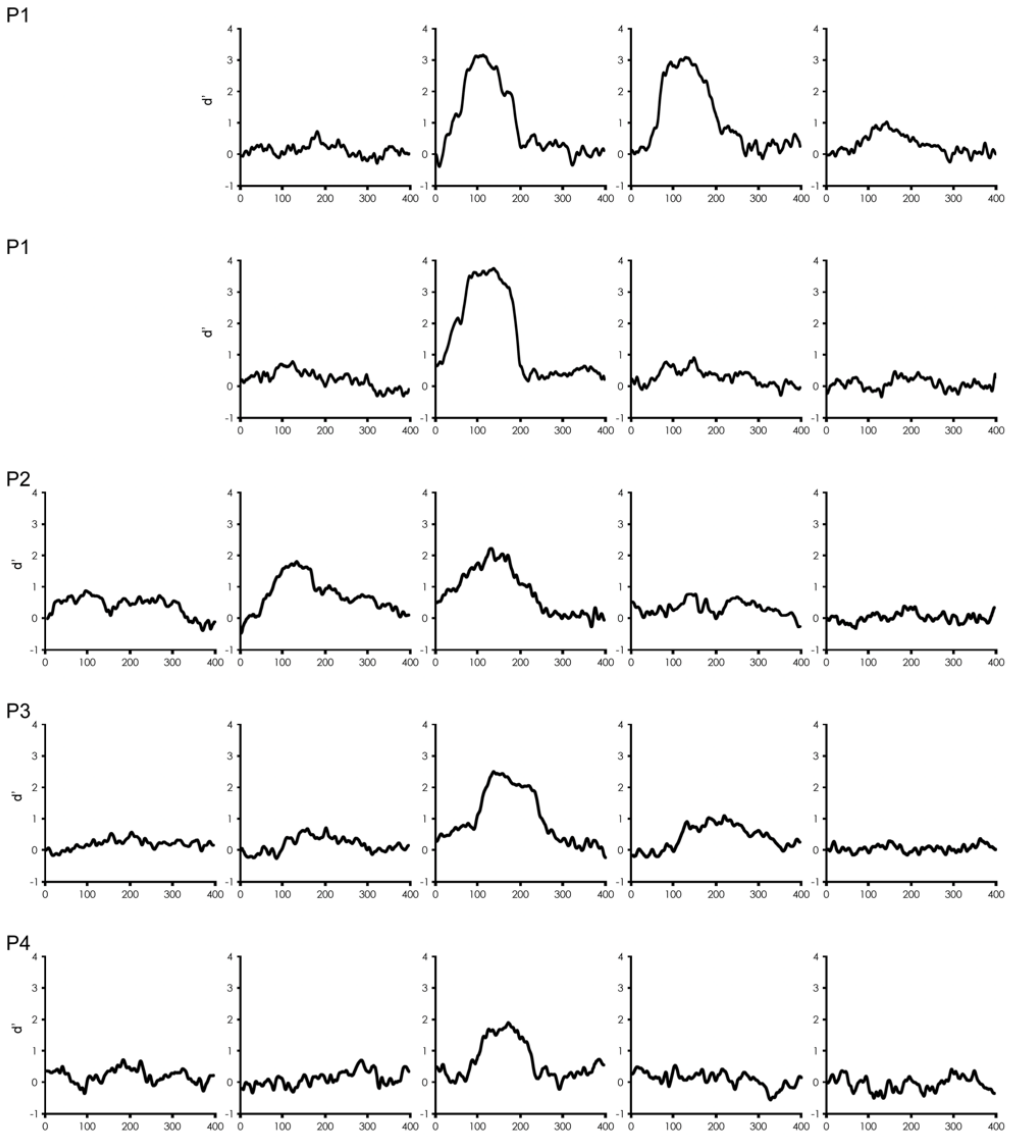


Figure 3.7: **Face classification accuracy in electrodes used in the study and their neighbors.** (see Figure 3.1 for locations of electrodes and these neighbors) Face classification accuracy over time as measured by d' (plotted against the beginning of the 100 ms sliding window) for all electrodes used in the study and their neighboring electrodes. There was 1 cm between the centers of neighboring electrodes. The first column of electrodes represents the most medial and/or posterior electrode on its strip.

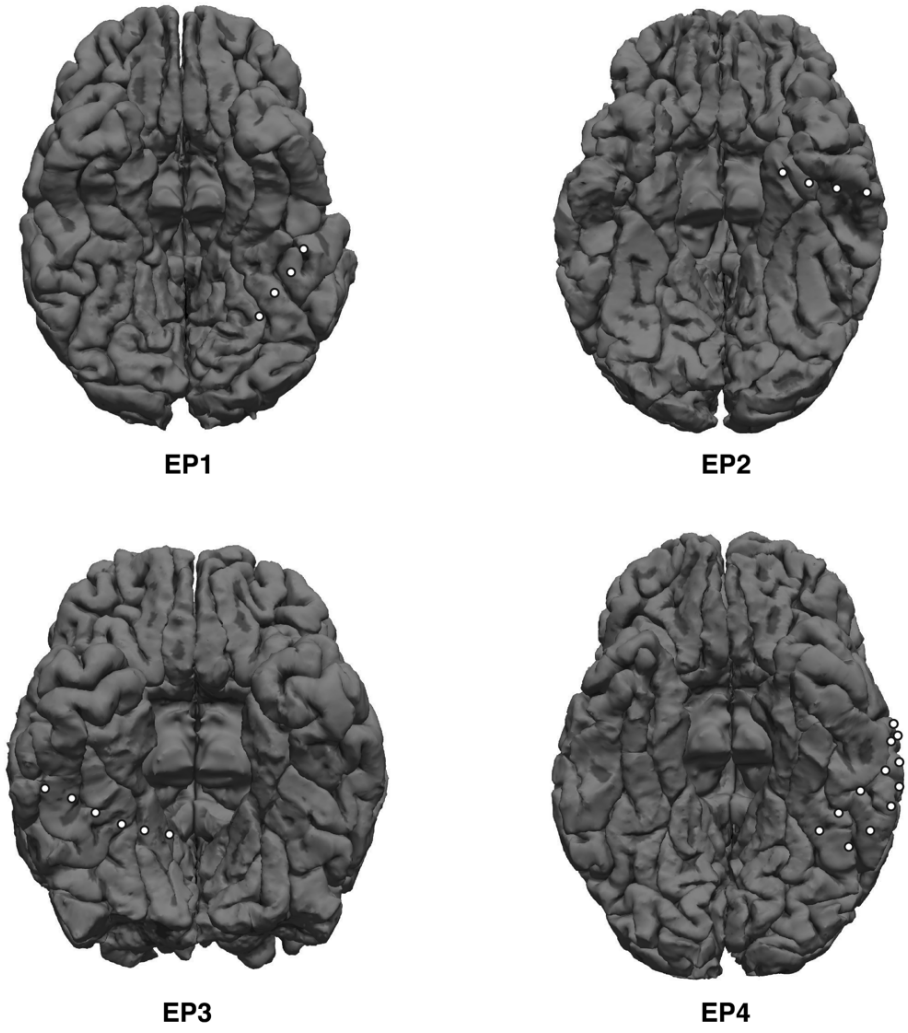


Figure 3.8: **Electrode localization for 4 participants excluded from the main analyses due to lack of face sensitive activity.**

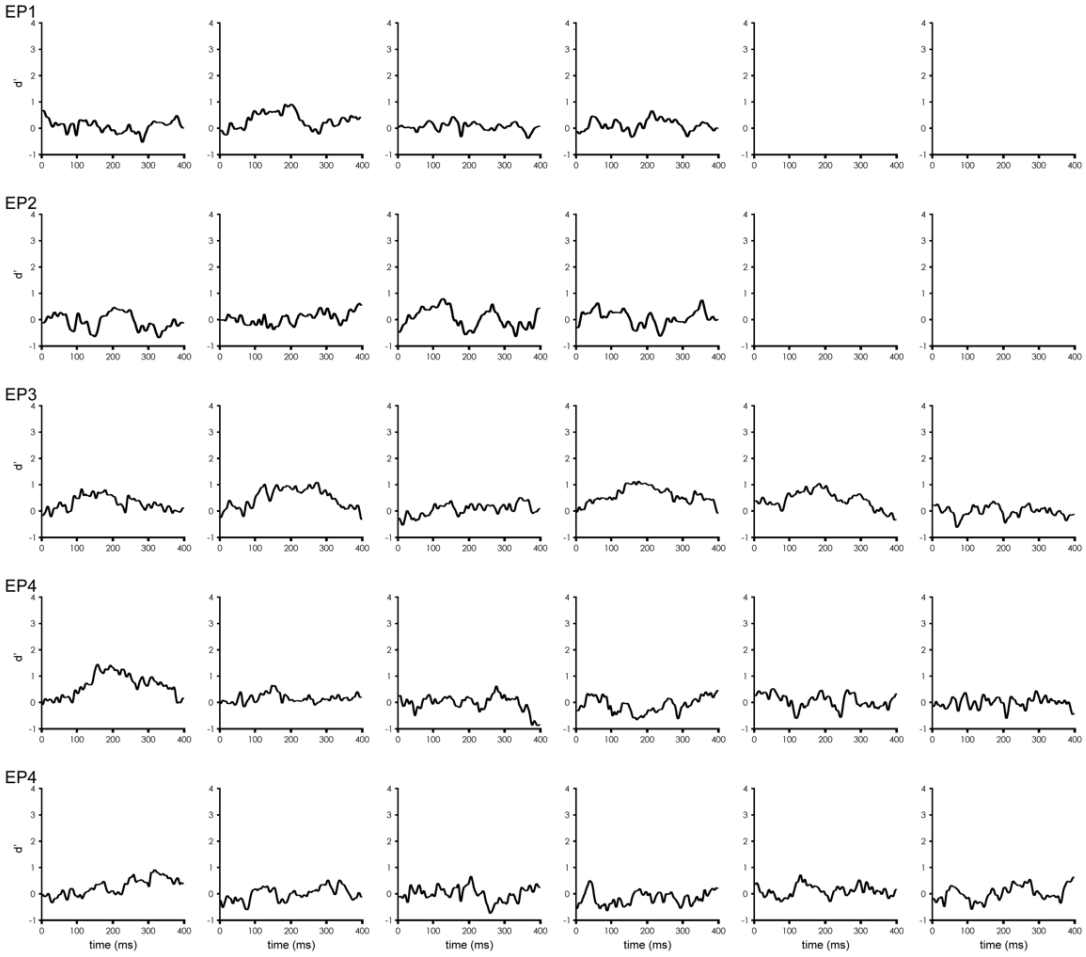


Figure 3.9: **Face classification accuracy in electrodes from participants excluded due to lack of face sensitive electrodes.** (see Figure 3.8 for locations of electrodes and these neighbors) Face classification accuracy over time as measured by d' (plotted against the beginning of the 100 ms sliding window) for all electrodes used in the study and their neighboring electrodes. There was 1 cm between the centers of neighboring electrodes. None of these show significant face sensitivity ($p > .05$ corrected for multiple comparisons, this corresponds to a peak d' of .97) except for the first electrode in EP4 (peak $d' = 1.5$). However, the signal recorded from this electrode was excluded because faces evoked substantially less activity than the other stimulus categories used in experiment 1 in this electrode (see Figure 3.10B for ERP from this channel and exclusion criteria in methods section).

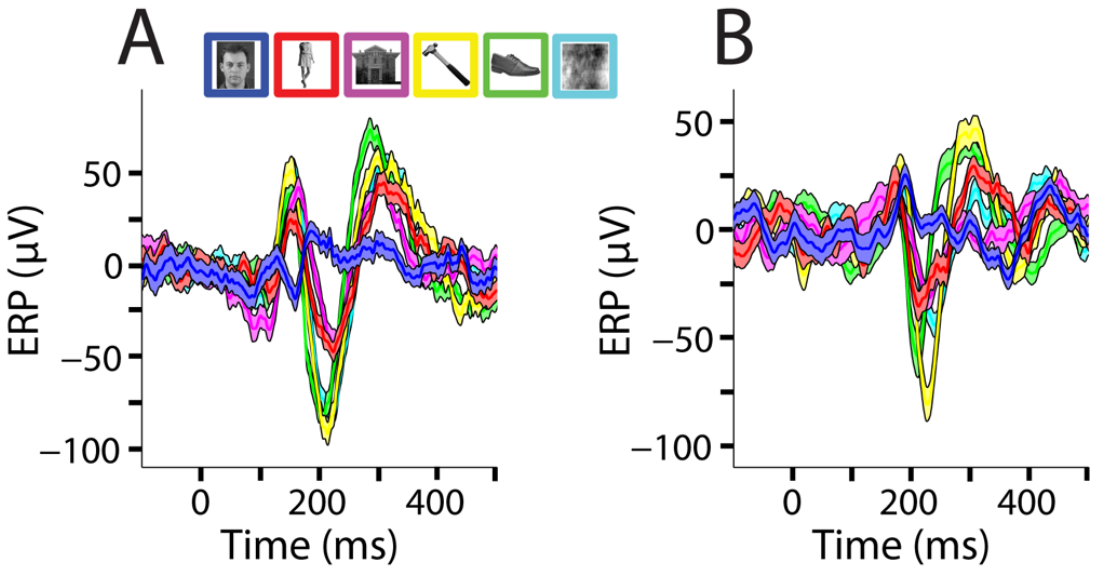


Figure 3.10: **ERPs from electrodes with significant d' due to faces showing less activity than other categories.** (A) ERP from electrode in the second column of P1 in Figure 3.7. (B) ERP from electrode in the first column, top row of EP4 in Figure 3.9. These electrodes were excluded from the analyses in the main text because they were not deemed to be in FFA due to the lower amplitude signal for faces relative to other categories.

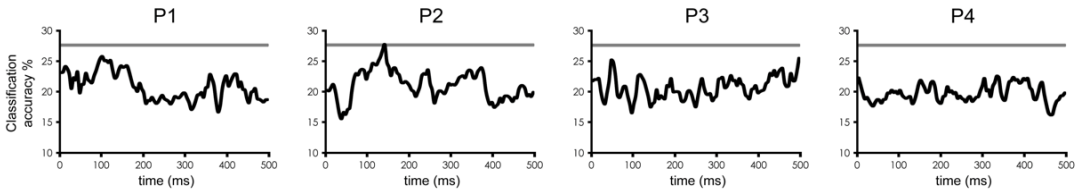


Figure 3.11: **Face expression classification.** Five-way classification accuracy for facial expressions (angry, fearful, sad, happy, and neutral) over time in experiment 2. Grey line indicates $p < .05$ corrected for multiple comparisons based on the permutation test.

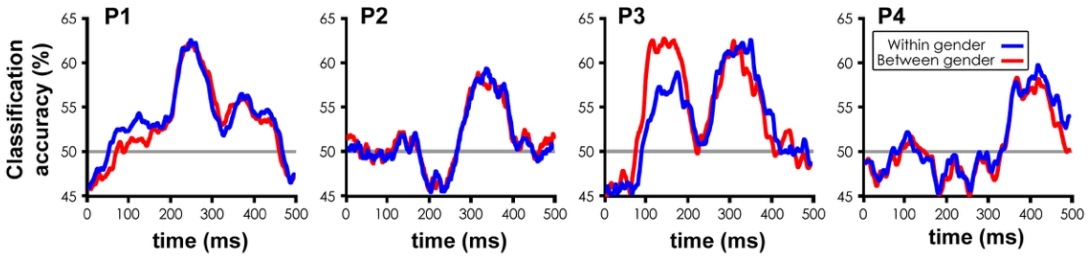


Figure 3.12: **Effects of task demands on face individuation.** Time course of individuation level face classification accuracy divided by within (blue) and across (red) gender classification in each participant. This shows, given two faces, how accurately we could predict which one the participant was viewing based on the neural data plotted against the beginning of the 100 ms sliding window. For within gender classification, all training and test faces were the same gender and for between gender classification, the two training faces were of different genders. If individuation was driven by task demands, only between gender classification would be greater than chance. The similarity between within and between gender classification suggests that individuation during the 200-500 ms time period was not driven by task demands.

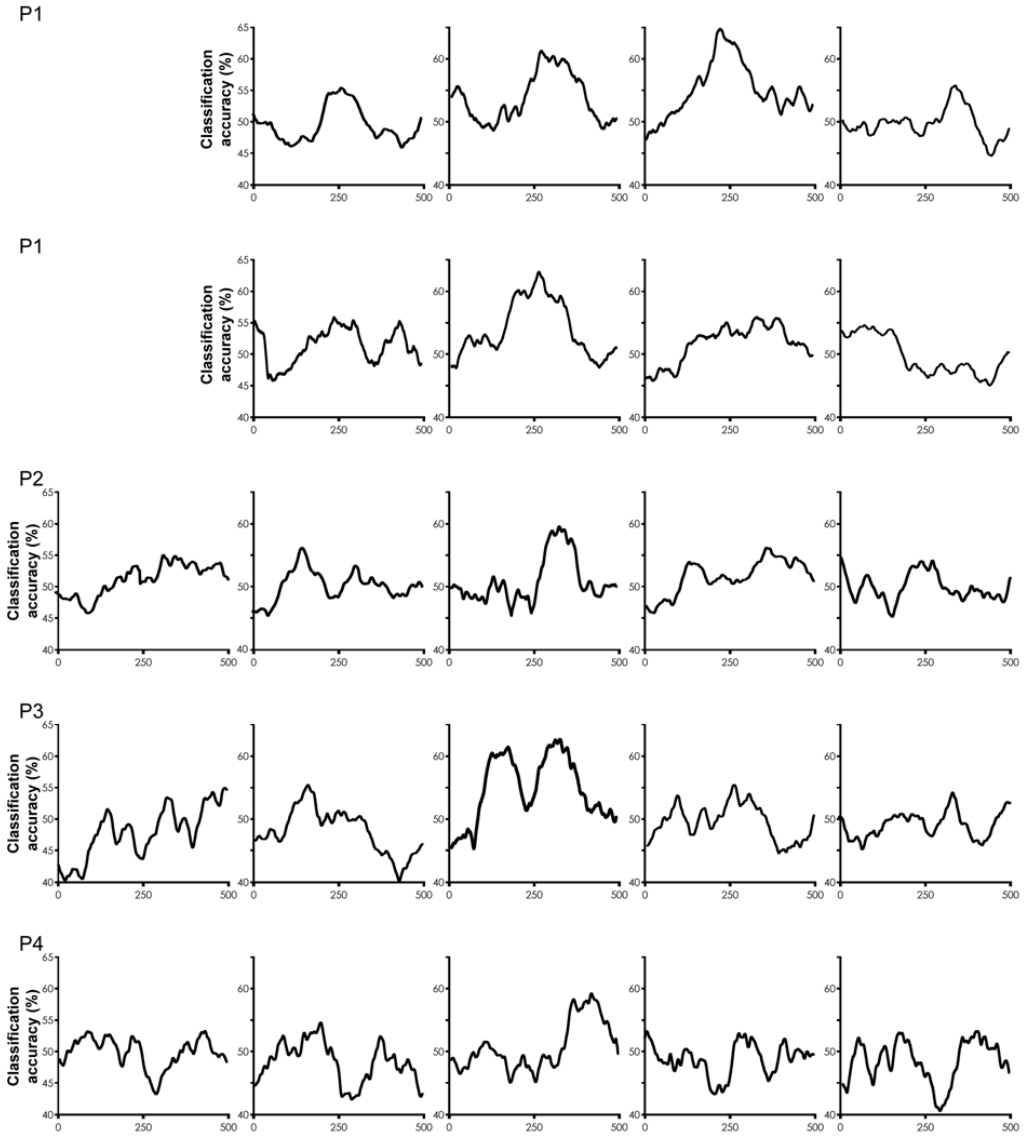


Figure 3.13: **Face individuation in all electrodes from P1-P4.** Time course of individual level face classification accuracy based on single trial voltage potentials in each participant. This shows, given two faces, how accurately we could predict which one the participant was viewing based on the neural data, plotted against the beginning of the 100 ms sliding window. $p = .05$, corrected for multiple time comparisons is at 57%. The layout of electrodes is the same as in Figure 3.7. In P1 for the analyses in the main text, the signals second and third electrode from the top row and the second electrode in the second row were averaged prior to classification, canonical correlation analysis, gamma power analysis, etc.

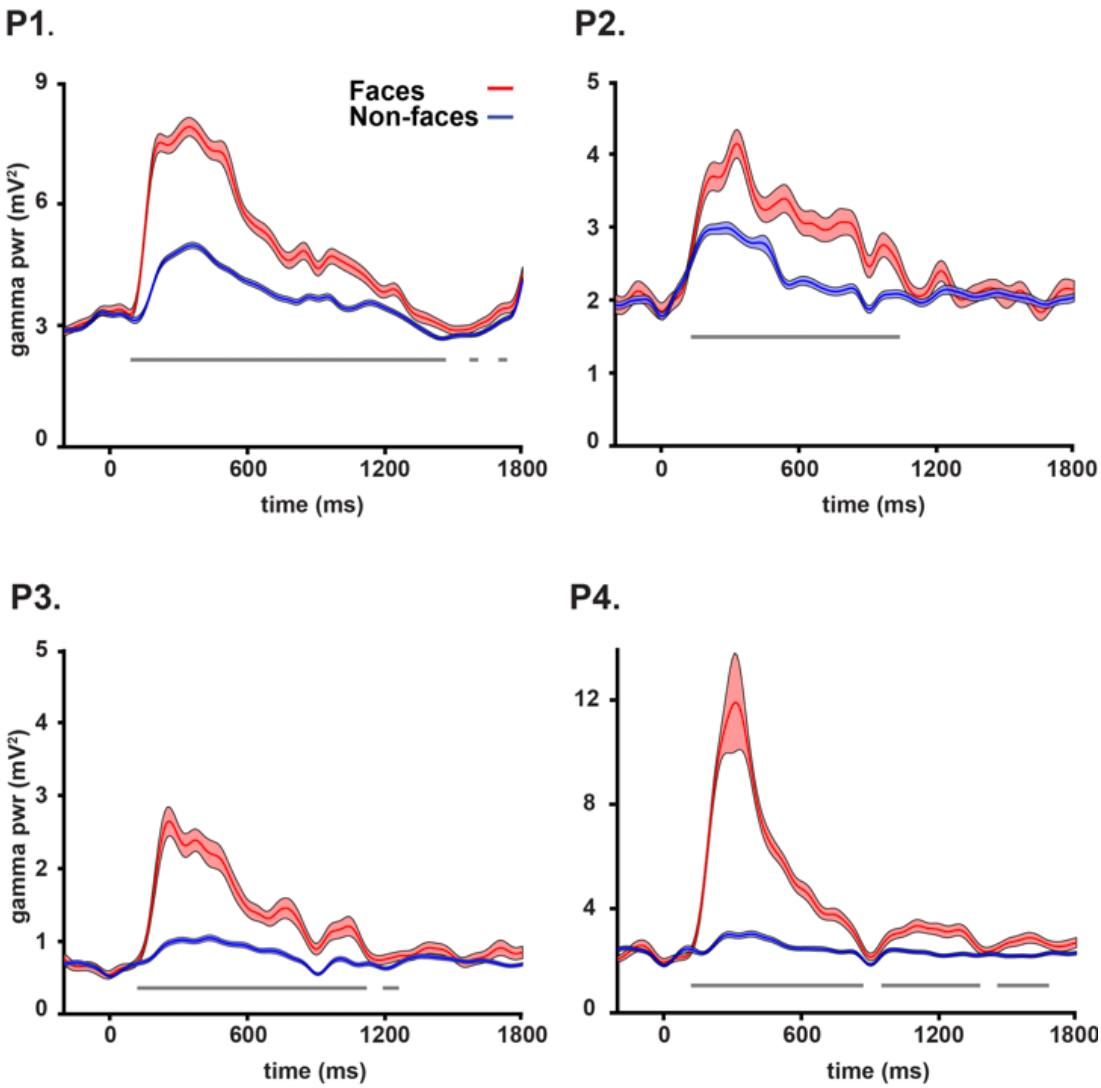


Figure 3.14: **Face specific gamma power in each participant.** Mean and standard error of gamma band (40-90 Hz) power for face and non-face trials in each participant in experiment 1. Grey bars indicate $p < .05$ using an across trial t-test between face and non-face objects.

Chapter 4

Spatiotemproal dynamics in human fusiform underlying facial expression perception

Though the fusiform is well-established as a key node in the face perception network, its role in facial expression processing remains unclear, due to competing models and discrepant findings. To help resolve this debate, we recorded from 17 subjects with intracranial electrodes implanted in face sensitive patches of the fusiform. Multivariate classification analysis showed that facial expression information is represented in fusiform activity, in the same regions that represent identity, though with a smaller effect size. Examination of the spatiotemporal dynamics revealed a functional distinction between posterior and mid-fusiform expression coding, with posterior fusiform showing an early peak of facial expression sensitivity at around 180 ms after subjects viewed a face and mid-fusiform showing a later and extended peak between 230 460 ms. These results support the hypothesis that the fusiform plays a role in facial expression perception and highlight a qualitative functional distinction between processing in posterior and mid-fusiform, with each contributing to temporally segregated stages of expression perception.

4.1 Introduction

Face perception, including detecting a face, recognizing face identity, assessing sex, age, emotion, attractiveness, and other characteristics associated with the face, is critical to social communication. An influential cognitive model of face processing distinguishes processes associated with recognizing the identity of a face from those associated with recognizing expression (Bruce and Young, 1986). A face sensitive region of the lateral fusiform gyrus, sometimes called the fusiform face area, is a critical node in the face processing network (Calder and Young, 2005; Duchaine and Yovel, 2015; Haxby et al., 2000; Ishai, 2008) that has been shown to be involved in identity perception (Barton, 2008; Barton et al., 2002; Ghuman et al., 2014; Goesaert and de Beeck, 2013; Nestor et al., 2011). What role, if any, the fusiform plays in face expression processing continues to be debated, particularly given the hypothesized cognitive distinction between identity and expression perception.

Results demonstrating relative insensitivity of the fusiform to face dynamics (Pitcher et al., 2011), reduced fusiform activity for attention to gaze direction (Hoffman and Haxby, 2000), and findings showing insensitivity of the fusiform to expression (Breiter et al., 1996; Foley et al., 2012; Streit et al., 1999; Thomas et al., 2001; Whalen et al., 1998) led to a model that proposed that this area was involved strictly in identity perception and not expression processing (Haxby et al., 2000). This model provided neuroscientific grounding for the earlier cognitive model that hypothesized a strong division between identity and expression perception (Bruce and Young, 1986). Recently, imaging studies have increasingly suggested that fusiform is involved in expression coding (Achaibou et al., 2015; Bishop et al., 2015; Fox et al., 2009; Ganel et al., 2005; Vuilleumier et al., 2001; Xu and Biederman, 2010). Positive findings for fusiform sensitivity to expression have led to the competing hypothesis that the division of face processing is not for identity and expression, but rather form/structure and motion (Duchaine and Yovel, 2015). However, mixed results have been reported in examining whether the same patches of the fusiform that code for identity also code for expression (see Zhang et al. (2016) for a study that examined both, but saw negative results for expression coding). Furthermore, some studies show the fusiform has an expression-independent identity code (Ghuman et al., 2014; Nestor et al., 2011). Taken together, prior results provide some, but not unequivocal, evidence for a role of the fusiform in expression processing.

Beyond whether the fusiform responds differentially to expression, one key question is whether the fusiform intrinsically codes for expression or if differential responses are due to task-related and/or top-down modulation of fusiform activity (Dehaene and Cohen, 2011). Assessing this requires a method with high temporal resolution to distinguish between early, more bottom-up biased activity, and later activity that likely involved recurrent interactions. Furthermore, a passive viewing or incidental task is required to exclude biases introduced by variable task demands across stimuli. The low temporal resolution of fMRI makes it difficult to disentangle early bottom-up processing from later top-down and recurrent processing. Some previous intracranial electroencephalography (iEEG) studies have used an explicit expression identification task, making task effects difficult to exclude (Kawasaki et al., 2012; Müsch et al., 2014; Tsuchiya et al., 2008). Those that have used an implicit task have shown mixed results regarding whether early fusiform response is sensitive to expression (Müsch et al., 2014; Pourtois et al., 2010). Furthermore, iEEG studies often lack sufficient subjects and population-level analysis to allow for a generalizable interpretation.

To help mediate between these two models and clarify the role of the fusiform in facial expression perception, iEEG was recorded from 17 subjects with a total of 31 face sensitive electrodes in face sensitive patches of the fusiform gyrus while these subjects viewed faces with neutral, happy, sad, angry, and fearful expressions in a gender discrimination task. Multivariate temporal pattern analysis (MTPA) on the data from these electrodes was used to analyze the temporal dynamics of neural activity with respect to facial expression sensitivity in fusiform. In a subset of 7 subjects, identity coding was examined in the same electrodes also using MTPA. In addition to examining the overall patterns across all electrodes, the responses from mid- and posterior fusiform, as well as the left and right hemisphere, were compared. To supplement these iEEG results, a meta-analysis of 64 neuroimaging studies was done examining facial expression sensitivity in the fusiform. The results support the view that fusiform response is sensitive to facial expression and suggest that the posterior and mid-fusiform regions play qualitatively

different roles in facial expression processing.

4.2 Methods

4.2.1 Participants

The experimental protocols were approved by the Institutional Review Board of the University of Pittsburgh. Written informed consent was obtained from all participants.

17 human subjects (8 male, 9 female) underwent surgical placement of subdural electrocorticographic electrodes or stereoelectroencephalography (together electrocorticography and stereoelectroencephalography are referred to here as iEEG) as standard of care for seizure onset zone localization. The ages of the subjects ranged from 19 to 65 years old (mean = 37.9, SD = 12.7). None of the participants showed evidence of epileptic activity on the fusiform electrodes used in this study nor any ictal events during experimental sessions.

4.2.2 Experiment design

In this study, each subject participated in two experiments. Experiment 1 was a functional localizer experiment and Experiment 2 was a face perception experiment. The experimental paradigms and the data pre-processing methods were similar to those described previously by Ghuman et al. (2014).

Stimuli

In Experiment 1, 180 images of faces (50% male), bodies (50% male), words, hammers, houses, and phase scrambled faces were used as visual stimuli. Each of the six categories contained 30 images. Phase scrambled faces were created in MATLAB^{TM} by taking the 2-dimensional spatial Fourier spectrum of each of the face images, extracting the phase, adding random phases, recombining the phase and amplitude, and taking the inverse 2-dimensional spatial Fourier spectrum.

In Experiment 2, face stimuli were taken from the Karolinska Directed Emotional Faces stimulus set (Lundqvist et al., 1998). Frontal views and 5 different facial expressions (fearful, angry, happy, sad, and neutral) from 70 faces (50% male) in the database were used, which yielded a total of 350 unique images. A short version of Experiment 2 used a subset of 40 faces (50% male) from the same database, which yielded a total of 200 unique images. 4 subjects participated in the long version of the experiment, and all other subjects participated in the short version of the experiment.

Paradigms

In Experiment 1, each image was presented for 900 ms with 900 ms inter-trial interval during which a fixation cross was presented at the center of the screen ($\sim 10^\circ \times 10^\circ$ of visual angle). At random, 1/3 of the time an image would be repeated, which yielded 480 independent trials in each session. Participants were instructed to press a button on a button box when an image was repeated (1-back).

In Experiment 2, each face image was presented for 1500 ms with 500 ms inter-trial interval during which a fixation cross was presented at the center of the screen. This yielded 200 independent trials per session. Faces subtended approximately 5 degrees of visual angle in width. Subjects were instructed to report whether the face was male or female via button press on a button box.

Paradigms were programmed in MATLAB^{TM} using Psychtoolbox and custom written code. All stimuli were presented on an LCD computer screen placed approximately 150 cm from participants heads.

All of the participants performed one session of Experiment 1. 9 of the subjects performed one session of Experiment 2, and the other 8 participants performed two or more sessions of Experiment 2.

4.2.3 Data preprocessing

The electrophysiological activity was recorded using iEEG electrodes at 1000 Hz. Single-trial potential signal was extracted by band-passing filtering the raw data between 0.2-115 Hz using a fourth order Butterworth filter to remove slow and linear drift, and high frequency noise. The 60 Hz line noise was removed using a fourth order Butterworth filter with 55-65 Hz stop-band. Power spectrum density (PSD) at 2-100 Hz with bin size of 2 Hz and time-step size of 10 ms was estimated for each trial using multi-taper power spectrum analysis with Hann tapers, using FieldTrip toolbox (Oostenveld et al., 2011). For each channel, the neural activity between 50 and 300 ms prior to stimulus onset was used as baseline, and the PSD at each frequency was then z-scored with respect to the mean and variance of the baseline activity to correct for the power scaling over frequency at each channel. The broadband signal was extracted as mean z-scored PSD across 40-100 Hz. Event-related potential (ERP) and event-related broadband signal (ERBB), both time-locked to the onset of stimulus from each trial, were used in the following data analysis. Specifically, the ERP signal is sampled at 1000 Hz and the ERBB is sampled at 100 Hz.

To reduce potential artifacts in the data, raw data were inspected for ictal events, and none were found during experimental recordings. Trials with maximum amplitude 5 standard deviations above the mean across all the trials were eliminated. In addition, trials with a change of more than 25 μV between consecutive sampling points were eliminated. These criteria resulted in the elimination of less than 1% of trials.

4.2.4 Electrode localization

Coregistration of grid electrodes and electrode strips was adapted from the method of Hermes et al. (2010). Electrode contacts were segmented from high resolution post-operative CT scans of patients coregistered with anatomical MRI scans before neurosurgery and electrode implantation. The Hermes method accounts for shifts in electrode location due to the deformation of the cortex by utilizing reconstructions of the cortical surface with FreeSurferTM software and co-registering these reconstructions with a high-resolution post-operative CT scan. SEEG electrodes were localized with Brainstorm software (Tadel et al., 2011) using post-operative MRI co-registered with pre-operative MRI images.

4.2.5 Electrode selection

Face sensitive electrodes were selected based on both anatomical and functional constraints. Anatomical constraint was based upon the localization of the electrodes on the reconstruction using post-implantation MRI. In addition, multivariate temporal pattern analysis (MTPA) was used to functionally select the electrodes that showed sensitivity to faces, comparing to other conditions in the localizer experiment (see below for MTPA details). Specifically, three criteria were used to screen and select the electrodes of interest: (1) electrodes of interest were restricted to those that were located in the mid-fusiform sulcus, on the fusiform gyrus, or in the sulci adjacent to fusiform gyrus; (2) electrodes were selected such that their peak 6-way classification d' score for faces (see below for how this was calculated) exceeded 0.5 ($p < 0.01$ based on a permutation test, as described below); (3) electrodes were selected such that the peak amplitude of the mean ERP and/or mean ERBB for faces was larger than the peak of mean ERP and/or ERBB for the other non-face object categories in the time window of 0 500 ms after stimulus onset. Dual functional criteria are used because criterion (2) insures only that faces give rise to statistically different activity from other categories, but not necessarily activity that is greater in magnitude. Combining criteria (2) and (3) insures that face activity is both statistically significantly different from other categories and greater magnitude in the electrodes of interest.

4.2.6 Multivariate temporal pattern analysis (MTPA)

Multivariate methods were used instead of traditional univariate statistics because of their superior sensitivity Ghuman et al. (2014); Haxby et al. (2014); Hirshorn et al. (2016). In this study, MTPA was applied to decode the coding of stimulus condition in the recorded neural activity. The timecourse of the decoding accuracy was estimated by classification using a sliding time window of 100 ms. Previous studies have demonstrated that both the low-frequency and the high-frequency neural activity contribute to the coding of facial information (Furl et al., 2017; Ghuman et al., 2014; Miller et al., 2016), therefore, both ERP and ERBB signals in the time window are combined as input features for the MTPA classifier. According to our preprocessing protocol, the ERP signal is sampled at 1000 Hz and the ERBB is sampled at 100 Hz, which yields 110 temporal features in each 100 ms time window (100 voltage potentials for ERP and 10 normalized mean power-spectrum density for ERBB). The 110 dimensional data were then used as input for the classifier. The goal of the classifier was to learn the patterns of the data distributions in such 110-dimensional space for different conditions and to decode the conditions of the corresponding stimuli from the testing trials. The classifier was trained on each electrode of each subject separately to assess the electrode sensitivity to faces and facial expressions. For Experiment 1, it was a 6-way classification problem and we specifically focused on the sensitivity of face category against other non-face categories. Therefore, we used the sensitivity index (d') for face category against all other non-face category as the metric of face sensitivity. d' was calculated as $d' = \Psi^{-1}(\text{true positive rate}) - \Psi^{-1}(\text{false alarm rate})$ where $\Psi^{-1}(x)$ is the inverse of the Gaussian cumulative distribution function. d' was used because it is an unbiased measure of effect size and one that takes into both the true positive and false positive rates. It also has the advantage that it is an effect size measure that has similar interpretation as Cohen's d (Cohen, 1988; Sawilowsky, 2009) while also being applicable to multivariate classification. In addition,

we provide full receiver-operator characteristic (ROC) curves for completeness and as validation of d' values. For Experiment 2, averaged pair-wise classification between every possible pair of facial expressions (10 pairs in total) was used.

The choice of the classifier is an empirical problem. The performance of the classifier depends on whether the assumptions of the classifier approximate the underlying truth of the data. Additionally, the complexity of the model and the size of the dataset affect performance (bias-variance trade-off). In this study, we employed Naïve Bayes (NB) classifiers, which assumes that each of the input features are conditionally independent from one another, and are Gaussian distributed. The classification accuracy of the classifier was estimated through 5-fold cross-validation. Specifically, all the trials were randomly and evenly spited into five folds. In each cross-validation loop, the classifier was trained based on four folds and the performance was evaluated on the left out fold. The overall performance was estimated by averaging across all the 5 cross-validation loops. In general, different classifiers gave similar results. Specifically, we evaluated the performance of different classifiers (NB, support vector machines, and random forests) on a small subset of the data, and NB classifier tended to perform better than other commonly used classifiers in the current experiment, but other classifiers also gave similar results. In addition, our previous experience (Hirshorn et al., 2016) with similar datasets also suggested that NB performed reasonably well in such classification analysis. We therefore used NB throughout the work presented here. The advantage of the Naïve Bayes classifier in the current study is likely due to intrinsic properties of the high dimensional problem (Bickel and Levina, 2004) that make a high-bias low-variance classifier (i.e. NB classifier) preferable compared to the low-bias high-variance classifiers (i.e. support vector machines).

4.2.7 Permutation testing

Permutation testing was used to determine the significance of the sensitivity index d'. For each permutation, the condition labels of all the trials were randomly permuted and the same procedure as described above was used to calculate the d' for each permutation. The permutation was repeated for a total of 1000 times. The d' of each permutation was used as the test statistic and the null distribution of the test statistic was estimated using the histogram of the permutation test.

4.2.8 K-means clustering

K-means clustering was used to cluster the electrodes in to groups based on both functional and anatomical features (Kaufman and Rousseeuw, 2009). Specifically, we applied k-means clustering algorithm to the electrodes in a 2D feature space of MNI y-coordinate and the peak classification accuracy time. Note that each dimension was normalized through z-scoring in order to account for different scales in space and time. Silhouette coefficients were used to evaluate the performance of models with different values of k (Kaufman and Rousseeuw, 2009). For each point x, the Silhouette coefficient is defined as $(b-a)/\max(a,b)$, where a is the mean intra-cluster distance from x to each points in the same cluster, and b is the mean inter-cluster distance from x to each points in the nearest cluster that x does not belong to.

4.2.9 Clustering analysis

We applied k-means clustering to the electrodes in the 2D space of MNI y-coordinate and peak classification time for facial expressions, with different values of k, and evaluated the model performance by computing the Bayes information criterion (BIC) and the mean Silhouette coefficient (SC) across all points.

Following (Kass and Wasserman, 1995; Pelleg et al., 2000), the BIC was estimated using Schwartz criterion. Specifically,

$$BIC = l(D|\hat{\theta}) - \frac{p}{2}\log N$$

, where $l(D|\hat{\theta})$ is the log-likelihood of the data under the assumption of k-means (spherical Gaussian) taken at the maximum likelihood estimation of parameters $\hat{\theta}$, p is the total number of parameters in the model, and N is the total number of data points.

Following (Kaufman and Rousseeuw, 2009), the Silhouette value for the i-th point was computed as $S_i = (b_i - a_i)/\max(a_i, b_i)$, where a_i is the average within cluster distance for the i-th point, and b_i is the minimum average between cluster distance for the i-th point (minimized over all other clusters). The mean SC was then estimated by averaging the Silhouette value over all data points.

4.2.10 Facial feature analysis

The facial features from the stimulus images were extracted following the similar process as (Ghuman et al., 2014). Anatomical landmarks for each picture were first determined by IntraFace (Xiong and De la Torre, 2013), which marks 49 points on the face along the eyebrows, down the bridge of the nose, along the base of the nose, and outlining the eyes and mouth. Based on these landmarks we calculated 17 facial feature dimensions listed in Table 4.4. The values for these 17 feature dimensions were normalized by subtracting the mean and dividing by the standard deviation across the all the pictures. The mean representation of each expression in facial feature space was computed by averaging across all 70 faces of the same expression.

4.2.11 Representational similarity analysis (RSA)

RSA was used to analyze the neural representational space for expressions (Kriegeskorte and Kievit, 2013). With pair-wised classification accuracy between each pair of facial expressions, we constructed the representational dissimilarity matrix (RDM) of the neural representation of facial expressions, with the element in the i-th column of the j-th row in the matrix corresponding to the pairwise classification accuracy between the i-th and j-th facial expressions. The corresponding RDM in the facial feature space was constructed by assessing the Euclidean distance between the vectors for the i-th and the j-th facial expressions averaged over all identities in the 17-dimensional facial feature space (Figure 4.7 top left).

4.2.12 Meta-analysis

Activation likelihood estimation (ALE, (Eickhoff et al., 2012; Laird et al., 2005)) was used for the meta-analysis of the neuroimaging literature. We first searched the online database of neuroimaging studies on Neurosynth.org and found around 300 imaging studies with the keyword "facial expressions". We then further narrowed the list down to 64 fMRI by only including the studies that had a direct full brain mapping by contrasting between emotional facial expressions, e.g. fear vs neutral, happy vs sad, etc. We only took into account the reported activation foci for the contrast between facial expressions. Then all of the activation foci in those relevant full brain map results were collected and extracted as 3D coordinates in MNI space. In the ALE, each of the extracted foci was assigned as the center of a Gaussian distribution, whose variance was scaled by the number of subjects in the corresponding experiment. These Gaussian distributions were then combined to build a full brain map of ALE. The ALE map was corrected for multiple comparison using cluster-based permutation test. Then we performed a spatial permutation test with 1000 permutations to construct a null distribution of the full brain activation. The ALE and the corresponding statistical analysis were performed based on GingerALE 2.3.6 (Eickhoff et al., 2009; Turkeltaub et al., 2012).

4.3 Results

4.3.1 Electrode selection and face sensitivity

The locations of the 31 fusiform electrodes from 17 participants sensitive to faces are shown in Figure 4.1A and Table 1. The averaged event-related potential (ERP) and event-related broadband gamma activity (ERBB) responses (see Methods for detailed definitions of ERP and ERBB) for each category across all channels are shown in Figure 4.1C and Figure 4.1D respectively. The averaged sensitivity index (d') for faces peaked at 160 ms ($d' = 1.22$, $p < 0.01$ in every channel, Figure 4.1B). Consistent with previous findings (Allison et al., 1999; Eimer, 2000c, 2011; Ghuman et al., 2014), strong sensitivity for faces was observed in the fusiform around 100-400 ms after stimulus onset.

4.3.2 Facial expression classification at group level

For each participant, the classification accuracy between each pair of facial expressions was estimated using 5-fold cross-validation (see Methods for details). As shown in Figure 4.2B, the averaged timecourse peaked at 190 ms after stimulus onset (average decoding at peak $d' = 0.12$, $p < 0.05$, Bonferroni corrected for multiple comparisons). In addition to the grand average, on the single electrode level, 21 out of the 31 electrodes from 12 out of 17 subjects showed a significant peak in their individual timecourses ($p < 0.05$, permutation test corrected for multiple comparisons). The locations of the significant electrodes are shown in Figure 4.2A and all electrodes are listed in Table 4.4.

The effect size for the mean peak expression classification is relatively low. This is in part because the electrodes consisted of two distinct populations with different timecourses (see below). Additionally, due to the variability in electrode position, iEEG effect sizes can be lower

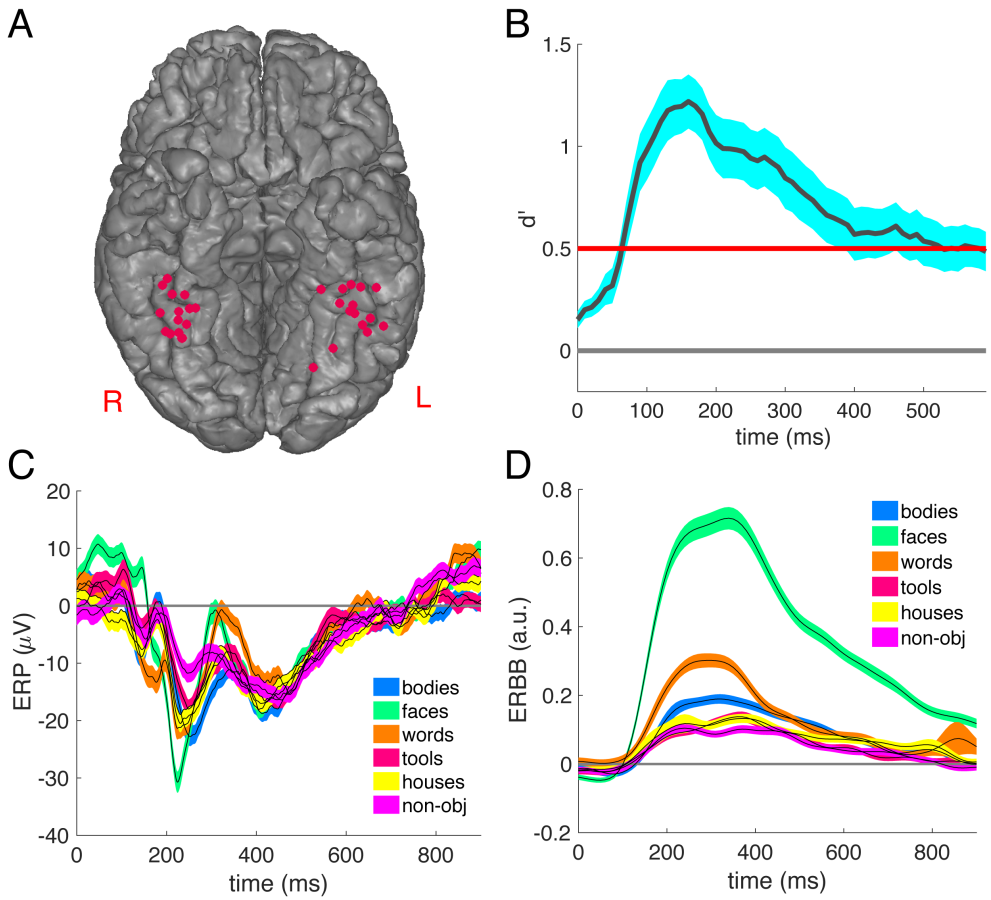


Figure 4.1: **The face sensitive electrodes in the fusiform. A)** The localization of the 31 face sensitive electrodes in (or close to) fusiform area, mapped onto a common space based on MNI coordinates. We moved depth electrode locations to the nearest location on the overlying cortical surface, in order to visualize all the electrodes. **B)** The timecourse of the sensitivity index (d') for faces versus the other categories in the six-way classification averaged across all 31 fusiform electrodes. The shaded areas indicate standard error of the mean across electrodes. The red line corresponds to $p < 0.01$ with Bonferroni correction for multiple comparisons across 60 time points. **C)** The ERP for each category averaged across all face sensitive fusiform electrodes. The shaded areas indicate standard error of the mean. **D)** The ERBB for each category averaged across all face sensitive fusiform electrodes. The shaded areas indicate standard error of the mean.

in some cases than what would be seen with electrodes optimally placed over face patches. To assess whether this was the case, we examined the correspondence between face category decoding and expression decoding based on the logic that placement closer to face patches should lead to higher face category decoding accuracy. A significant positive correlation between the decoding accuracy (d') for face category and the decoding accuracy (d') for facial expressions was seen (Pearson correlation $r = 0.57$, $N = 21$, $p = 0.007$). This suggests that electrode position relative to face patches in the fusiform can explain some of the effect size variability

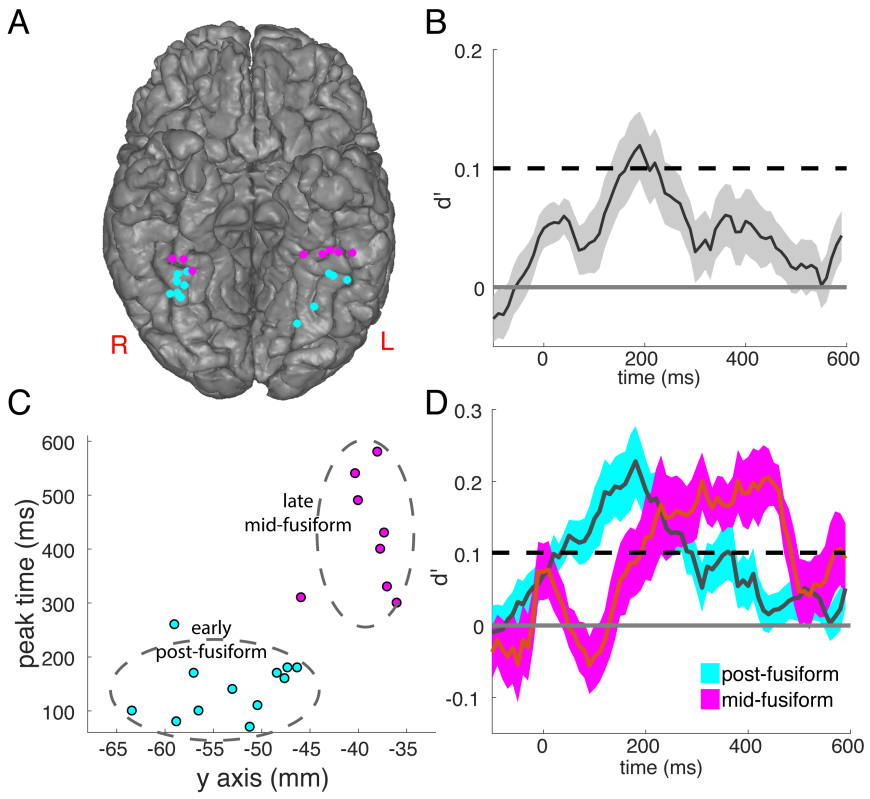


Figure 4.2: **The timecourse of the facial expression classification in fusiform. A)** The locations of the electrodes with significant face expression decoding accuracy, with the posterior fusiform group colored in cyan and the mid-fusiform group colored in magenta. **B)** The timecourse of mean and standard error for pairwise classification between different face expressions in all 31 fusiform electrodes. The shaded areas indicate standard error of the mean across electrodes. Dashed line: $p = 0.05$ threshold with Bonferroni correction for 60 time points [600 ms with 10 ms stepsize]) **C)** The time of the peak classification accuracy was plotted against the MNI y-coordinate for each single electrode with significant expression classification accuracy. K-means clustering partitions these electrodes into posterior and mid- fusiform groups. Dashed oval represent the 2-σ contour using the mean and standard deviation along the MNI x- and y-axes. **D)** The mean and standard error for pairwise classification between different face expressions in posterior fusiform electrodes and mid- fusiform electrodes. The posterior group peaked at 180 ms after stimulus onset and the mid-fusiform group had an extended peak starting at 230 ms and extending to 450 ms (both $p < 0.05$, binomial test, Bonferroni corrected; dashed line: $p = 0.05$ threshold with Bonferroni correction for 60 time points [600 ms with 10 ms stepsize]). The shaded areas indicate standard error of the mean across electrodes. See below for receiver operator characteristic (ROC) curves validating classification analysis (Figure 4.4).

for expression classification. That suggests the true effect size for expression classification for optimal electrode placement may be closer to what was seen for electrodes with higher accuracy (0.4-0.6, see Table 4.4) rather than the mean across all electrodes.

4.3.3 Spatiotemporal dynamics of facial expression decoding

The next question we addressed was whether spatiotemporal dynamics of facial expression representation in fusiform was location dependent. Specifically, we compared the dynamics of expression sensitivity between left and right hemispheres, and between posterior and mid- fusiform regions for electrodes showing significant expression sensitivity.

We first analyzed the lateralization effect for the expression coding in fusiform. The mean timecourses of decoding accuracy for left and right fusiform did not differ at the $p < 0.05$ uncorrected level at any time point (Figure 4.5).

In contrast substantial differences were seen in the timing and representation of expression coding between posterior and mid- fusiform. This was first illustrated by plotting the time of the peak decoding accuracy in each individual electrode against the corresponding MNI y-coordinate of the electrode (Figure 4.2C). A qualitative difference was seen between the peak times for electrodes posterior to approximately $y = -45$ compared to those anterior to that, rather than a continuous relationship between y-coordinate and peak time. This was quantified by a clustering analysis using both Bayesian information criterion (BIC) (Kass and Wasserman, 1995) and Silhouette analysis (Kaufman and Rousseeuw, 2009)(Figure 4.6), which both showed evidence for a cluster-structure in the data (Bayes factor > 20) with $k = 2$ as the optimal number of clusters (mean Sillhouette coefficient $= 0.59$). The 2 clusters corresponded to the posterior and mid-fusiform (Figure 4.2C). The border between these data-driven clusters corresponds well with prior functional and anatomical evidence showing that the mid-fusiform face patch falls within a 1 cm disk centered around the anterior tip of mid-fusiform sulcus (MFS; which falls at $y = -40$ in MNI coordinates) with high probability (Weiner et al., 2014). That would make the border between the mid-fusiform and posterior fusiform face patch approximately $y = -45$ in MNI coordinates, which is very close to the border produced by the clustering analysis ($y = -45.9$).

The timecourse of the posterior and mid- fusiform clusters were then examined in detail. As shown in Figure 4.2D, the timecourse of decoding accuracy in the posterior group peaked at 180 ms after stimulus onset and the timecourse of mid-fusiform group first peaked at 230 ms and the peak extended until approximately 450 ms after stimulus onset.

4.3.4 Comparison of the contributions from ERP and ERBB features to the classification

Here we compared the classification results using both ERP and ERBB vs using ERP or ERBB alone. As shown in Figure 4.3, both ERP and ERBB contributed to the expression decoding (left panel has higher d' than the other two panels). The posterior d' peak improves from 0.19 with only ERP features to 0.23 combining both ERP and ERBB features. The mid-fusiform d' peak improves from 0.19 with only ERP features to 0.21 combining both ERP and ERBB features. ERP features made greater contribution to the expression discrimination than ERBB (the middle panel has larger d' than the right panel, and the results in the middle panel are very close to results in the left panel). Due to the $1/f$ decay in the power spectrum, the ERP signal is dominated by low frequency components (mainly alpha and beta bands). This suggests that it is the low frequency components in ERP that mainly contributes to the facial expressions representation in the fusiform (Furl et al., 2017).

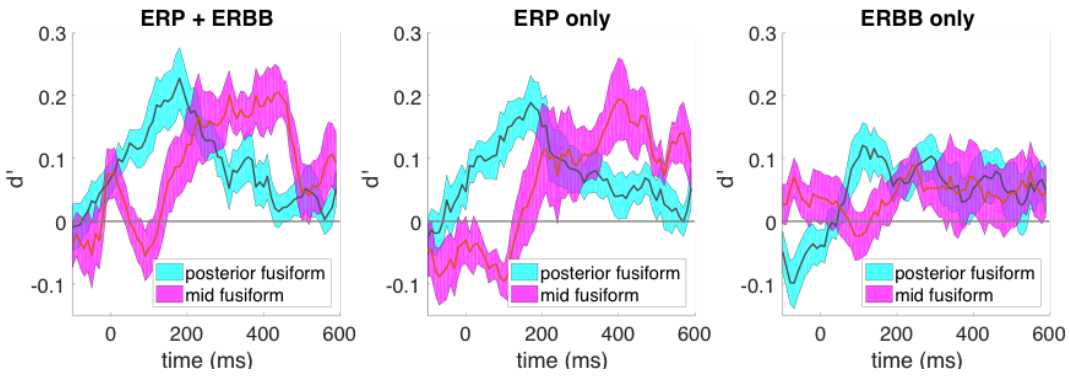
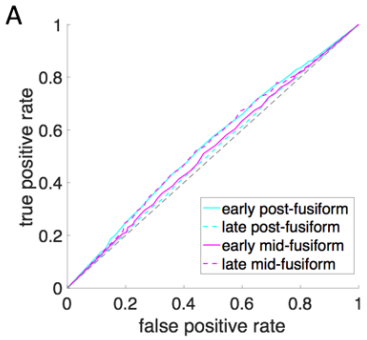


Figure 4.3: **Comparison of the contributions from ERP and ERBB features.** The mean and standard error for pairwise classification between different face expressions in posterior fusiform electrodes and mid- fusiform electrodes, using both ERP and ERBB features (left), using only ERP features (middle), and using only ERBB features (right).



B

AUC (mean ± s.e.)	Early (150 – 200 ms)	Late (400-450 ms)
Post-fusiform	0.54 ± 0.008	0.51 ± 0.005
Mid-fusiform	0.52 ± 0.008	0.54 ± 0.009

Figure 4.4: The mean ROC curve and area-under-curve (AUC) for posterior fusiform electrodes and mid-fusiform electrodes at early (150-200 ms after stim onset) and late stage (400-450 ms after stim onset).

4.3.5 Selection of models for k-means clustering

We applied k-means clustering to the electrodes in the 2D space of MNI y-coordinate and peak classification time for facial expressions, with different values of k, and evaluate the model performance by computing the Bayes information criterion (BIC) and the mean Silhouette coefficient (SC) across all points.

As shown in Figure 4.6, for $k = 1$, $BIC = -61.28$; for $k = 2$, $BIC = -54.63$. Therefore, Bayes factor between the hypothesis ($\mathbf{H}_1$) that there is a cluster structure ($k = 2$) and the null hypothesis ($\mathbf{H}_0$) that there is no cluster structure ($k = 1$) can be approximated as $BF = \exp\left((BIC_1 - BIC_0)/2\right)$. This approximation yields a $BF > 20$, which suggests a strong evidence of $\mathbf{H}_1$ over $\mathbf{H}_0$. In other words, there is a strong clustering structure in the data.

Moreover, for $k = 2$, $BIC = -54.63$, the mean $SC = 0.601$; for $k = 3$, $BIC = -56.29$, the mean $SC = 0.490$; for $k = 4$, $BIC = -58.54$, mean $SC = 0.428$. Both BIC and mean SC suggest that $k = 2$ is the optimal number of clusters. Therefore, $k = 2$ was used in the study.

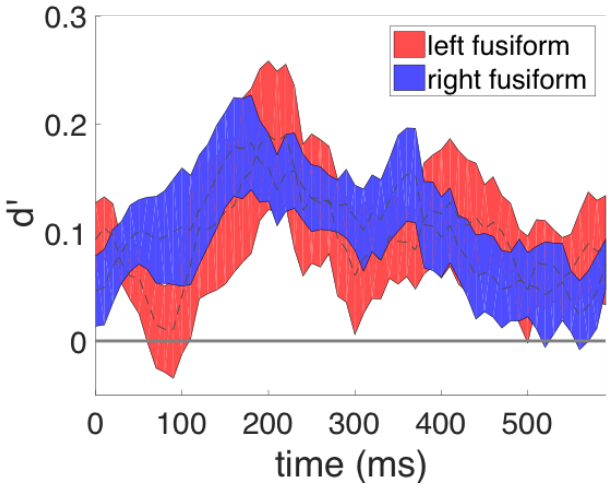


Figure 4.5: **The mean and standard error for classification between different face expressions in left and right fusiform electrodes.** The timecourse of the left fusiform peaked at 220 ms after stimulus onset with mean $d' = 0.19$, and the timecourse of the right fusiform peaked at 180 ms after stimulus onset with mean $d' = 0.18$ (both $p < 0.05$, binomial test, Bonferroni corrected).

4.3.6 Representational similarity analysis

A recent meta-analysis suggests that fusiform is particularly sensitive to the contrast between specific pairs of expressions (Vytal and Hamann, 2010). To examine this in iEEG data, the representation dissimilarity matrices (RDMs) for facial expressions in the early and late activity in posterior and mid-fusiform were computed Figure 4.7. No contrasts between expressions showed significant differences in posterior fusiform in the late window or in mid-fusiform in the early window ($p > 0.1$ in all cases, T-test), as expected due to the corresponding low overall classification accuracy. In the early posterior fusiform, expressions of negative emotions (fearful, angry) were dissimilar to happy and neutral expressions ($p < 0.05$ in each case, T-test), but not very distinguishable from one another. In the late mid-fusiform activity, happy and neutral expressions were both distinguishable from expressions of negative emotions and from each other ($p < 0.05$ in each case, T-test). The results showed partial consistency with a previous meta-analysis based on neuroimaging studies (consistent in angry vs. neutral, fearful vs. neutral, fearful vs. happy, and fear vs sad) (Vytal and Hamann, 2010). However, the previous meta-analysis also reported statistical significance for the contrasts of fearful vs. angry and angry vs. sad, which were absent in our results.

One question is the degree to which the representation in fusiform reflects the structural properties of the facial expressions subjects were viewing. To examine this question, an 17-dimensional facial feature space was constructed based on a computer vision algorithm (Xiong and De la Torre, 2013). The features characterize structural and spatial frequency properties of each image, e.g. eye width, eyebrow length, nose height, eye-mouth width ratio, skin tone, etc. An RDM was then built between the expressions in this feature space and compared to the neural feature spaces. There was a significant correlation between posterior fusiform representation

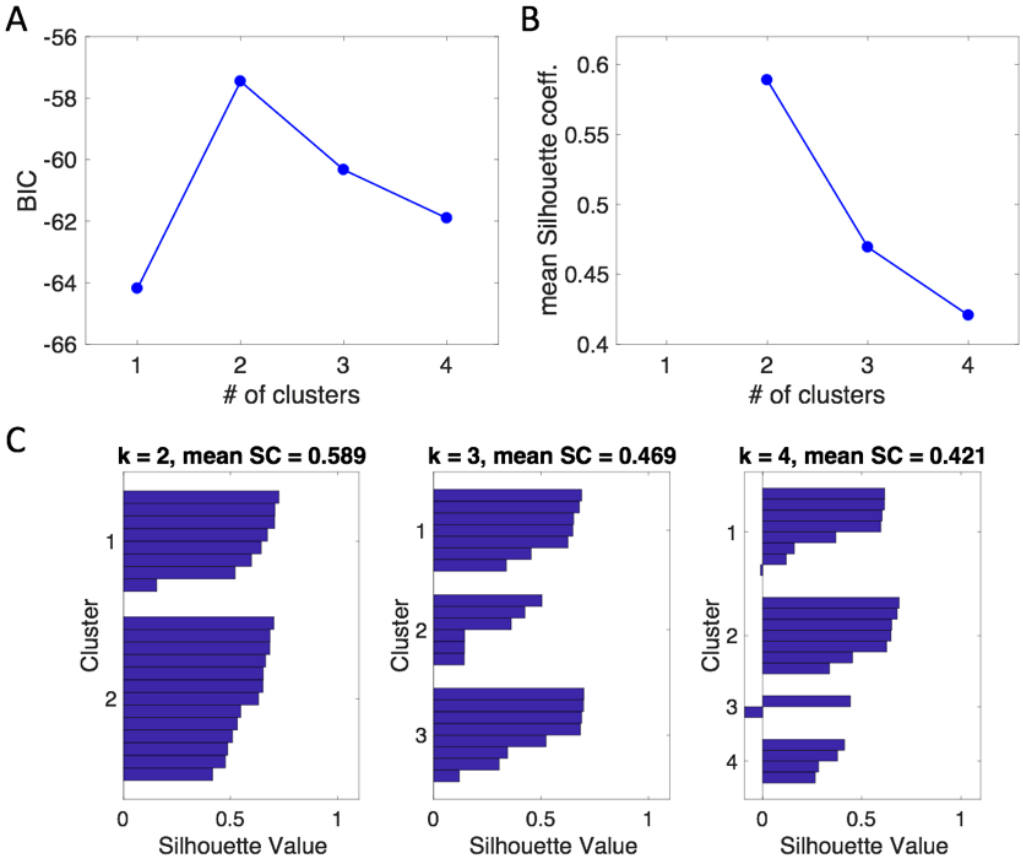


Figure 4.6: **Clustering analysis.** **A**) BIC of k-means models with different values of k ($k = 1, 2, 3, 4$). **B**) Mean SC of k-means models with different values of k ($k = 2, 3, 4$, note that SC is not applicable for $k = 1$). **C**) The distribution of Silhouette Coefficients (SC) with different values of k in k-means clustering. From left to right, $k = 2, 3$, and 4.

space in the early time window (Spearman's rho $= 0.24$, $p < 0.05$, permutation test). The correlation between mid- fusiform representation space in the late time window and the facial feature space was smaller and did not reach statistical significance (Spearman's rho $= 0.15$, $p > 0.1$, permutation test).

4.3.7 Comparison to facial identity classification

Given the strongly supported hypothesis the fusiform plays a central role in face identity recognition, the effect size of identity and expression coding in the fusiform was compared. Due to the relatively few repetitions of individual faces, individuation was examined in only the 7 subjects that had sufficient repetitions of each face identity allowing for multivariate classification of identity across expression; identity decoding was previously reported for 4 of these subjects in a recent study (Ghuman et al., 2014). Across the 7 total subjects (3 here and 4 reported previously), the mean peak $d' = 0.50$ for face identity decoding was significantly greater than the

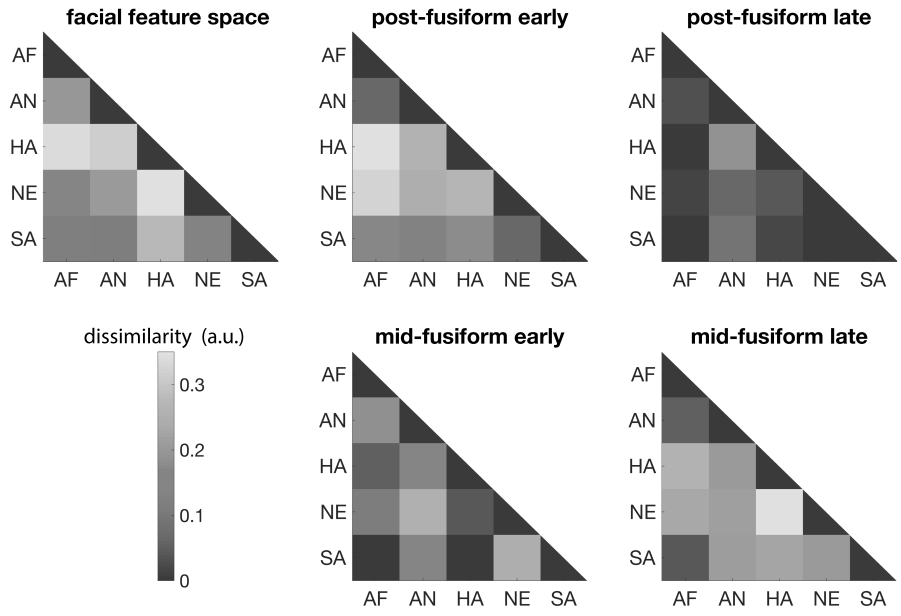


Figure 4.7: **Representational similarity analysis (RSA) between the facial feature space and the representational spaces of posterior and mid- fusiform at both early and late stages. Top row**: representational dissimilarity matrices (RDM) of facial expressions in the facial feature space (left), RDM of posterior fusiform at early stage (middle), RDM of posterior fusiform at late stage (right). **Bottom row**: RDM of mid-fusiform at early stage (middle), RDM of mid-fusiform at late stage (right). Abbreviations: AF fearful, AN angry, HA happy, NE neutral, SA sad.

mean peak accuracy for facial expression decoding in the exact same set of electrodes (mean peak $d' = 0.20$; $t(6) = 3.7821$, $p = 0.0092$). With regards to the timing of identity (mean peak time $= 314$ ms) versus expression sensitivity, the posterior peak time for expression classification was significantly earlier than the peak time for identity ($t(18) = 4.45$, $p = 0.0003$). The mid-fusiform extended peak time for expression classification overlapped with the peak time for identity.

4.3.8 Meta-analysis of the neuroimaging literature

In the broad neuroimaging literature, we found 64 fMRI studies with full brain contrasts between face expressions (See Table 4.4). Among the 64 studies, 24 studies report at least one significant focus of fusiform sensitivity to differences in expressions (See Figure 4.8 for activation map). A total of 999 significant foci were reported in those experiments for contrasts between different facial expressions (Figure 4.8). A full brain activation likelihood estimation (ALE) was performed and significance was assessed using a cluster-based permutation test. 4 significant clusters were found at the $p < 0.01$ threshold, none of which included the fusiform. The MNI coordinates for the center and the corresponding label names of the 4 clusters are shown in Table 4.4.

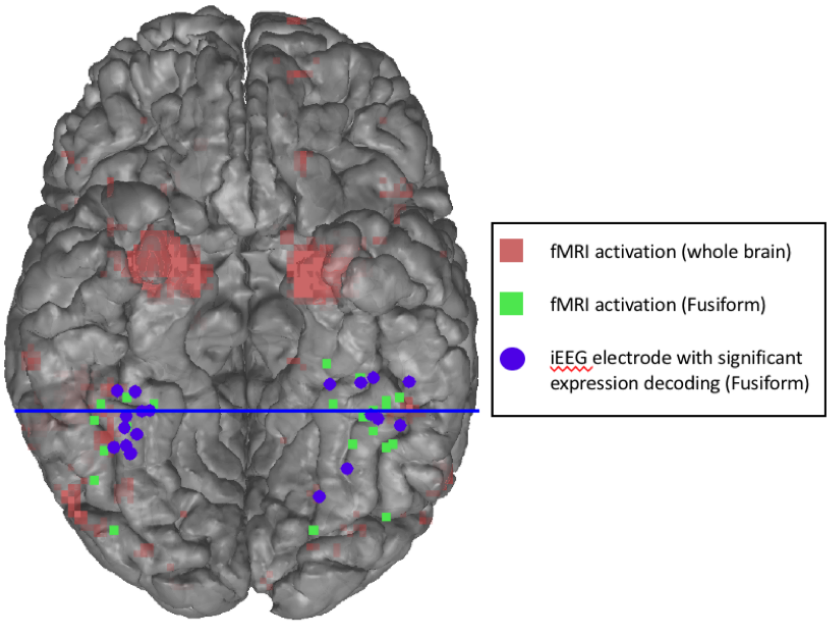


Figure 4.8: **Activation map for facial expressions.** (Red) Whole brain activation map from all 64 relevant fMRI studies. (Green square) Voxels in fusiform reported in 24/64 of the fMRI studies that have significant contrast between facial expressions. (Blue dots) iEEG electrodes in fusiform that have significant facial expression decoding. (Blue line) the border between posterior and mid-fusiform clusters based upon clustering analysis in the iEEG electrodes.

4.4 Discussion

Multivariate classification methods were used to evaluate the encoding of facial expressions recorded from electrodes placed directly in face sensitive fusiform cortex. Though the effect size for expression classification is smaller than for identity classification, the results support a role for the fusiform in the processing of facial expressions. Electrodes that were sensitive to expression were also sensitive to identity, suggesting a shared neural substrate for identity and expression coding in the fusiform. The results also show that the posterior and mid- fusiform are dynamically involved in distinct stages of facial expression processing and have different representations of expressions. The differential representation and magnitude of the temporal displacement between the sensitivity in posterior and mid-fusiform suggests these are qualitatively distinct stages of facial expression processing and not merely a consequence of transmission or information processing delay along a feedforward hierarchy.

Fusiform is sensitive to facial expression

The results here show that the fusiform is sensitive to expression, though the effect size for classification of expression in the fusiform using iEEG is small-to-medium (Cohen, 1988). The results also suggest that the same patches of the fusiform that are sensitive to expression are sensitive to identity as well. Given the variability of the effect size due to the proximity of elec-

trode placement relative to face patches, the relative effect size may be more informative than the absolute effect size. The magnitude for facial expression classification is approximately half what was seen for face identity classification. This suggests that while fusiform contributes to facial expression perception, it is to a lesser degree than face identity processing. Greater involvement in identity than expression perception is expected for a region involved in structural processing of faces because identity relies on this information more than expression because expression perception also relies on facial dynamics. These results support models that hypothesize fusiform involvement in form/structural processing, at least for posterior fusiform (see discussion on spatially and temporally segregated stages of processing below), which can support facial expression processing (Calder and Young, 2005; Duchaine and Yovel, 2015). These results do not support models that hypothesize a strong division between facial identity and expression processing (Bruce and Young, 1986; Haxby et al., 2000).

To test what a brain region codes for one must examine its response for early, bottom-up activation during an incidental task or passive viewing (Dehaene and Cohen, 2011), otherwise it is difficult to disentangle effects of task demands and top-down modulation. Indeed, previous studies have demonstrated that extended fusiform activity, particularly in the broadband gamma range, is modulated by task-related information (Engell and McCarthy, 2010; Ghuman et al., 2014). Some previous iEEG studies of expression coding in the fusiform have used an explicit expression judgment task and examined only broadband gamma activity, making it difficult to draw definitive conclusions about fusiform expression coding from these results (Kawasaki et al., 2012; Tsuchiya et al., 2008). One previous study that used an implicit task did not show evidence of expression sensitivity during the early stage of activity in the fusiform (Müsch et al., 2014); another did show evidence of expression sensitivity, though it reported results only from a single subject (Pourtois et al., 2010). The results here show in a large iEEG sample that the early response of the fusiform most sensitive to bottom-up processing is modulated by expression, at least for the posterior fusiform.

The effect size for facial expression classification is consistent with mixed findings in the neuroimaging literature for expression sensitivity in the fusiform (Harris et al., 2014; Harry et al., 2013; Haxby et al., 2000; Skerry and Saxe, 2014; Tsuchiya et al., 2008; Zhang et al., 2016). IEEG generally has greater sensitivity and lower noise than non-invasive measures of brain activity. Methods with lower sensitivity, such as fMRI, would be expected to have a substantial false negative rate for facial expression coding in the fusiform. To quantify fMRI sensitivity to expression we performed a meta-analysis on 64 studies. Of these studies, 24 reported at least one expression sensitive loci in the fusiform. However, at the meta-analytic level, no significant cluster of expression sensitivity was seen in the fusiform after whole brain analysis (see Table 4.4, and Figure 4.8). Thus, consistent with the iEEG effect size for expression decoding in the fusiform seen here, there is some suggestion in the fMRI literature for expression sensitivity in the fusiform, but it is relatively small in magnitude and does not achieve statistical significance at the whole brain level.

Multiple, spatially and temporally segregated stages of face expression processing in the fusiform

Using a data-driven analysis, posterior and mid-fusiform face patches were shown to contribute differentially to expression processing. The dividing point between post-fusiform and mid-fusiform electrodes found in a data-driven manner is consistent with the anatomical border for the posterior and mid-fusiform face patches previously described, suggesting a strong coupling between the anatomical and functional divisions in fusiform (Weiner et al., 2014). While posterior and mid-fusiform have been shown to be cytoarchitectonically distinct regions each with separate face sensitive patches (Freiwald and Tsao, 2010; Weiner et al., 2014, 2017), functional differences between these patches have remained elusive in the literature. The results here suggest that these anatomical and physiological distinctions correspond to functional distinctions in the role of these areas in face processing, as reflected in qualitatively different temporal dynamics in these regions for facial expression processing. Specifically, posterior fusiform participates in a relatively early stage of facial expression processing that may be related to structural encoding of faces. Mid-fusiform demonstrates a distinct pattern of extended dynamics and participates in a later stage of processing that may be related to a more abstract and/or multifaceted representation of expression and emotion. These results support the revised model of fusiform function that posits the fusiform contributes to structural encoding of facial expression during the initial stages of processing (Calder and Young, 2005; Duchaine and Yovel, 2015), with the notable addition that it may be primarily posterior fusiform contributes to structural processing.

The early time period of expression sensitivity in posterior fusiform overlaps with strong face sensitive activity measured non-invasively around 170 ms after viewing a face, which is thought to reflect structural encoding of face information (Bentin and Deouell, 2000; Blau et al., 2007; Eimer, 2000a,c, 2011). Face sensitive activity in this time window has been shown to be insensitive to attention and is thought to reflect a "rapid, feed-forward phase of face-selective processing." (Furey et al., 2006) Additionally, a face adaptation study showed that activity in this window reflects the actual facial expression rather than the perceived (adapted) expression (Furl et al., 2007). Consistent with these previous findings, the RSA results here show that the early posterior activity is significantly correlated to the physical/structural features of the face.

The expression sensitivity in mid-fusiform onset began later than the posterior fusiform (around 230 ms), and remained active until $\sim$450 ms after viewing a face. Face sensitive activity in this time window has been shown to be sensitive to face familiarity and to attention (Eimer, 2000b; Eimer et al., 2003). Previous studies and the results presented here show that face identity can be decoded from the activity in this later time window in mid-fusiform (Ghuman et al., 2014; Vida et al., 2017) and reflects a distributed code for identity among regions of the face processing network (Li et al., 2017). Thus, this later activity may relate to integration of multiple kinds of face information, such integration of identity and expression. Additionally, the previously mentioned face adaptation study showed that activity in this window reflects the subjectively perceived facial expression after adaptation (Furl et al., 2007). The RSA analysis here showed that the activity in this time window in mid-fusiform was not significantly correlated with physical similarity of the facial expressions. This lack of correlation with the physical features of the space, combined with the result that mid-fusiform activity does show significant expression decoding, suggests that the representation in mid-fusiform may reflect a more conceptual repre-

sentation of expression. Taken together, these results and prior findings suggest the mid-fusiform expression sensitivity in this later window reflect a more abstract and subjective representation of expression and may be related to integration of multiple face cues, including identity and expression. This abstract and multifaceted representation is likely to reflect processes arising from interactions across the face processing network (Ishai, 2008).

To conclude, the results presented here support the hypothesis that the fusiform contributes to expression processing (Calder and Young, 2005; Duchaine and Yovel, 2015). The finding that the same part of the fusiform is sensitive to both identity and expression contradicts models that hypothesize separate pathways for their processing (Bruce and Young, 1986; Haxby et al., 2000) and instead supports the hypothesis the fusiform supports structural encoding of faces in service of both identity and expression (Duchaine and Yovel, 2015). The results also show there is a qualitative distinction between face processing in posterior and mid-fusiform, with each contributing to temporally and functionally distinct stages of expression processing. This distinct contribution of these two fusiform patches suggest that the structural and cytoarchitectonic differences between posterior and mid-fusiform are associated with functional differences between the contributions of these areas to face perception. Taken together, the results here illustrate the dynamic role the fusiform plays in multiple stages of facial expression processing.

Table 4.1: MNI coordinates and facial expression sensitivity (d') for all face sensitive electrodes. Electrode ID is labeled by subject number (SX) and electrode from that subject (a, b, etc.). Sensitivity to expression defined as $p < 0.05$ decoding accuracy corrected for multiple comparisons.

Electrode ID	X (mm)	Y (mm)	Z (mm)	Peak time (ms)	Peak d'	Sensitive to expressions
S1a	35	-59	-22	260	0.29	Y
S1b	33	-53	-22	150	0.31	Y
S1c	42	-56	-26	200	0.20	N
S2a	40	-57	-23	170	0.34	Y
S3a	-33	-44	-31	580	0.18	N
S4a	-38	-36	-30	440	0.12	N
S5a	-38	-36	-20	300	0.25	Y
S5b	-42	-37	-19	330	0.25	Y
S6a	34	-40	-11	540	0.24	Y
S6b	39	-40	-10	490	0.33	Y
S7a	36	-57	-21	100	0.42	Y
S8a	-22	-72	-9	100	0.23	Y
S8b	-40	-48	-23	170	0.38	Y
S9a	32	-46	-7	180	0.34	Y
S9b	36	-48	-8	160	0.40	Y
S10a	29	-46	-15	310	0.31	Y
S11a	-25	-38	-17	580	0.36	Y
S11b	-34	-38	-18	400	0.46	Y
S11c	-49	-37	-20	430	0.27	Y
S12a	41	-33	-19	70	0.06	N
S12b	37	-51	-9	70	0.22	Y
S12c	35	-59	-4	80	0.23	Y
S13a	43	-36	-13	400	0.11	N
S13b	44	-48	-11	190	0.10	N
S14a	-52	-54	-17	30	0.14	N
S15a	-37	-47	-10	180	0.64	Y
S16a	-39	-45	-11	160	0.03	N
S17a	-43	-53	-26	90	0.20	N
S17b	-46	-50	-28	110	0.31	Y
S17c	-30	-63	-20	120	0.27	Y
S17d	-45	-56	-25	40	0.13	N

Table 4.4: A summary list for the 64 neuroimaging studies included in the meta-analysis (the 24 studies that report significant emotional sensitivity in fusiform are marked with *).

	title	authors	journal	year
1*	A common neural code for perceived and inferred emotion.	Skerry AE, Saxe R	Journal of neuroscience	2014

2	A left amygdala mediated network for rapid orienting to masked fearful faces.	Carlson JM, Reinke KS, Habib R	Neuropsychologia	2009
3	A neural network reflecting individual differences in cognitive processing of emotions during perceptual decision making.	Meriau K, Wartenburger I, Kazzer P, Prehn K, Lammers CH, van der Meer E, Villringer A, Heekeren HR	NeuroImage	2006
4	Affect-specific activation of shared networks for perception and execution of facial expressions.	Kircher T, Pohl A, Krach S, Thimm M, Schulte-Ruther M, Anders S, Mathiak K	Social cognitive and affective neuroscience	2013
5	Amygdala activation at 3T in response to human and avatar facial expressions of emotions.	Moser E, Derntl B, Robinson S, Fink B, Gur RC, Grammer K	Journal of neuroscience methods	2007
6	Amygdala integrates emotional expression and gaze direction in response to dynamic facial expressions.	Sato W, Kochiyama T, Uono S, Yoshikawa S	NeuroImage	2010
7	Amygdala reactivity predicts automatic negative evaluations for facial emotions.	Dannlowski U, Ohrmann P, Bauer J, Kugel H, Arolt V, Heindel W, Suslow T	Psychiatry research	2007
8	Amygdala response to facial expressions in children and adults.	Thomas KM, Drevets WC, Whalen PJ, Eccard CH, Dahl RE, Ryan ND, Casey BJ	Biological psychiatry	2001
9	Amygdala response to facial expressions reflects emotional learning.	Hooker CI, Germine LT, Knight RT, D'Esposito M	Journal of neuroscience	2006
10	Anxiety predicts a differential neural response to attended and unattended facial signals of anger and fear.	Ewbank MP, Lawrence AD, Passamonti L, Keane J, Peers PV, Calder AJ	NeuroImage	2009
11	Automatic emotion processing as a function of trait emotional awareness: an fMRI study.	Lichev V, Sacher J, Ihme K, Rosenberg N, Quirin M, Lepsien J, Pampel A, Rufer M, Grabe HJ, Kugel H, Kersting A, Villringer A, Lane RD, Suslow T	Social cognitive and affective neuroscience	2014
12	Beyond threat: amygdala reactivity across multiple expressions of facial affect.	Fitzgerald DA, Angstadt M, Jelsone LM, Nathan PJ, Phan KL	NeuroImage	2006

13*	Binding action and emotion in social understanding.	Ferri F, Ebisch SJ, Costantini M, Salone A, Arciero G, Mazzola V, Ferro FM, Romani GL, Gallese V	PloS one	2013
14*	Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust.	Wicker B, Keysers C, Plailly J, Royet JP, Gallese V, Rizzolatti G	Neuron	2003
15	Brain networks involved in haptic and visual identification of facial expressions of emotion: an fMRI study.	Kitada R, Johnsrude IS, Kochiyama T, Lederman SJ	NeuroImage	2010
16*	Brain responses to dynamic facial expressions of pain.	Simon D, Craig KD, Miltner WH, Rainville P	Pain	2006
17	Brain responses to facial expressions of pain: emotional or motor mirroring?	Budell L, Jackson P, Rainville P	NeuroImage	2010
18*	Cerebral integration of verbal and nonverbal emotional cues: impact of individual nonverbal dominance.	Jacob H, Kreifelts B, Bruck C, Erb M, Hosl F, Wildgruber D	NeuroImage	2012
19	Cerebral regulation of facial expressions of pain.	Kunz M, Chen JI, Lautenbacher S, Vachon-Presseau E, Rainville P	Journal of neuroscience	2011
20	Classification images reveal the information sensitivity of brain voxels in fMRI.	Smith FW, Muckli L, Brennan D, Pernet C, Smith ML, Belin P, Gosselin F, Hadley DM, Cavanagh J, Schyns PG	NeuroImage	2008
21	Converging evidence for the advantage of dynamic facial expressions.	Arsalidou M, Morris D, Taylor MJ	Brain topography	2011
22*	Decoding of affective facial expressions in the context of emotional situations.	Sommer M, Dohnel K, Meinhardt J, Hajak G	Neuropsychologia	2008
23	Dynamic facial expressions evoke distinct activation in the face perception network: a connectivity analysis study.	Foley E, Rippon G, Thai NJ, Longe O, Senior C	Journal of cognitive neuroscience	2012
24*	Dynamic stimuli demonstrate a categorical representation of facial expression in the amygdala.	Harris RJ, Young AW, Andrews TJ	Neuropsychologia	2014

25*	Emotions in motion: dynamic compared to static facial expressions of disgust and happiness reveal more widespread emotion-specific activations.	Trautmann SA, Fehr T, Herrmann M	Brain research	2009
26	Enhanced neural activity in response to dynamic facial expressions of emotion: an fMRI study.	Sato W, Kochiyama T, Yoshikawa S, Naito E, Matsumura M	Cognitive brain research	2004
27*	Facial emotion modulates the neural mechanisms responsible for short interval time perception.	Tipples J, Brattan V, Johnston P	Brain topography	2015
28*	Facial expression and gaze-direction in human superior temporal sulcus.	Engell AD, Haxby JV	Neuropsychologia	2007
29	Facial expressions and complex IAPS pictures: common and differential networks.	Britton JC, Taylor SF, Sudheimer KD, Liberzon I	NeuroImage	2006
30	Frontal lobe networks for effective processing of ambiguously expressed emotions in humans.	Nomura M, Iidaka T, Kakehi K, Tsukiura T, Hasegawa T, Maeda Y, Matsue Y	Neuroscience letters	2003
31	Functional imaging of face and hand imitation: towards a motor theory of empathy.	Leslie KR, Johnson-Frey SH, Grafton ST	NeuroImage	2004
32*	Functional neuroanatomy of perceiving surprised faces.	Schroeder U, Hennenlotter A, Erhard P, Haslinger B, Stahl R, Lange KW, Ceballos-Baumann AO	Human brain mapping	2004
33	Functional responses and structural connections of cortical areas for processing faces and voices in the superior temporal sulcus.	Ethofer T, Bretscher J, Wiethoff S, Bisch J, Schlipf S, Wildgruber D, Kreifelts B	NeuroImage	2013
34	Incongruence effects in crossmodal emotional integration.	Muller VI, Habel U, Derntl B, Schneider F, Zilles K, Turetsky BI, Eickhoff SB	NeuroImage	2011
35*	Integration of cross-modal emotional information in the human brain: an fMRI study.	Park JY, Gu BM, Kang DH, Shin YW, Choi CH, Lee JM, Kwon JS	Cortex	2010
36*	Investigating the brain basis of facial expression perception using multi-voxel pattern analysis.	Wegrzyn M, Riehle M, Labudda K, Woermann F, Baumgartner F, Pollmann S, Bien CG, Kissler J	Cortex	2015

37	Is a neutral expression also a neutral stimulus? A study with functional magnetic resonance.	Carvajal F, Rubio S, Serrano JM, Rios-Lago M, Alvarez-Linera J, Pacheco L, Martin P	Experimental brain research	2013
38*	Leaving a bad taste in your mouth but not in my insula.	von dem Hagen EA, Beaver JD, Ewbank MP, Keane J, Passamonti L, Lawrence AD, Calder AJ	Social cognitive and affective neuroscience	2009
39	Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge.	Whalen PJ, Rauch SL, Etcoff NL, McInerney SC, Lee MB, Jenike MA	Journal of neuroscience	1998
40	Mind your left: spatial bias in subcortical fear processing.	Siman-Tov T, Papo D, Gadoth N, Schonberg T, Mendelsohn A, Perry D, Hendler T	Journal of cognitive neuroscience	2009
41*	Multiple mechanisms of consciousness: the neural correlates of emotional awareness.	Amting JM, Greening SG, Mitchell DG	Journal of neuroscience	2010
42	Neural mechanism for judging the appropriateness of facial affect.	Kim JW, Kim JJ, Jeong BS, Ki SW, Im DM, Lee SJ, Lee HS	Cognitive brain research	2005
43*	Neural mechanism of unconscious perception of surprised facial expression.	Duan X, Dai Q, Gong Q, Chen H	NeuroImage	2010
44	Neural responses to ambiguity involve domain-general and domain-specific emotion processing systems.	Neta M, Kelley WM, Whalen PJ	Journal of cognitive neuroscience	2013
45*	Nonconscious emotional processing involves distinct neural pathways for pictures and videos.	Faivre N, Charron S, Roux P, Lehericy S, Kouider S	Neuropsychologia	2012
46*	Orbitofrontal and hippocampal contributions to memory for face-name associations: the rewarding power of a smile.	Tsukiura T, Cabeza R	Neuropsychologia	2008
47	Orbitofrontal Cortex Reactivity to Angry Facial Expression in a Social Interaction Correlates with Aggressive Behavior.	Beyer F, Munte TF, Gottlich M, Kramer UM	Cerebral cortex	2014
48	Positive facial affect - an fMRI study on the involvement of insula and amygdala.	Pohl A, Anders S, Schulte-Ruther M, Mathiak K, Kircher T	PloS one	2013

49	Preferential amygdala reactivity to the negative assessment of neutral faces.	Blasi G, Hariri AR, Alce G, Taurisano P, Sambataro F, Das S, Bertolino A, Weinberger DR, Mattay VS	Biological psychiatry	2009
50	Pupillary contagion: central mechanisms engaged in sadness processing.	Harrison NA, Singer T, Rotshtein P, Dolan RJ, Critchley HD	Social cognitive and affective neuroscience	2006
51*	Reduced emotion processing efficiency in healthy males relative to females.	Weisenbach SL, Rapport LJ, Briceno EM, Haase BD, Vederman AC, Bieliauskas LA, Welsh RC, Starkman MN, McInnis MG, Zubieta JK, Langenecker SA	Social cognitive and affective neuroscience	2014
52*	Segregating intra-amygdalar responses to dynamic facial emotion with cytoarchitectonic maximum probability maps.	Hurlemann R, Rehme AK, Diessel M, Kukolja J, Maier W, Walter H, Cohen MX	Journal of neuroscience methods	2008
53	Similarities and differences in perceiving threat from dynamic faces and bodies. An fMRI study.	Kret ME, Pichon S, Grezes J, de Gelder B	NeuroImage	2011
54	Stop looking angry and smile, please: start and stop of the very same facial expression differentially activate threat- and reward-related brain networks.	Muhlberger A, Wieser MJ, Gerdes AB, Frey MC, Weyers P, Pauli P	Social cognitive and affective neuroscience	2011
55	Temporal pole activity during perception of sad faces, but not happy faces, correlates with neuroticism trait.	Jimura K, Konishi S, Miyashita Y	Neuroscience letters	2009
56*	The amygdala and FFA track both social and non-social face dimensions.	Said CP, Dotsch R, Todorov A	Neuropsychologia	2010
57	The amygdala processes the emotional significance of facial expressions: an fMRI investigation using the interaction between expression and face direction.	Sato W, Yoshikawa S, Kochiyama T, Matsumura M	NeuroImage	2004
58	The behavioral and neural effect of emotional primes on intertemporal decisions.	Luo S, Ainslie G, Monterosso J	Social cognitive and affective neuroscience	2014

59*	The changing face of emotion: age-related patterns of amygdala activation to salient faces.	Todd RM, Evans JW, Morris D, Lewis MD, Taylor MJ	Social cognitive and affective neuroscience	2011
60*	The functional correlates of face perception and recognition of emotional facial expressions as evidenced by fMRI.	Jehna M, Neuper C, Ischebeck A, Loitfelder M, Ropele S, Langkammer C, Ebner F, Fuchs S, Schmidt R, Fazekas F, Enzinger C	Brain research	2011
61	The highly sensitive brain: an fMRI study of sensory processing sensitivity and response to others' emotions.	Acevedo BP, Aron EN, Aron A, Sangster MD, Collins N, Brown LL	Brain and behavior	2014
62	The Kuleshov Effect: the influence of contextual framing on emotional attributions.	Mobbs D, Weiskopf N, Lau HC, Featherstone E, Dolan RJ, Frith CD	Social cognitive and affective neuroscience	2006
63	The stimuli drive the response: an fMRI study of youth processing adult or child emotional face stimuli.	Marusak HA, Carre JM, Thomason ME	NeuroImage	2013
64*	Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain.	Botvinick M, Jha AP, Bylsma LM, Fabian SA, Solomon PE, Prkachin KM	NeuroImage	2005

Table 4.2: 17 features used for the facial feature space

Feature #	Feature name
1	eyebrow length
2	inter-eyebrow distance
3	eye width
4	inter-eyes distance
5	vertical distance between eyes and nosetip
6	horizontal length of the nose
7	distance between nose and upper lip
8	face height
9	face width
10	eye height
11	width of the mouth
12	intense of red on cheeks
13	intense of green on cheeks
14	intense of blue on cheeks
15	contrast polarity between eyes and nose
16	eye area
17	eye mouth ratio

Table 4.3: The MNI coordinates for the weighted center, volume, and the corresponding label name of the significant clusters in the ALE map from meta-analysis

Cluster #	X (mm)	Y (mm)	Z (mm)	Volume (mm^3)	Lateralization	Label
1	23	-3	-18	6088	right	amygdala
2	-22	-4	-18	4640	left	amygdala
3	56	-42	5	2520	right	middle/superior temporal gyrus
4	3	11	53	1464	right	superior frontal gyrus

Chapter 5

Decoding the patterns of neural interactions underlying categorical and individual image perception

In this chapter, we move from analyzing representation dynamics within a specific region of interest to a larger scale. Specifically, we are interested in the representational structure of the interactions between brain areas. The lack of multivariate methods for decoding the representational content of interregional neural communication has left it difficult to know what information is represented in distributed brain circuit interactions, in addition to the information represented in each of the local brain area. In this chapter, we address this gap and present a novel method termed Multi-Connection Pattern Analysis (MCPA), which works by learning mappings between the activity patterns of the populations as a factor of the information being processed. These maps are used to predict the activity from one neural population based on the activity from the other population. Successful MCPA-based decoding indicates the involvement of distributed computational processing and provides a framework for probing the representational structure of the interaction. Simulations demonstrate the efficacy of MCPA in realistic circumstances. In addition, we demonstrate that MCPA can be applied to different signal modalities to evaluate a variety of hypothesis associated with information coding in neural communications. We apply MCPA to fMRI and human intracranial electrophysiological data to provide a proof-of-concept of the utility of this method for decoding individual natural images and faces in functional connectivity data. We further use a MCPA-based representational similarity analysis to illustrate how MCPA may be used to test computational models of information transfer among regions of the visual processing stream. Thus, MCPA can be used to assess the information represented in the coupled activity of interacting neural circuits and probe the underlying principles of information transformation between regions.

5.1 Introduction

Since at least the seminal studies of Hubel and Wiesel (Hubel and Wiesel, 1959) the computational role that neurons and neural populations play in processing has defined, and has been

defined by, how they are tuned to represent information. The classical approach to address this question has been to determine how the activity recorded from different neurons or neural populations varies in response to parametric changes in the information being processed. Single unit studies have revealed tuning curves for neurons from different areas in the visual system responsive to features ranging from the orientation of a line, shapes, and even high level properties such as properties of the face (Desimone et al., 1984; Hubel and Wiesel, 1959; Tsao et al., 2006). Multivariate methods, especially pattern classification methods from modern statistics and machine learning, such as multivariate pattern analysis (MVPA), have gained popularity in recent years and have been used to study neural population tuning and the information represented via population coding in neuroimaging and multiunit activity (Cox and Savoy, 2003; Ghuman et al., 2014; Haxby et al., 2001; Haynes and Rees, 2006; Hirshorn et al., 2016; Kamitani and Tong, 2005; Poldrack, 2011; Polyn et al., 2005). These methods allow one to go beyond examining involvement in a particular neural process by probing the nature of the representational space contained in the pattern of population activity (Edelman et al., 1998; Haxby et al., 2014; Kriegeskorte and Kievit, 2013).

Neural populations do not act in isolation, rather the brain is highly interconnected and cognitive processes occur through the interaction of multiple populations. Indeed, many models of neural processing suggest that information is not represented solely in the activity of local neural populations, but rather at the level of recurrent interactions between regions (Grossberg, 1982; Kveraga et al., 2007; Lee and Mumford, 2003). However previous studies only focused on the information representation within a specific population (Freiwald et al., 2009; Ghuman et al., 2014; Haxby et al., 2014; Hirshorn et al., 2016; Nestor et al., 2011; Tsao et al., 2006), as no current multivariate methods allow one to directly assess what information is represented in the pattern of functional connections between distinct and interacting neural populations with practical amounts of data. Such a method would allow one to assess the content and organization of the information represented in the neural interaction. Thus, it remains unknown whether functional connections passively transfer information between encapsulated modules (Fodor, 1983) or whether these interactions play an adaptive computational role in processing. Note that our definition of non-adaptive information transfer is equivalent to a static linear projection where no computational ”work” is done in the interaction between the regions and therefore no information is added (from an information theory perspective). Adaptive information transfer is one in which computational work related to the behavioral state or condition is performed and therefore state or condition specific information is added through the interaction between regions; this is equivalent to a non-linear function.

Univariate methods that go beyond assessing the degree of coupling between populations to assess changes in the relationship between the activity as a factor of condition also examine adaptive communication between regions. For example the psychophysiological interactions (PPI; (Friston et al., 1997)) and dynamic causal modeling methods (Friston et al., 2003) are sensitive to adaptive interregional communication. Multivariate methods, however, in comparison to univariate methods, allow for more sensitive detection of cognitive states, relating brain activity to behavior on a trial-by-trial basis, and characterizing the structure of the neural code (Norman et al., 2006). Thus, a multivariate pattern analysis method for functional connectivity analysis is critical for decoding the representational structure of interregional interactions.

In this paper, we introduce a multivariate analysis algorithm combining functional connec-

tivity and pattern recognition analyses that we term Multi-Connection Pattern Analysis (MCPA). MCPA works by learning the discriminant information represented in the shared activity between distinct neural populations by combining multivariate correlational methods with pattern classification techniques from machine learning in a novel way. Much the way that MVPA goes beyond a t-test or ANOVA by building a multivariate model of local activity that is then used for single-trial prediction and classification, MCPA goes beyond PPI by building a multivariate connectivity model that is then used for single-trial prediction and classification. This single-trial prediction and classification makes MCPA distinct from previous connectivity approaches that only statistically test the absolute or relative functional connectivity between two populations (Cribben et al., 2012; Finn et al., 2015; Richiardi et al., 2011; Shirer et al., 2012; Wang et al., 2015) and allows for a detailed probe of the representational structure of the interaction.

The MCPA method consists of an integrated process of learning connectivity maps based on the pattern of coupled activity between two populations A and B conditioned on the stimulus information and using these maps to classify the information representation in shared activity between A and B in test data. The rationale for MCPA is that if the activity in one area can be predicted based on the activity in the other area and the mapping that allows for this prediction is sensitive to the information being processed, then this suggests that the areas are communicating with one another and the communication pattern is sensitive to the information being processed. Thus, MCPA simultaneously asks two questions: 1) Are the multivariate patterns of activity from two neural populations correlated? (i.e. is there functional connectivity?) and 2) Does the connectivity pattern change based on the information being processed? This is operationalized by learning a connectivity map that maximizes the multivariate correlation between the activities of the two populations in each condition. This map can be thought of like the regression weights that transform the activity pattern in area A to the activity pattern in area B (properly termed "canonical coefficients" because a canonical correlation analysis [CCA] is used to learn the map). These maps are then used to generate the predictions as part of the classification algorithm. Specifically, a prediction of the activity pattern in one region is generated for each condition based on the activity pattern in the other region projected through each mapping. Single trial classification is achieved by comparing these predicted activity patterns with the true activity pattern (see Figure 5.1 for illustration). With MCPA single trial classification based on multivariate functional connectivity patterns is achieved allowing the nature of the representational space of the interaction to be probed.

We present a number of simulations to validate MCPA for a realistic range of signal-to-noise ratios (SNR) and to show that MCPA is insensitive to local information processing. We apply MCPA to examine the inter-regional representation for natural visual stimuli in visual cortex using functional magnetic resonance imaging (fMRI) data. Specifically, we show that the interactions between regions of the visual stream (V1, V2, V3, V4, and lateral occipital cortex [LO]) are sensitive to information about individual natural images. We combine MCPA with representational similarity analysis to demonstrate that MCPA can be used to evaluate computational models and make inferences regarding the underlying neural mechanism of information transferring. To demonstrate MCPA's applicability to electrophysiological signals and multivariate oscillatory synchrony, we use MCPA to examine the circuit-level representation for faces using intracranial electroencephalography (iEEG) data. Specifically, we show that the interaction between the occipital face area (OFA) and the fusiform face area (FFA) represents information

about individual faces. Despite the potential caveat of small effect size due to the limited size of dataset, these results demonstrate that MCPA can be used to probe the nature of representational space resulting from processing distributed across neural regions.

5.2 Methods

5.2.1 Overview

The MCPA method consists of a learning phase and a test phase (as in machine learning, where a model is first learned, then tested). In the learning phase, the connectivity maps for each condition that characterize the pattern of shared activity between two populations is learned. In the test phase, these maps are used to generate predictions of the activity in one population based on the activity in the other population as a factor of condition and these predictions are tested against the true activity in the two populations. Similar to linear regression where one can generate a prediction for the single variable A given the single variable B based on the line that correlates A and B, MCPA employs a canonical correlation model (a generalization of multivariate linear regression) and produces a mapping model for each condition as a hyperplane that correlates multidimensional vectors A and B. Thus one can generate a prediction of the observation in multivariate space A given the observation in multivariate space B on a single trials basis. In this sense, MCPA is more analogous to a machine learning classifier combined with a multivariate extension of PPI (Friston et al., 1997) rather than being analogous to correlation-based functional connectivity measures.

The general framework of MCPA is to learn the connectivity map between the populations for each task or stimulus condition separately based on training data. Specifically, given two neural populations (referred to as A and B), the neural activity of the two populations can be represented by feature vectors in multi-dimensional spaces (Haxby et al., 2014). The actual physical meaning of the vectors would vary depending on modality, for example spike counts for a population of single unit recordings; time point features for event-related potentials (ERP) or event-related fields; time-frequency features for electroencephalography, electrocorticography (ECoG) or magnetoencephalography; or single voxel blood-oxygen-level dependent (BOLD) responses for functional magnetic resonance imaging. A mapping between A and B is calculated based on any shared information between them for each condition on the training subset of the data. This mapping can be any kind of linear transformation, such as any combination of projections, scalings, rotations, reflections, shears, or squeezes.

These mappings are then tested as to their sensitivity to the differential information being processed between cognitive conditions by determining if the neural activity can be classified based on the mappings. Specifically, for each new test data trial, the maps are used to predict the neural activity in one area based on the activity in the other area and these predictions are compared to the true condition of the data. The trained information-mapping model that fits the data better is selected and the trial is classified into the corresponding condition. This allows one to test whether the mappings were sensitive to the differential information being represented in the neural interaction in the two conditions.

The flow of the MCPA framework is demonstrated in Figure 5.1 and Algorithm 1. An

implementation of MCPA with sample data and scripts in MATLAB are freely available at `https://github.com/yuanningli/MCPA`.

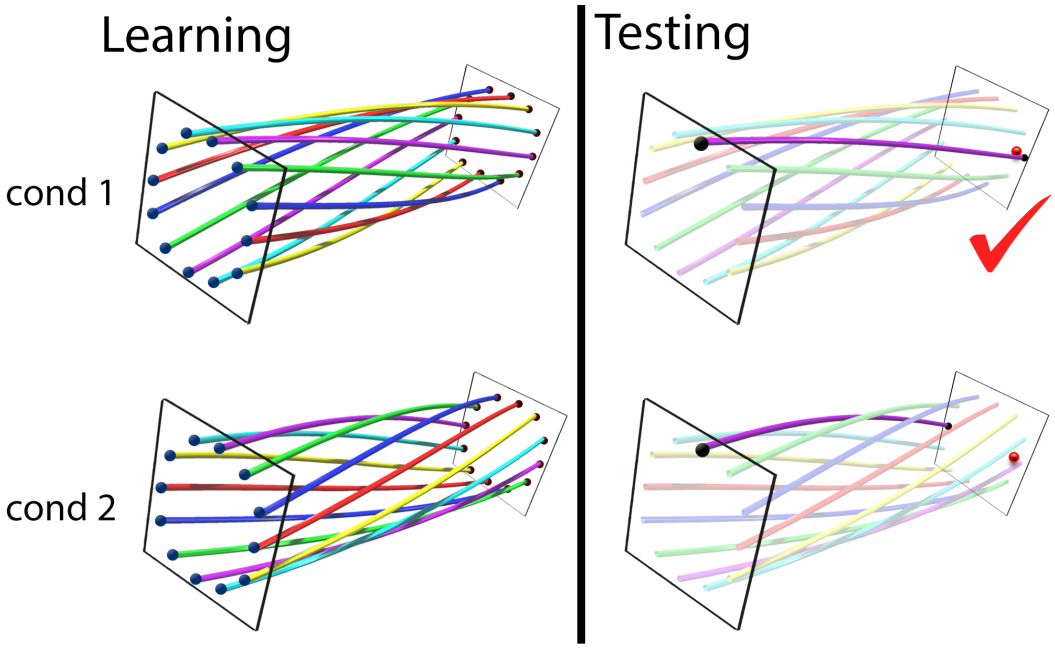


Figure 5.1: **Illustration of the connectivity map and classifier of MCPA. The MCPA framework is demonstrated as a two-phase process: learning and testing. Top left**: an illustration of the learned functional information mapping between two populations under condition 1. The representational state spaces of the two populations are shown as two planes and each pair of blue and red dots correspond to an observed data point from the populations. The functional information mapping is demonstrated as the colored pipes that project points from one space onto another (in this case, a 90 degree clockwise rotation). **Bottom left**: an illustration of the learned functional information mapping between two populations under condition 2 (in this case, a 90 degree counterclockwise rotation). **Top right**: an illustration of the predicted signal by mapping the observed neural activity from one population onto another using the mapping patterns learned from condition 1. The real signal in the second population is shown by the red dot. **Bottom right**: an illustration of the predicted signal by mapping the observed neural activity from one population onto another using the mapping patterns learned from condition 2. In this case, MCPA would classify the activity as arising from condition 1 because of the better match between the predicted and real signal.

Algorithm 1: Multi-Connection Pattern Analysis (MCPA)

Data: training data: matrices $\{\mathbf{X}_A^{(i)} \in \mathbb{R}^{m_A \times n_i}, \mathbf{X}_B^{(i)} \in \mathbb{R}^{m_B \times n_i}\}$ for training observation in ROI-A and ROI-B under condition i, $i = 1, ..., k$;
testing data: $\mathbf{x}_A \in \mathbb{R}^{m_A}$, $\mathbf{x}_B \in \mathbb{R}^{m_A}$ for observation in ROI-A and ROI-B
Result: Prediction of condition y for observation $(\mathbf{x}_A, \mathbf{x}_B)$

Learning phase:
for $i \leftarrow 1$ **to** k **do**
    Apply CCA on $\{\mathbf{X}_A^{(i)}, \mathbf{X}_B^{(i)}\}$ to get linear mapping function $\mathbf{R}_{BA}^{(i)}, \mathbf{R}_{AB}^{(i)}$;
Testing phase:
for $i \leftarrow 1$ **to** k **do**
    Use $\mathbf{x}_A$ and $\mathbf{R}_{BA}^{(i)}$ to reconstruct activity in ROI-B under condition i, $\widehat{\mathbf{x}}_B^{(i)} = \mathbf{R}_{BA}^{(i)}\mathbf{x}_A$;
    Use $\mathbf{x}_B$ and $\mathbf{R}_{AB}^{(i)}$ to reconstruct activity in ROI-A under condition i, $\widehat{\mathbf{x}}_A^{(i)} = \mathbf{R}_{AB}^{(i)}\mathbf{x}_B$;
Assign the condition that gives maximum average correlation coefficient as:
$y = \underset{i \in \{1,..,k\}}{\text{argmax}}\ \text{corr}(\widehat{\mathbf{x}}_B^{(i)}, \mathbf{x}_B) + \text{corr}(\widehat{\mathbf{x}}_A^{(i)}, \mathbf{x}_A)$

5.2.2 Connectivity Map

The first phase of MCPA is to build the connectivity map between populations. The neural signal in each population can be decomposed into two parts: the part that encodes shared information, and the part that encodes non-shared local information (including any non-shared measurement noise, shared measurement noise, such as movement artifacts in fMRI, can result in artifactually inflated connectivity, but for well-balanced and randomized experiments should not differ between conditions and therefore does not affect MCPA discrimination). We assume that the parts of the neural activities that represent the shared information in the two populations are linearly correlated (though, this can easily be extended by the introduction of a non-linear kernel). The model can be described as follows:

$$\mathbf{C} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \quad \min\{m_A, m_B\} \geq d \geq 1 \tag{5.1}$$

$$\mathbf{A}|\mathbf{C} = \mathbf{W}_A\mathbf{C} + \mathbf{D}, \quad \mathbf{D} \sim \mathcal{N}(\boldsymbol{\mu}_A, \boldsymbol{\Psi}_A), \quad \mathbf{W}_A \in \mathbb{R}^{m_A \times d},\ \boldsymbol{\Psi}_A \succeq \mathbf{0} \tag{5.2}$$

$$\mathbf{B}|\mathbf{C} = \mathbf{W}_B\mathbf{C} + \mathbf{E}, \quad \mathbf{E} \sim \mathcal{N}(\boldsymbol{\mu}_B, \boldsymbol{\Psi}_B), \quad \mathbf{W}_B \in \mathbb{R}^{m_B \times d},\ \boldsymbol{\Psi}_B \succeq \mathbf{0} \tag{5.3}$$

where $\mathbf{A} \in \mathbb{R}^{m_A}$ and $\mathbf{B} \in \mathbb{R}^{m_B}$ are the population activity vectors in A and B respectively, $\mathbf{C} \in \mathbb{R}^d$ is the common activity, $\mathbf{D} \in \mathbb{R}^{m_A}$ and $\mathbf{E} \in \mathbb{R}^{m_B}$ are local activities, m_A, m_B are the dimensionality of activity vector in population A and B respectively. Without loss of generality, $\boldsymbol{\mu}_A = \boldsymbol{\mu}_B = 0$ can be assumed. The activity in population A can be decomposed into shared activity $\mathbf{W}_A\mathbf{C}$ and local activity $\mathbf{D}$, while activity in B can be decomposed into shared activity $\mathbf{W}_B\mathbf{C}$ and local activity $\mathbf{E}$. The shared discriminant information only lies in the mapping matrix $\mathbf{W}_A$ and $\mathbf{W}_B$ since $\mathbf{C}$ always follows the standard multivariate normal distribution (though correlation measures that do not assume normally distributed data can also be applied with minor modifications to the calculation). In statistics, canonical correlation analysis (CCA) is optimally designed for such a model and estimate the linear mappings (Bach and Jordan, 2005; Hardoon

et al., 2004). In brief, let $\mathbf{S}$ be the covariance matrix

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{AA} & \mathbf{S}_{AB} \\ \mathbf{S}_{AB}^T & \mathbf{S}_{BB} \end{bmatrix} = \mathbb{E}\left[\begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} \quad \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix}^T \right] \tag{5.4}$$

Therefore $\mathbf{W}_A$ and $\mathbf{W}_B$ can be estimated by solving the following eigen problem

$$\begin{cases} \mathbf{S}_{AA}^{-1}\mathbf{S}_{AB}\mathbf{S}_{BB}^{-1}\mathbf{S}_{BA}\mathbf{U}_A = \rho^2\mathbf{U}_A \\ \mathbf{S}_{BB}^{-1}\mathbf{S}_{BA}\mathbf{S}_{AA}^{-1}\mathbf{S}_{AB}\mathbf{U}_B = \rho^2\mathbf{U}_B \end{cases} \tag{5.5}$$

And we have

$$\begin{cases} \mathbf{W}_A = \mathbf{S}_{AA}\mathbf{U}_{Ad}\mathbf{M}_1 \\ \mathbf{W}_B = \mathbf{S}_{BB}\mathbf{U}_{Bd}\mathbf{M}_2 \end{cases} \tag{5.6}$$

where $\mathbf{U}_{Ad}$ and $\mathbf{U}_{Bd}$ are the first d columns of canonical directions $\mathbf{U}_A$ and $\mathbf{U}_B$, and $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{d\times d}$ are arbitrary matrices such that $\mathbf{M}_1\mathbf{M}_2^T = \mathbf{P}_d$, $\mathbf{P}_d$ is the diagonal matrix with the first d elements of $\mathbf{P} = \mathbf{U}_B^T\mathbf{S}_{BA}\mathbf{U}_A$. Therefore, $\mathbf{M}_1$ and $\mathbf{M}_2$ are just matrices used to normalize the projection of A and B onto the latent space. So $\mathbf{M}_1$ and $\mathbf{M}_2$ can take arbitrary value as long as $\mathbf{M}_1\mathbf{M}_2^T = \mathbf{P}_d$, where $\mathbf{P}_d$ is the diagonal matrix representing the variance along each of the d latent dimensions. Therefore, we can just take $\mathbf{M}_1 = \mathbf{M}_2 = \mathbf{P}_d^{1/2}$.

With $\mathbf{W}_A$ and $\mathbf{W}_B$, the shared information $\mathbf{C}$ can be estimated using its posterior mean $\mathbb{E}(\mathbf{C}|\mathbf{A})$ and $\mathbb{E}(\mathbf{C}|\mathbf{B})$, where $\mathbf{E}(\mathbf{C}|\mathbf{A}) = \mathbf{M}_1^T\mathbf{U}_A^T\mathbf{A}$ and $\mathbb{E}(\mathbf{C}|\mathbf{B}) = \mathbf{M}_2^T\mathbf{U}_B^T\mathbf{B}$. Let $\mathbf{M}_1 = \mathbf{M}_2$ and equate $\mathbb{E}(\mathbf{C}|\mathbf{A})$ and $\mathbb{E}(\mathbf{C}|\mathbf{B})$, this shared information can be used as a relay to build the bidirectional mapping between A and B. Specifically,

$$\mathbf{M}_1^T\mathbf{U}_A^T\mathbf{A} = \mathbf{M}_2^T\mathbf{U}_B^T\mathbf{B} \tag{5.7}$$

$$\Rightarrow \quad \begin{cases} \widehat{\mathbf{B}} = (\mathbf{M}_2^T\mathbf{U}_B^T)^\dagger\mathbf{M}_1^T\mathbf{U}_A^T\mathbf{A} = \mathbf{U}_B^{T\dagger}\mathbf{U}_A^T\mathbf{A} = \mathbf{R}_{BA}\mathbf{A} \\ \widehat{\mathbf{A}} = (\mathbf{M}_1^T\mathbf{U}_A^T)^\dagger\mathbf{M}_2^T\mathbf{U}_B^T\mathbf{B} = \mathbf{U}_A^{T\dagger}\mathbf{U}_B^T\mathbf{B} = \mathbf{R}_{AB}\mathbf{B} \end{cases} \tag{5.8}$$

where

$$\mathbf{R}_{BA} = \mathbf{U}_B^{T\dagger}\mathbf{U}_A^T = \mathbf{U}_B(\mathbf{U}_B^T\mathbf{U}_B)^{-1}\mathbf{U}_A^T \tag{5.9}$$

$$\mathbf{R}_{AB} = \mathbf{U}_A^{T\dagger}\mathbf{U}_B^T = \mathbf{U}_A(\mathbf{U}_A^T\mathbf{U}_A)^{-1}\mathbf{U}_B^T \tag{5.10}$$

To ensure stability, we use ℓ_2 regularization, which yields the ridge estimator (Tikhonov and Arsenin, 1977; Vinod, 1976)

$$\begin{cases} (\mathbf{S}_{AA} + \lambda_1\mathbf{I}_{m_A})^{-1}\mathbf{S}_{AB}(\mathbf{S}_{BB} + \lambda_2\mathbf{I}_{m_B})^{-1}\mathbf{S}_{BA}\mathbf{U}_A = \rho^2\mathbf{U}_A \\ (\mathbf{S}_{BB} + \lambda_2\mathbf{I}_{m_B})^{-1}\mathbf{S}_{BA}(\mathbf{S}_{AA} + \lambda_1\mathbf{I}_{m_A})^{-1}\mathbf{S}_{AB}\mathbf{U}_B = \rho^2\mathbf{U}_B \end{cases} \tag{5.11}$$

and

$$\mathbf{R}_{BA} = \mathbf{U}_B(\mathbf{U}_B^T\mathbf{U}_B + \lambda_3\mathbf{I}_d)^{-1}\mathbf{U}_A^T \tag{5.12}$$

$$\mathbf{R}_{AB} = \mathbf{U}_A(\mathbf{U}_A^T\mathbf{U}_A + \lambda_4\mathbf{I}_d)^{-1}\mathbf{U}_B^T \tag{5.13}$$

In the first step, the connectivity map is estimated for each condition separately. Suppose we have n_1 trials in condition 1 and n_2 trials in condition 2 in the training set, the training data for the two conditions are represented in matrices as $[\mathbf{X}_A^{(1)}, \mathbf{X}_B^{(1)}]^T$ and $[\mathbf{X}_A^{(2)}, \mathbf{X}_B^{(2)}]^T$ respectively, where $\mathbf{X}_A^{(1)} \in \mathbb{R}^{m_A \times n_1}$, $\mathbf{X}_B^{(1)} \in \mathbb{R}^{m_b \times n_2}$ are the population activity for A and B under condition 1 respectively, and $\mathbf{X}_A^{(2)} \in \mathbb{R}^{m_A \times n_2}$, $\mathbf{X}_B^{(2)} \in \mathbb{R}^{m_b \times n_2}$ are the population activity for A and B under condition 2 respectively. The testing data vector is then represented as $[\mathbf{x}_A^T, \mathbf{x}_B^T]^T$, where $\mathbf{x}_A \in \mathbb{R}^{m_A}$ and $\mathbf{x}_B \in \mathbb{R}^{m_B}$ are population activities in A and B respectively. Using CCA, the estimations of the mapping matrices with respect to different conditions are $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$. To sum up, by building the connectivity map, linear mapping function $\mathbf{R}$ is estimated from the data for each condition so that the activity of the two populations can be directly linked through bidirectional functional connectivity that captures only the shared information.

5.2.3 Classification

The second phase of MCPA is a pattern classifier that takes in the activity from one population and predicts the activity in a second population based on the learned connectivity maps conditioned upon the stimulus condition or cognitive state. The testing data is classified into the condition to which the corresponding model most accurately predicts the true activity in the second population.

The activity from one population is projected to another using the learned CCA model,

$$\mathbf{x}_B^{(i)} = \mathbf{R}_{BA}^{(i)} \mathbf{x}_A \tag{5.14}$$

$$\mathbf{x}_A^{(i)} = \mathbf{R}_{AB}^{(i)} \mathbf{x}_B \tag{5.15}$$

The predicted projections $\mathbf{x}_B^{(i)}$ are compared to the real observation $\mathbf{x}_B$, and then the testing trial is labeled to the condition where the predicted and real data match most closely. Cosine similarity (correlation) is used as the measurement of the goodness of prediction. The mapping is bidirectional, so A can be projected to B and vice versa. In practice, the similarities from the two directions are averaged in order to find the condition that gives maximum average correlation coefficient. Therefore, we have

$$\widehat{y}_{\text{pred}} = \underset{i}{\text{argmax}} \quad \frac{\mathbf{x}_B^T \mathbf{x}_B^{(i)}}{\|\mathbf{x}_B\| \|\mathbf{x}_B^{(i)}\|} + \frac{\mathbf{x}_A^T \mathbf{x}_A^{(i)}}{\|\mathbf{x}_A\| \|\mathbf{x}_A^{(i)}\|} \tag{5.16}$$

5.2.4 Simulated experiments

Simulations to evaluate the general performance of MCPA

To test the performance of MCPA, we used BOLD signal recorded from areas V1 and V2 to simulate shared and local activity in two populations and tested the performance of MCPA on synthetic data as a factor of the number of dimensions in each population and signal-to-noise ratio (SNR; Figure 5.2a). We further evaluated three control experiments to demonstrate that MCPA is insensitive to the presence or change in the local information. For the first simulation (Figure 5.2a), we sampled from the empirical distribution of BOLD signal recorded from area V1

in the visual cortex and used it as the shared activity, and independently sampled signal from the empirical distributions of activity in V1 and V2 as the local unshared activity. (See fMRI method described below for experiment details). The shared activity for both conditions in population A was drawn from the empirical distribution of the first d principal components of V1 activity to mimic a d-dimensional normal distribution $\mathbf{Y}_A^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_d)$, for $i = 1, 2$, where $\mathbf{\Sigma}_d$ is a diagonal matrix with the j-th element in the diagonal as σ_j^2. The shared activity in population B under two different conditions were generated by rotating $\mathbf{Y}_A$ with different rotation matrices separately, $\mathbf{Y}_B^{(i)} = \mathbf{R}(i)\mathbf{Y}_A^{(i)}$, where $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ were two d-by-d random rotation matrices corresponding to the information mapping functions under condition 1 and 2 respectively, and for simplicity, $\mathbf{R}^{(i)}$ is orthonormal with $\mathbf{R}^{(i)^T}\mathbf{R}^{(i)} = \mathbf{I}_d$. In addition to the shared activity, local activity in A and B was randomly drawn from the empirical distributions of the first d principal components of V1 and V2 activity respectively and multiplied by a factor of σ to simulate white noise $\mathbf{E}^{(i)} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{\Sigma}_d)$.

The two important parameters here are the dimensionality d and the variance σ^2. SNR was used to characterize the ratio between the variance of shared activity and variance of local activity, and the logarithmic decibel scale $\text{SNR}_{\text{dB}} = -10\log_{10}(\sigma^2)$ was used. To cover the wide range of possible data recorded from different brain regions and different measurement modalities, we tested the performance of MCPA with d ranging from 2 to 25 and SNR ranging from -20 dB to 20 dB (σ^2 ranged from 0.01 to 100). Note that each of the d dimensions contain independent information about the conditions and have the same SNR. Thus the overall SNR does not change, but the amount of pooled information does change with d. For each particular setup of parameters, the rotation matrices $\mathbf{R}^{(i)}$ were randomly generated first, then 200 trials were randomly sampled for each condition and evenly split into training set and testing set. MCPA was trained using the training set and tested on the testing set to estimate the corresponding true positive rate (TPR) and false positive rate (FPR) for the binary classification. The sensitivity index d' was then calculated as $d' = \Phi^{-1}(\text{TPR}) - \Phi^{-1}(\text{FPR})$, where $\Phi^{-1}(x)$ is the inverse function of the cdf of standard normal distribution. This process was repeated 100 times and the mean and standard errors across these 100 simulations were calculated. Note that the only discriminant information about the two conditions is the pattern of interactions between the two populations, and neither of the two populations contains local discriminant information about the two conditions in its own activity. We further tested and confirmed this by trying to classify the local activity in populations A and B (see below). To avoid an infinity d' value, with 100 testing trials per condition, the maximum and minimum for TPR or FPR were set to be 0.99 and 0.01, which made the maximum possible d' to be 4.65.

The MCPA method captures the pattern of correlation between neural activities from populations and is invariant to the discriminant information encoded in local covariance. To see this, we took the simulation data described above and applied MVPA (naïve Bayes) to each of the two populations separately. Note that in each of the two populations, we set the two conditions to have the same mean and covariance. As a result, there should be no local discriminant information within any of the two populations alone.

Robustness of MCPA to non-informative dimensions

In addition to the existing simulations that evaluate the influence of SNR and informative dimensionality on the performance of MCPA, we evaluated the influence of having non-informative dimensionality on the performance of MCPA (Figure 5.2b). Specifically, we simulated 10 informative dimensions and simulate $P = 30$ additional dimensions that are not informative for discrimination and apply MCPA to this simulated data without PCA. We changed the number of training samples available for MCPA and evaluated the performance of MCPA as a factor of the ratio between number of dimensions and the number of training samples per condition. The intuition is that, with a fixed amount of informative dimensions, when the number of training samples decreases, the model would suffer from overfitting and the performance would decay.

Control simulations

For the first control simulation (Figure 5.2c), we fixed the dimensionality at $d = 10$ and SNR at 0 dB ($\sigma^2 = 1$). For condition 1, $\mathbf{X}_A^{(1)}, \mathbf{X}_B^{(1)}$ were drawn independently from the empirical distributions of the first d principal components of area V1 and area V2 using the corresponding empirical distributions; for condition 2, $\mathbf{X}_A^{(2)}, \mathbf{X}_B^{(2)}$ were drawn independently from the same distribution in the empirical distributions of the first d principal components of area V1 and area V2. Then we changed the local variance in one of the conditions. For the features in population A and B under condition 1, we used $\mathbf{X}_A^{(1)'} = k\mathbf{X}_A^{(1)}$, and $\mathbf{X}_B^{(1)'} = k\mathbf{X}_B^{(1)}$, where k ranged from 1 to 9. Thus, in both populations, the variance of condition 1 was different from the variance of condition 2, and such difference would increase as k became larger. Therefore, there was no information shared between the two populations under either condition, but each of the population had discriminant information about the conditions encoded in the variance for any $k \neq 1$.

For the second control simulation (Figure 5.2d), we fixed the dimensionality at 10 and SNR at 0 dB ($\sigma^2 = 1$) and kept the rotation matrices of different conditions different from each other. As a result, the amount of shared discriminant information represented in the patterns of interactions stayed constant. Then we changed the local variance in one of the conditions. For the features in population A under condition 1, we used $\mathbf{X}_A^{(1)'} = k\mathbf{X}_A^{(1)}$, where k ranged from 1 to 9. Thus, population A, the variance of condition 1 was different from the variance of condition 2, and such difference would increase as k became larger. According to our construction of MCPA, it should only pick up the discriminant information contained in the interactions and should be insensitive to the changes in local discriminant information from any of the two populations.

For the third control simulation (Figure 5.2e), we introduced local discriminant information into the two populations to demonstrate that MCPA is insensitive to the presence of constantly correlated local information (Figure 5.2e). We fixed the dimensionality at 10 and SNR at 0 dB ($\sigma^2 = 1$) and kept the rotation matrices constant for different conditions. As a result, the amount of shared discriminant information represented in the patterns of interactions was 0. Then we changed the local variance in one of the conditions. For the features in population A and B under condition 1, we used $\mathbf{X}_A^{(1)'} = k\mathbf{X}_A^{(1)}$, and $\mathbf{X}_B^{(1)'} = k\mathbf{X}_B^{(1)}$, where k ranged from 1 to 9. Thus, in both populations, the variance of condition 1 was different from the variance of condition 2, and such difference would increase as k became larger. Notably, such local information was

actually correlated through interactions between the populations. However, since the pattern of interaction did not vary as the condition changed, there was no discriminant information about the conditions represented in the interactions. According to our construction of MCPA, it should not pick up any discriminant information in this control case.

5.2.5 Examining visual cortex coding for natural images using MCPA

fMRI methods

The fMRI dataset was taken from CRCNS.org Kay et al. (2011). See (Kay et al., 2008; Naselaris et al., 2009) for details regarding subjects, stimuli, MRI parameters, data collection, and data preprocessing. In the experiment, two subjects performed passive natural image viewing tasks while BOLD signals were recorded from the brain. The experiment contains two stages: a training stage and a validation stage. In the training stage, two separate trials were recorded in each subject. In each trial, a total of 1750 images were presented to the subject, which yields a total of 3500 presentations of images (3500 = 1750 images $\times$ 2 repeats). In the validation stage, another 120 images were presented to the subject in 13 repeated trials, which yields a total of 1560 presentations (1560 = 120 images $\times$ 13 repeats). The single-trial response for each voxel was estimated using deconvolution method and used for the following analysis. The voxels were assigned to 5 visual areas (V1, V2, V3, V4, and lateral occipital [LO]) based on retinotopic mapping data from separate scans (Kay et al., 2008; Naselaris et al., 2009).

Categorical image classification

To control for repetition of each individual image and to increase the image number being used, we used the data from the training stage for the categorical image classification. The 1750 images were manually sorted into 8 categories (animals, buildings, humans, natural scenes, textures, food, indoor scenes, and manmade objects). In order to maintain enough statistical power, only categories with more than 100 images were used in the analysis. As a result, 3 categories (food, indoor scenes, and manmade objects) were excluded.

For each pair of ROIs, namely V1-V2, V2-V3, V3-V4, and V4-LO, MCPA was applied to classify the functional connectivity patterns for each possible pair of image categories (total of 10 pairs). For each specific pair of categories, BOLD signal from all the voxels in the ROIs were used as features in MCPA. Principal Component Analysis (PCA) was used to reduce the dimensionality to P, where P corresponds to the number of PCs that capture 90% of variation in the data, which yielded $\sim$100-200 PCs. Leave-one-trial-out cross-validation was used in order to estimate the classification accuracy. This procedure was repeated for all 10 pairs. Classification accuracy and the corresponding sensitivity index d' were used to quantify the performance of MCPA.

Single image classification using MCPA

For single image classification the 13 repetitions of each individual image from the validation stage data was used.

For each pair of ROIs, namely V1-V2, V2-V3, V3-V4, and V4-LO, MCPA was applied to classify the functional connectivity patterns for each possible pair of images (total of 7140 pairs). For each specific pair of categories, BOLD signal from all the voxels in the ROIs were used as features in MCPA. Considering the limited number of trials in each condition, PCA was first used with the data from the training stage to reduce the representation dimensionality to 10. Because the top PCs that explain most variations may contain variance not related to the stimuli, the 10 PCs were selected from the top 50 PCs, based on maximizing the between-trial correlations for single images. As a result, we reduced the dimensionality of the validation data from more than 1000 to 10 based on the training dataset, which was completely independent from all the validation data that was used in the learning and testing stages of MCPA. Leave-one-out cross-validation was then used in order to estimate the classification accuracy. This procedure was repeated for all 7140 pairs. Classification accuracy and the corresponding sensitivity index d were used to quantify the performance of MCPA.

MVPA analysis

MVPA was applied to classify the neural activity within each ROI (V1, V2, V3, V4, and LO) or from a pair of ROIs simultaneously (V1-V2, V2-V3, V3-V4, and V4-LO) for each possible pair of categories (total of 10 pairs). The same features extracted from all the voxels within the ROI, as described above, were used in MVPA analysis. Naïve Bayes classifier was used as the linear classifier and leave-one-out cross-validation was used in order to estimate the classification accuracy. This procedure was repeated for all 10 pairs. Classification accuracy and the corresponding sensitivity index d were used to quantify the performance of MVPA.

Permutation test

Permutation testing was used to evaluate the significance of the classification accuracy d'. For each permutation, the condition labels of all the trials were randomly permuted and the same procedure as described above was used to calculate the classification accuracy (d') for each permutation. The permutation was repeated for a total of 1000 times. The classification accuracy (d') of each permutation was used as the test statistic and the null distribution of the test statistic was estimated using the histogram of the permutation test.

Representational similarity analysis

Based on the classification results, for each classification analysis, the representational dissimilarity matrix (RDM) was constructed such that the j-th element in the i-th row, m_{ij}, equals the dissimilarity (classification accuracy) between the condition i and condition j in the corresponding representational space defined by the analysis. Spearmans rank correlation was used to compare representational dissimilarity matrices in order to account for outliers and non-normality in the data.

Psychophysiological interactions

PPI (Friston et al., 1997) was used to analyze the pattern of interactions between V1 and V4 for each pair of image categories (total of 10). The response in each ROI was extracted by taking the first principal component across all voxels. The PPI model can be written as

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{5.17}$$

where y is the response in ROI1, x_1 is the response in ROI2, x_2 is the categorical condition (1 or -1), and x_3 is the psychophysiological interaction ($x_3 = x_1 x_2$).

HMAX model and connectivity patterns

The implementation of HMAX model by Serre et al. (2007b) was used. Each image was fed into the network and the activations in the four layers (S1, C1, S2, and C2) were recorded. At each patch size level, for image k ($k = 1, 2, , 120$), the activation pattern in simple layer i ($i = 1, 2$) is recorded as $\mathbf{S}_i^k$, which is a square matrix with retinotopic mapping to the image space. On the other hand, the activation pattern in complex layer i ($i = 1, 2$) is represented as vector C_i^k with each element representing the activation of one single unit (for C1, this is achieved by concatenating all the units in the layer into one vector). The activation of each unit in the complex layer was calculated by taking a maximum over its corresponding pool of units in the previous simple layer. For each complex unit, we recorded the location of the corresponding maximum activation simple unit. As a result, we got a N_i-by-2 connectivity matrix $\mathbf{V}_i^k$ for complex layer Ci for image k, where N_i is the total number of units in Ci and each row is the 2-D coordinate of the corresponding maximum activation simple unit. Thus, the connectivity pattern between simple layer Si and complex layer Ci for image k was described by such connectivity matrix $\mathbf{V}_i^k$. Considering all pairs of images, the RDM of the connectivity pattern $\mathbf{M}_i$ is calculated by taking the Frobenius norm of the difference between each pair of connectivity matrix, i.e. $\mathbf{M}_i(j,k) = \|\mathbf{V}_i^j - \mathbf{V}_i^k\|_F$.

The representation space for each single layer was then extracted by concatenating all units in the layer into one vector. The RDM of each single layer was calculated using the Euclidean distance between the corresponding activation vectors of the images.

Representation similarity analysis and permutation test

Permutation test was used to determine the statistical significance of the correlation between the RDM from MCPA and the RDM from HMAX. Specifically, for each pair of ROIs (i.e. V1-V2, V2-V3, V3-V4, and V4-LO), we calculated the corresponding 120-by-120 RDM for all the images from MCPA and averaged across the two subjects, noted as $\mathbf{M}^{\text{ROI1-ROI2}}$, where ROI1-ROI2 = V1-V2, V2-V3, V3-V4, or V4-LO. Then we used the RDMs of HMAX ($\mathbf{M}_i$, $i = 1, 2$) described in the previous part and calculate the Spearman's rank correlation between $\mathbf{M}^{\text{ROI1-ROI2}}$ and $\mathbf{M}_i$. As a result, we have $\rho_i^{\text{ROI1-ROI2}} = corr(\mathbf{M}^{\text{ROI1-ROI2}}, \mathbf{M}_i)$. Then to compare the correlation from different layers in HMAX to MCPA, we use $\Delta\rho^{\text{ROI1-ROI2}} = \rho_1^{\text{ROI1-ROI2}} - \rho_2^{\text{ROI1-ROI2}}$ as the test statistic. For each permutation, the labels of the 120 images were randomly permuted and the above procedure was repeated. With a total of 1000 permutations, we got the empirical

distribution of the test statistic for the null hypothesis that there is no difference between the two correlations. A p-value for the real test statistic can then be estimated.

5.2.6 Examining OFA-FFA coding for individual faces using MCPA

Subject

A human subject underwent surgical placement of iEEG depth electrodes (stereotactic electroencephalography) into the right temporal lobe as standard of care for surgical epilepsy localization. The subject a 56 year-old male. No epileptiform discharges or other evidence of epileptic activity were recorded from the electrode contacts used in this study. The experimental protocols were approved by the Institutional Review Board of the University of Pittsburgh. Written informed consent was obtained from the participant.

Stimuli

In the localizer experiment, 180 images of faces (50% male), bodies (50% male), words, hammers, houses, and phase scrambled faces were used as a functional localizer. Each category contained 30 images. Phase scrambled faces were created in Matlab by taking the 2-dimensional spatial Fourier spectrum of each of the face images, extracting the phase, adding random phases, recombining the phase and amplitude, and taking the inverse 2-dimensional spatial Fourier spectrum. Each image was presented in pseudorandom order and repeated once in each session.

Faces in the individuation experiment were taken from the Karolinska Directed Emotional Faces stimulus set (Lundqvist et al., 1998). Frontal views and 5 different facial expressions (happy, sad, angry, fearful, and neutral) from all 70 faces (50% male) in the database were used, which yielded a total of 350 face images, each presented once in random order during a session. The patient participated in a total of 3 sessions.

All stimuli were presented on an LCD computer screen placed approximately 2 meters from participants' heads.

Experimental paradigms

In the localizer experiment, each image was presented for 900 ms with 900 ms inter-trial interval during which a fixation cross was presented at the center of the screen ($\sim 10^\circ \times 10^\circ$ of visual angle). At random, 25% of the time an image would be repeated. Participants were instructed to press a button on a button box when an image was repeated (1-back). Only the first presentations of repeated images were used in the analysis.

In the individuation experiment, each face was presented for 1500 ms with 500 ms inter-trial interval during which a fixation cross was presented at the center of the screen. Faces subtended approximately 5 degrees of visual angle in width. Subjects were instructed to report whether the face was male or female via button press on a button box.

Paradigms were programmed in MATLAB^{TM} using Psychtoolbox and custom written code.

Data preprocessing

The electrophysiological activity in OFA and FFA were recorded simultaneously using iEEG electrodes at 1000 Hz. They were subsequently bandpass filtered offline from 1-170 Hz using a fifth order Butterworth filter to remove slow and linear drift, the 180 Hz harmonic of the line noise, and high frequency noise. The 60 Hz line noise and the 120 Hz harmonic noise were removed using DFT filter. To reduce potential artifacts in the data, trials with maximum amplitude 5 standard deviations above the mean across the rest of the trials were eliminated. In addition, trials with a change of more than 25 μV between consecutive sampling points were eliminated. These criteria resulted in the elimination of less than 1% of trials.

As the last step of the data preprocessing, we extracted wavelet features using Morlet wavelets. The number of cycles of the wavelet was set to be 7. The entire epoch length of the data was 1500ms (-500 $\sim$ 1000 ms relative to stimulus onset). To avoid numerical issues in MATLAB, the lowest frequency was set at 7 Hz. The wavelet features were estimated using FieldTrip^{TM} toolbox. Finally, we took all the wavelet features at 7-100 Hz, with 1 Hz steps, at every 10 ms as features, which yielded a 94-dimensional feature vector at every time point. All the wavelets were normalized to the baseline by subtracting the mean value and divided by the standard deviation of the data from 350ms to 50ms before stimulus onset.

Electrode selection

Face sensitive electrodes were selected based on anatomical and functional considerations. Electrodes of interest were restricted to those that were located in or near the fusiform gyrus or inferior occipital cortex. In addition, MVPA was used to functionally select the electrodes that showed sensitivity to faces, comparing to other conditions in the localizer experiment. Specifically, electrodes were selected such that their peak 6-way classification d' score exceeded 1 ($p < 0.001$ based on a permutation test, as described below) and the event related potential (ERP) for faces was larger than the ERP for the other non-face object categories.

There were 12 contacts on a depth electrode on the ventral temporal lobe extending along the anterior-posterior axis. Among all the contacts, only three (the 1st, 6th and 7th contacts, see Figure 5.5a for the location of these contacts) satisfied the criterion described above (see Figure S1 for d' timecourses from all contacts on the depth electrode). The first contact was near the mid-fusiform gyrus while the other two were near posterior end of the fusiform gyrus/anterior end of the inferior occipital cortex. Hence we used the data from the first electrode as FFA signal and the averaged data across the 6th and 7th electrodes as the OFA signal (see Figure S2 for averaged ERP data in the two areas). The post-operative structural MRI scan did not allow us to carefully distinguish the precise localization of the "OFA" electrodes and it may be that these electrodes are in fact in the posterior fusiform and properly labeled "FFA-1" according to the recent nomenclature introduced by Weiner and Grill-Spector (2010). However, considering OFA and FFA-1 are contiguous with one another and it has not been determined what, if any, functional distinction there is between the two, we use "OFA" for the label of the electrodes out of convenience.

MCPA Analysis

MCPA was applied to classify the OFA-FFA connectivity for each possible pair of faces (total of 2415 pairs). For each specific pair of faces, averaged wavelet features within a 50 ms time window were used as features in MCPA. Principal Component Analysis (PCA) was used to reduce the dimensionality from 94 to P, where P corresponds to the number of PCs that capture 95% of variation in the data, the typical value of P is around 7$\sim$8. Leave-one-trial-out cross-validation was used in order to estimate the classification accuracy. This procedure was repeated for all 2415 pairs and all time windows slid with 10 ms step between 0 and 600ms after stimulus onset. Similar to previous simulations, d' was used to quantify the performance of MCPA.

Permutation test was used to determine the significance of the d' timecourse of MCPA (Maris and Oostenveld, 2007). During each permutation, the condition labels of all the trials were randomly permuted and the same procedure as described above was used to calculate the timecourse of d' for each permutation. The permutation was repeated for a total of 1000 times. The mean d' during 200-500 ms of each permutation was used as the test statistic and the null distribution of the test statistic was estimated using the histogram of the permutation test. The time window 200-500 ms was chosen based on the fact that the sensitivity of facial identity was only presented in OFA and FFA roughly 200 -500 ms after stimulus onset (Ghuman et al., 2014).

5.3 Results

5.3.1 Simulations

We used simulations to test and verify the performance and properties of MCPA on synthetic data. Specifically, synthetic data generated based on real fMRI data representing neural activity of two distinct populations and the information represented in the interaction between those populations was manipulated to construct different testing conditions.

In the first simulation, we evaluated the ability of MCPA to detect information represented in the functional connectivity pattern when it was present as a factor of the SNR and the number of dimensions of the data. The mean and standard error of the sensitivity index (d') from 100 simulation runs for each particular setup (dimensionality and SNR) are shown in Figure 5.2a. The performance of the MCPA classifier increased when SNR or effective dimensionality increased. Classification accuracy saturated to the maximum when SNR and number of dimensions were high enough (SNR $>$ 5dB, dimensionality $>$ 10). The performance of MCPA was significantly higher than chance ($p < 0.01$, permutation test) for SNRs above -5 dB for all cases where the dimensionality was higher than 2, when the pattern of the multivariate mapping between the activity was changed between conditions.

In addition, we examined how robust MCPA is to uninformative dimensions. This simulation assesses performance of MCPA as the number of training samples changes and approaches the total number of dimensions. In the evaluation with a fixed number of 10 informative dimension and 30 non-informative dimension (40 dimension in total), MCPA was shown to be highly robust to uninformative dimensions and gave significant classification accuracy until the ratio between the number of total dimensions and the number of training samples approaches $\sim 80\%$ (Figure 5.2b).

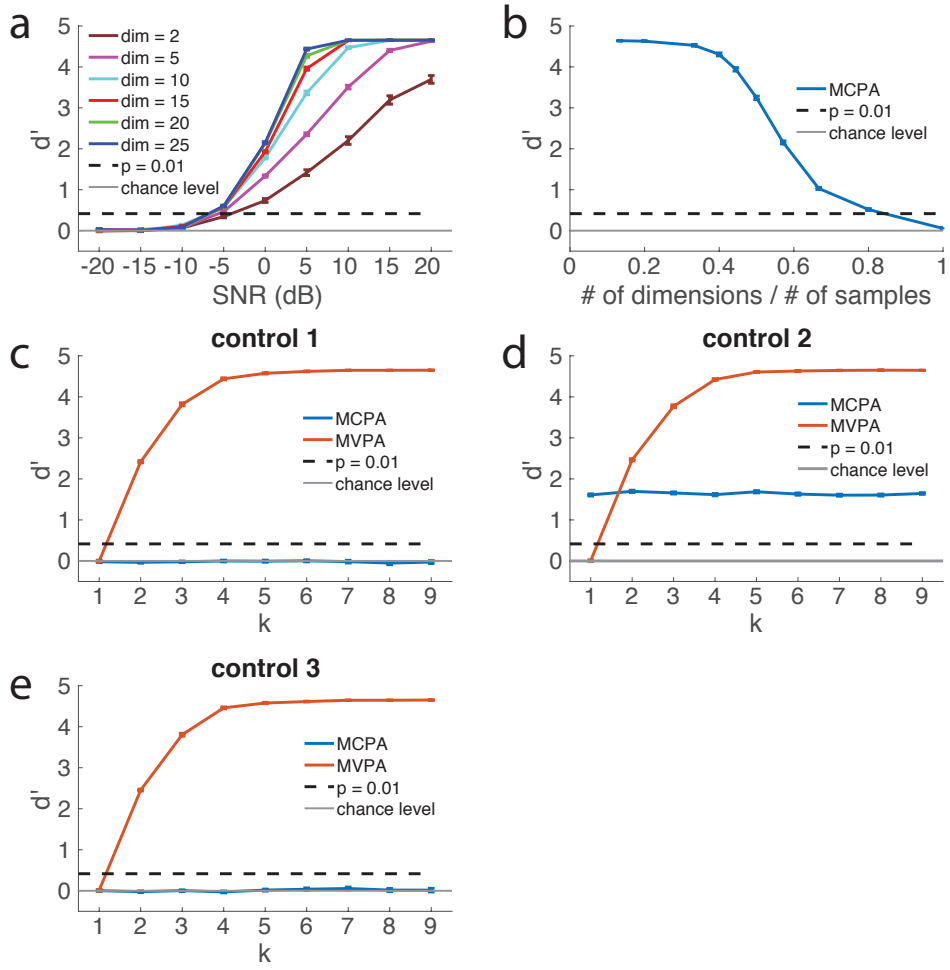


Figure 5.2: **Synthetic data and control simulation experiments.** The mean and standard error for 100 simulation runs are plotted. The horizontal gray line corresponds to chance level ($d' = 0$). The dashed line ($d' = 0.42$, corresponding accuracy 58.5%) corresponds to the chance threshold, $p = 0.01$, based on a permutation test. The maximum possible $d' = 4.65$ (equivalent to 99% accuracy because the d for 100% accuracy is infinity). **(a)** The sensitivity of MCPA for connectivity between two populations as a factor of SNR and the number of effective dimensions in each population. **(b)** The robustness of MCPA to non-informative dimensions. **(c)** The insensitivity of MCPA when there is variable local discriminant information, but no circuit-level information (control case 1). MCPA and MVPA were applied to control case 1. **(d)** The insensitivity of MCPA to changes in local discriminant information with fixed circuit-level information when there is both local and circuit-level information (control case 2). **(e)** The insensitivity of MCPA to variable local discriminant information when the circuit-level activity is correlated, but does not contain circuit-level information about what is being processed (control case 3).

The first control simulation was designed to confirm that when two unconnected populations both carry local discriminant information, MCPA would not be sensitive to that piece of informa-

tion. As shown in Figure 5.2c, MCPA did not show any significant classification accuracy above chance ($d' = 0$) as changed. On the other hand, the MVPA classifier that only took the data from local activity showed significant classification accuracy above chance level and the performance increased as local discriminant information increased.

The second control simulation was designed to test if MCPA would be insensitive to changes in local discriminant information when there was constant information coded in neural communication. Local discriminant information was injected into the populations by varying the ratio of the standard deviation (k) between the two conditions. When MVPA was applied to the local activity, increasing classification accuracy was seen as k became larger (Figure 5.2d). This result confirmed that discriminant information was indeed encoded in the local activity in the simulation. On the other hand, the performance of MCPA did not change with the level of local discriminant information (d' stayed around 1.65 for all cases, corresponding to accuracy = 79%), demonstrating that MCPA is only sensitive to changes in information contained in neural interactions.

The final control simulation tested whether MCPA is simply sensitive to the presence of functional connectivity between two populations per se or is only sensitive to whether the functional connectivity contains discriminant information. Specifically, are local discriminant information in two populations, and a correlation between their activity, sufficient for MCPA decoding? It should not be, considering that MCPA requires that the pattern of the mapping between the populations to change as a factor of the information being processed (see Figure 5.1). The final control simulation was designed to assess whether MCPA is sensitive to the case where two populations communicate, but in a way that would not imply distributed computational processing. Specifically, neural activity in areas A and B were simulated such that local discrimination was possible in each population and the activity of the two populations was correlated, but the interaction between them was invariant to the information being processed. Figure 2e shows that in this case MCPA did not classify the activity above chance, despite significant correlation between the regions and significant local classification (MVPA). Thus, functional connectivity between the populations is a necessary, but not sufficient, condition for MCPA decoding. Therefore, MCPA is only sensitive to the case where the mapping itself changes with respect to the information being processed, which is a test of the presence of distributed neural computation.

5.3.2 Single image classification of visual cortex interactions using MCPA

To assess its performance on real neural data, MCPA was applied to Blood-oxygen-level-dependent (BOLD) fMRI measurements of human occipital visual areas, in two subjects (Subject 1 and Subject 2) during passive viewing of 13 repetitions of 120 natural images (Kay et al., 2008, 2011; Naselaris et al., 2009). MCPA was used for single-trial classification of these images for the interactions between V1-V2, V2-V3, V3-V4, and V4-lateral occipital (LO) cortex (e.g. 4 total region pairs $\times$ 2 subjects; see Figure 5 of Naselaris et al. (2009) for depictions of these regions in these subjects). Across the 8 pairs of regions the mean sensitivity index (d') of the single trial classification was 0.405 (s.d. = 0.094), with all of the pairs showing significant classification at $p < 0.01$ corrected for multiple comparisons (permutation test). In both subjects, MCPA classification accuracy declined going up the classic visual hierarchy. The classification accuracies are shown in Table 5.1.

Table 5.1: Mean d' and classification accuracy of MCPA for Subject 1 and Subject 2 (chance level: $d' = 0$, accuracy = 50%, * $p < 0.01$, permutation test)

Subject 1	V1-V2	V2-V3	V3-V4	V4-LO
d'	0.477 *	0.443 *	0.408 *	0.319 *
accuracy	58.5% *	57.9% *	57.3% *	55.7% *
Subject 2	V1-V2	V2-V3	V3-V4	V4-LO
d'	0.589 *	0.470 *	0.330 *	0.271 *
accuracy	60.3% *	58.5% *	55.9% *	54.9% *

5.3.3 Using MCPA-based RSA to test models of between-area information transformation

One important application of MCPA is to evaluate models and test theoretical hypotheses regarding the computational operation underlying how representations are transformed from one region to another. MCPA-based representational similarity analysis (RSA) can be used to compare the representational space derived from the interaction between brain regions to representational spaces derived from the transformation of representations in computational models. To illustrate this we compare the representational space for natural images in the same fMRI dataset described above to the representational space derived from the transformation between layers of the HMAX model of the visual processing stream (Riesenhuber and Poggio, 1999; Serre et al., 2007b), which is a representative neural network model derived from the classical framework of ventral visual pathway (Figure 1.1C). HMAX has four layers going from S1 to C1 to S2 to C2 along the hierarchy. The transformation of the representation between S1 and C1 (S1-C1 transformation) occurs through a local, non-linear max-pooling operation and the transformation between S2 and C2 (S2-C2 transformation) occurs through a more global non-linear max-pooling operation. We compared the representational dissimilarity matrices (RDMs) derived from these HMAX transformations to the RDMs derived from MCPA between V1-V2, V2-V3, V3-V4, and V4-LO. The transformation between C1 and S2 occurs through a passive filtering that does not give rise to an RDM because the transformation is effectively the same across all C1 representations.

As shown in Figure 5.3, we found that the RDM derived from the S1-C1 transformation in HMAX correlates with the V2-V3 RDM based upon MCPA of the fMRI data (mean Spearman's $\rho = 0.053$, $p < 0.05$, permutation test). Furthermore, the S1-C1 correlation to V2-V3 was significantly greater ($p < 0.05$, permutation test) than the S2-C2 correlation to V2-V3. The RDM derived from the S2-C2 transformation in HMAX correlates with the V4-LO RDM based upon MCPA of the fMRI data (mean Spearman's $\rho = 0.112$, $p = 0.002$, permutation test). Furthermore, the S2-C2 correlation to V4-LO was significantly greater ($p < 0.01$, permutation test) than the S1-C1 correlation to V4-LO. Additionally, none of the individual layers in HMAX showed a consistent significant correlation with the connectivity-based RDM from MCPA. Taken together, these results suggest that the interaction between the lower layers of the neural visual hierarchy reflects an operation more like the operation between the lower layers of the model of the visual hierarchy than between higher layers of the model. Furthermore, the interaction between higher layers of the neural visual hierarchy reflects an operation more like the operation between higher layers of the model than between lower layers of the model.

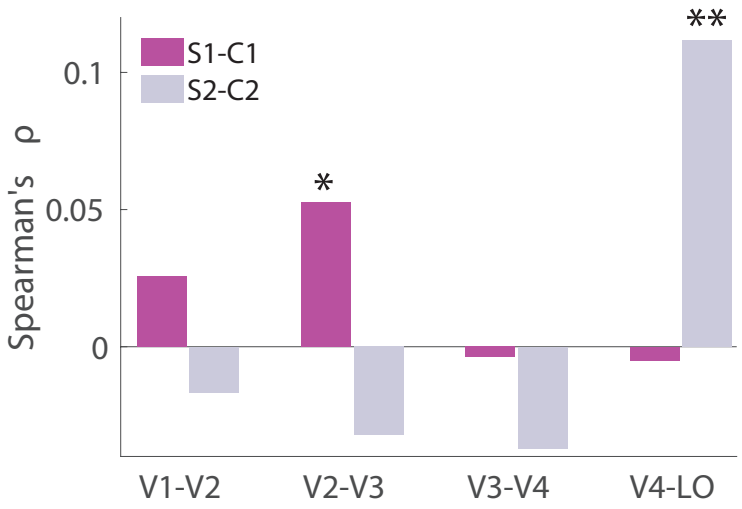


Figure 5.3: **Correlating MCPA and HMAX.** Correlation coefficients between the between-layer connectivity patterns in HMAX (S1-C1, and S2-C2) and the between-area connectivity patterns in fMRI data extracted by MCPA (V1-V2, V2-V3, V3-V4, and V4-LO) were plotted. The correlation was evaluated by Spearman's rank correlation coefficients. For S1-C1, correlation peaked at V2-V3, mean Spearman's $\rho = 0.053$ (* $p = 0.036$, permutation test within each subject, and p-values were combined using Fisher's method). For S2-C2, correlation peaked at V4-LO, mean Spearman's $\rho = 0.112$ (** $p = 0.001$, permutation test within each subject, and p-values were combined using Fisher's method).

5.3.4 Comparing the between region representation to the local representation

To assess whether the information represented in the between region interactions reflected a distinct computational process or merely reflected the representation in either of the individual areas, RSA was performed. To increase our power, we performed this RSA at the category level (animals, buildings, humans, natural scenes, and textures) based on classification accuracy rather than the single image level because the dataset contained many more repetitions per category than per image (Figure 5.4). This yielded a total of 24 correlations (8 MCPA-based matrices correlated with each of the two regions that contribute to each MCPA and with MVPA that takes the two regions together). 20 out of the 24 correlations were negative, many showing large negative correlation coefficients (see Table 5.2 for details and see below for and effect size calculations [Wilks'λ] and statistical tests for the canonical correlations, mean Spearman's $\rho = -0.420$, $s.d. = 0.346$). In other words, categories that were relatively easy to decode based on the activity within regions using MVPA were relatively more difficult to decode based on the shared activity between that region and the other regions in the visual stream using MCPA and vice versa (Figure 5.4). This negative correlation suggests that the communication between regions represents information that has not been explained aspects by local computational processes.

Table 5.2: Spearman's rank correlation coefficients ρ between MCPA of ROI1-ROI2 and MVPA of ROI1-ROI2 in Subjects 1 and 2.

Subject 1	V1-V2	V2-V3	V3-V4	V4-LO
MVPA (both ROIs)	0.333	-0.527	-0.576	-0.309
MVPA (ROI1 only)	0.333	-0.055	-0.721	-0.442
MVPA (ROI2 only)	0.176	-0.370	-0.491	-0.442
Subject 2	V1-V2	V2-V3	V3-V4	V4-LO
MVPA (both ROIs)	-0.685	-0.673	-0.479	-0.527
MVPA (ROI1 only)	-0.539	-0.758	-0.782	-0.539
MVPA (ROI2 only)	-0.855	-0.794	-0.418	0.055

5.3.5 Comparing MCPA to PPI

To demonstrate the dominance of MCPA over classical univariate methods, we applied PPI to the same data to analyze categorical effective analysis between neighboring areas. As a comparison, 80 different pairs of categories (10 pairs of categories $\times$ 4 pairs of regions $\times$ 2 subjects) were analyzed using both PPI and MCPA. 4/80 PPI results were significant with $p < 0.05$ (uncorrected), while 13/80 MCPA results were significant with $p < 0.05$ (uncorrected). As a result, the number of significant MCPA results is significantly larger than the number of significant PPI results ($p < 0.01$, permutation test). Note that it is not clear how many of these 80 different pairs of categories are expected to be classifiable given that the regions examined are not category sensitive, other than LO. Thus, it is not clear if 13/80 is close to the number of category pairs that would be classifiable with perfect data or if this is a low percentage of that number, but the key point in the context of validating MCPA is that MCPA is more sensitive than univariate (PPI) methods.

5.3.6 Single face identity classification of OFA-FFA interactions using MCPA

To further assess its performance on electrophysiological data, MCPA was applied on intracranial electroencephalography (iEEG) data recorded from OFA and FFA in one human epileptic patient during a visual perception task (see Figure 5a for the electrode locations). MCPA was applied in the classification between each possible pair of faces. Previous studies on the time-course of face individuation (Ghuman et al., 2014) have demonstrated that the 250-450 ms time window is critical for the processing of face individuation information. For MCPA, as shown in Figure 5.5b, with a chance level of $d' = 0$ and corresponding accuracy = 50%, the classification accuracy was significantly above chance level across that time window (averaged $d' = 0.14$, mean classification accuracy 52.7%, $p < 0.01$, permutation test). The CCA weights for the FFA and OFA are plotted in Figure 5.5c, showing that 15-30 Hz in FFA and 25-40 Hz in OFA contributed most strongly to their interaction in response to individual faces, suggesting that there may be a degree of cross-frequency coupling involved in the OFA-FFA coding for faces. Using MVPA, classification accuracy was significantly above chance level across that time window in FFA (averaged $d' = 0.42$, mean accuracy 58%, $p < 0.01$, permutation test), replicating previous reports (Ghuman et al., 2014), classification accuracy was also above chance level across that

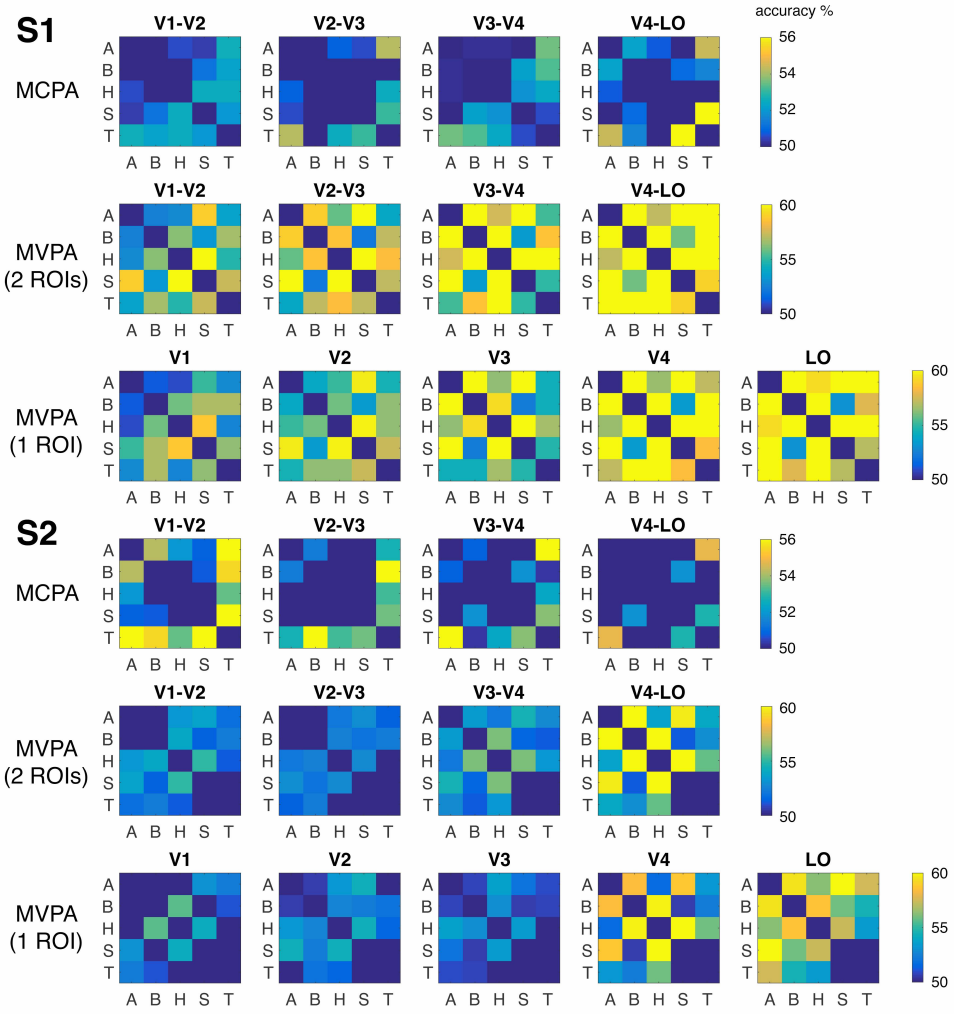


Figure 5.4: **MCPA and MVPA results for fMRI categorical data. RSA results based on MCPA and MVPA Subjects 1 and 2.** Categories: A-animals, B-buildings, H-humans, S-natural scenes, T-textures. **Row 1,3**: RSA based on MCPA for Subject 1, 2; **Row 2, 5**: RSA based on MVPA with two ROIs at a time of Subject 1, 2; **Row 3, 6**: RSA based on MVPA with one ROI at a time of Subject 1, 2 (chance level: accuracy = 50%).

time window in OFA (averaged $d' = 0.13$, mean accuracy 52.6%, $p < 0.05$, permutation test). In the early time window (50 250 ms), MCPA did not show significant classification accuracy (averaged $d' = 0.116$, mean accuracy 51.6%, $p > 0.1$, permutation test). See below for statistical testing of the single face canonical correlation models.

As a control analysis, we took a contact outside of the fusiform gyrus that did not show face sensitivity and performed the same analysis between the control contact and the OFA and FFA

contacts. As shown in Figure 5.5b, the averaged d' of MCPA between the control contact and both the OFA and FFA contacts was not significant above chance level ($d' = 0.074$ for control & FFA, accuracy = 51.2%, $d' = 0.012$ for control & OFA, accuracy = 50.3%, both $p > 0.1$, permutation test).

With the caveat that the effect size is small, the results support the hypothesis individual level face information is represented in the OFA-FFA interaction pattern.

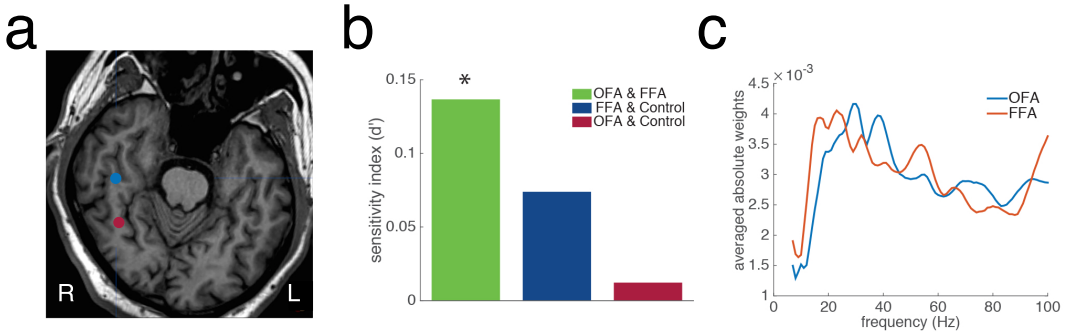


Figure 5.5: **iEEG experiments and MCPA results. (a)** Location of the electrodes of interest. The blue dot corresponds to the location of the FFA contact while the red dot corresponds to the location of the OFA contacts. **(b)** MCPA applied between (1) the OFA and FFA channels, (2) the FFA channel and the control channel, (3) the OFA channel and the control channel. The mean d' of pairwise face classification over all 2415 pair of faces across the 200-500 ms timewindow after stimulus onset is plotted. * $p < 0.01$, permutation test. **(c)** Averaged absolute loading weights in the functional connectivity model of MCPA for OFA and FFA across the frequency spectrum during the time window of 250-450 ms after stimulus onset. (chance level: $d' = 0$, accuracy = 50%)

5.3.7 Testing significance of CCA models

The significance of MCPA relies on two factors: the presence of functional connectivity and the discriminant information in the connectivity patterns. Both are necessary conditions for the significance of MCPA. Here we evaluate the significance of the CCA models in order to further support the MCPA results. This is particularly useful for interpreting non-significant or low accuracy MCPA results. If the CCA models are significant, but MCPA yields non-significant results, this suggests that the areas are functionally connected, but the connectivity pattern does not change with respect to condition (e.g. a passive linear filter). If MCPA does not show significant decoding and the CCA statistics are non-significant, this suggests that the areas are not functionally connected (or have weak functional connectivity at best.

For category-level decoding with the fMRI data, we have enough repetitions to perform parametric test. Specifically, we have $\sim 200 - 400$ trials per condition, and we use the first N PCs that account for 90% of the variance as the features for each ROI. The typical N ranges from

$\sim 100 - 200$. Therefore, for each pair of ROIs under each condition, we can evaluate the CCA model and get N_c canonical variate pairs, where $N_c = \min(N_1, N_2)$, and N_1, N_2 are the numbers of PCs extracted from the two ROIs under this categorical condition. As a result, we will have N_c Wilks'λ, noted as $\lambda_1, , \lambda_{N_c}$, where λ_i corresponds to the test statistic for the model including the i-th to the N_c-th canonical variate pairs. In addition, we can also compute the canonical correlation r_i for each pair of canonical variates. To get a p-value for each individual canonical correlation r_i, we performed a permutation test. Specifically, for all the trials in one category, we permute the order of the trials in ROI2 while maintaining the original order of the trials in ROI1. As a result, we get surrogate CCA models and we can then compute the corresponding canonical correlation r_i for each pair of canonical variates. We repeated this permutation process 1000 times to get the surrogate distribution of the canonical correlations, and then estimated the corresponding p-value. This yields sufficient small λ_1 (smaller than eps(0) in MATLAB) for all categories and all pairs of ROIs, meaning that the whole CCA model is significant for all categories and all pairs of ROIs ($p < eps(0)$). Note that eps(0) = 1e-324. As for each single canonical correlation, 86.9% of all the individual canonical correlations (5718/6580 canonical correlations) reach statistical significance of $p < 0.01$, based on the permutation test.

For the single image case in fMRI, because for each condition we have only 13 repetitions Wilk's λ and the Chi-square approximation are not reliable. Therefore, we use permutation testing and compute the averaged canonical correlation across all conditions as the test statistic for each pair of canonical variates. The procedure of permutation testing is similar to the one described for category-level decoding above. We first fixed the order of trials in ROI1 and permuted the ordering of the trials in ROI2 randomly. Then for each of the 120 images, we estimated the CCA model for each pair of ROIs and computed the corresponding canonical correlations for each pair of canonical variates. As a result, we can evaluate the significance of each canonical correlation averaged across all images using permutation. Since we have 2 subjects, each with 4 pairs of ROIs, and 10 canonical correlations for each pairs of ROIs, there are 80 averaged canonical correlations in total. Based on 1000 permutations, 64/80 canonical correlations are significant at $p < 0.01$. Moreover, 61/80 canonical correlations are significant at $p < 0.001$. (Note that with 1000 permutations, the minimum p value is $p < 0.001$ when the test statistic is larger than all 1000 empirical statistics computed from the permutations)

Similarly, for face individuation in iEEG data, we also use permutation test and compute the averaged canonical correlation across all conditions as the test statistic. As a result, the first 7 canonical correlations have $p < 0.001$, and the last canonical correlation has $p = 0.003$. The corresponding averaged canonical correlation values are $r = [1.00, 1.00, 0.97, 0.89, 0.75, 0.56, 0.33, 0.10]$.

5.3.8 Evaluating feature-selection using PCA

To assess the influence of selecting the most relevant PCs, the percent of variance explained by the subset of PCs are summarized in Table 5.3. As we discuss in the Methods part, the PCs that are most task-relevant are usually not just the top PCs with regards to variance explained. As was pointed out in a previous study using the same public fMRI data as we used here (Henriksson et al., 2015), a majority of the variance in the activity is due to intrinsic fluctuations. Therefore, we computed and selected the top PCs that are task relevant by correlating the first half trials with the second half trials in a separate run in this data set (what was called the training data

in the original Kay et al. (2008) paper). As expected, the task-relevant PCs explain a smaller portion of the total variance than the top PCs. As mentioned, we select PCs based on a set of data that is totally independent from the actual training and testing data for MCPA. This rules out the possibility of overfitting induced by the selection of PCs.

Table 5.3: Amount of variance explained by the subset of selected PCs

% of variance explained	V1	V2	V3	V4	LO
top 10 PCs	21.71	17.00	15.35	11.43	11.97
top 50 PCs	36.04	30.22	28.39	24.43	27.24
10 selected PCs	5.09	4.84	4.89	5.33	6.24

5.4 Discussion

This paper presents a novel method to assess the information represented in the patterns of interactions between two neural populations. MCPA works by learning the mapping between the activity patterns from the populations from a training data set, and then classifying the neural communication pattern using these maps in a test data set. Simulated data demonstrated that MCPA was sensitive to information represented in neural interaction for realistic SNR ranges. Furthermore, MCPA is only sensitive to the discriminant information represented through different patterns of interactions irrespective of the information encoded in the local populations. Applying this method to fMRI data demonstrated that the multivariate connectivity patterns between areas along the visual stream represent information about individual natural images. MCPA-based RSA showed that, at the category level, the representational structure of the interaction between regions is negatively correlated to the representational structure locally within each region. Furthermore, MCPA was used to test hypotheses from the HMAX model regarding the computational operation that transforms the representation between regions along the visual processing pathway. Finally, as an example with electrophysiological data, applying MCPA to iEEG data showed that the multivariate connectivity pattern between OFA and FFA represents information at the level of individual faces.

One practical consideration with MCPA is that CCA generally requires the number of trials to be substantially larger than the number of variables in the two areas. This is often not the case in neuroscientific studies and therefore dimensionality reductions may be required. In the optimal case, this dimensionality reduction would be performed in the canonical space reducing the number of canonical variables used in MCPA-based classification. However, we find that performing a PCA to reduce dimensionality prior to CCA generally performs better than reducing the dimensionality in the canonical space, which is in line with previous in neuroscientific studies using CCA (Karageorgiou et al., 2012; Smith et al., 2015). While it is not entirely clear why PCA before CCA performs better than dimensionality reduction using CCA alone, it is likely because CCA is known to be very sensitive to noise (Anderson, 1958; Gittins, 2012) and using PCA for dimensionality reduction can have the added benefit of noise reduction.

5.4.1 MCPA as assessing adaptive processing

Significant discrimination within each population and significant functional connectivity between them is not sufficient to produce MCPA and indeed local classification within each population is not even necessary (Figures 5.2a and 5.2e respectively). MCPA requires the pattern of connectivity (linear correlations) between the two populations to vary across the different conditions. In other words, MCPA is sensitive to both the degree of functional connectivity in the conditions and how distinct the mappings are across conditions. As an example, if the two populations interact, but the interaction behaves like a passive linear filter, mapping the activity between the populations in a similar way in all conditions, MCPA would not be sensitive to the interaction because the mapping does not change (Figure 5.2e). Instead, MCPA is more akin to testing for non-constant linear filtering or distributed, interactive computation that behaves as a non-linear process where the nature of the interaction adapts (from a linear perspective) as a factor of the information that is being processed. Recent studies demonstrate that neural populations in perceptual areas alter their response properties based on context, task demands, etc. (Gilbert and Li, 2013). These modulations of response properties suggest that lateral and long-distance interactions are adaptive and dynamic processes responsive to the type of information being processed. In this context adaptive is meant purely in the sense that the linear transformation between the multivariate activity in the two regions change as a factor of condition. As noted previously, this is equivalent to a non-linear filter and adaptive denotes that information is added to the representation in an information theoretic sense. Adaptive not necessarily imply active changing of connections in a neuroscientific sense. This type of adaptation can occur through a passive non-linear transfer function that accounts for the stimulus condition and the structural connectivity certainly does not change in the timeframes measured in functional neuroscientific studies. MCPA provides a platform for examining the role of interregional connectivity patterns in this type of adaptive process. Indeed, MCPA can be interpreted as testing whether distributed computational ”work” is being done in the interaction between the two populations (Friston et al., 1997) and the interaction does not just reflect a passive relay of information between two encapsulated modules (Fodor, 1983).

Passive linear filters do not allow for information to be added to the representation through computational work being done in the interaction between regions. Sensitivity to this type of computation is a central appeal of fully non-linear models of neural representation and neural interactions, such as deep neural network approaches. However, these approaches often require tens of thousands or even millions of trials before they achieve good performance (Goodfellow et al., 2016), which is impractical for most neuroscientific applications. MCPA is not sensitive to multivariate non-linear interactions within conditions, but is sensitive to multivariate non-linear relationships between the interregional interaction pattern and the conditions. This is effectively a piecewise linear approximation of the underlying nonlinear function relating the condition space to the interaction pattern between regions. This restriction relative to deep neural network and other non-linear function approximation approaches allows MCPA to perform well with reasonable numbers of trials (10s of trials in our examples), which is critical for being practically useful in neuroscience. Thus, one strength of MCPA is the ability to capture some key aspects of non-linear neural computations without requiring an impractical amount of data.

5.4.2 MCPA and representation space

In addition to allowing one to infer whether distributed computational work is being done in service of information processing, MCPA provides a platform for assessing its representational structure (Haxby et al., 2014). Much as MVPA has been used in representational similarity analyses to measure the structure of the representational space at the level of local neural populations (Edelman et al., 1998; Kriegeskorte, 2011; Kriegeskorte and Kievit, 2013), MCPA can be used to measure the structure of the representational space at the level of network interactions. Specifically, the representational geometry of the interaction can be mapped in terms of the similarity among the multivariate functional connectivity patterns corresponding to the brain states associated with varying input information. The representational structure can be compared to behavioral measures of the structure to make brain-behavior inferences and assess what aspects of behavior a neural interaction contributes to. It can also be compared to models of the structure to test theoretical hypotheses regarding the computational role of the neural interaction (Kriegeskorte, 2011; Kriegeskorte et al., 2008). By comparing the representational space in models to the neural representation, one can assess how well these models approximate the neural representation in both absolute and relative terms. Much the way MVPA-based RSA analyses have been used to examine these models at the level of individual brain regions (Kriegeskorte et al., 2008), RSA analyses can be used to assess how well the representation inferred by these models transfer functions fit the representation measured in the brain using MCPA.

The MCPA-based RSA analysis presented here relating the representational space derived from the interaction between regions of the visual processing stream to the transformation operations in HMAX is a concrete example of how MCPA can be used to test models of how representations are transformed between regions. This example also helps illustrate the underlying hypothesis being tested by MCPA: that there is a non-constant linear function that relates how the transformation of the activity between regions changes with respect to the experimental condition. A non-constant linear function is analogous to a local linear approximation of a non-linear function, as we have seen in the example of HMAX. The existence of this non-constant linear function is what allows for information to be added to the representation through distributed computational work. By comparing the MCPA-based representational space to models of this function, we can gain insight into what this transformation function might be. For example, in the case of the S1-C1 transformation HMAX, this function is a local, non-linear max-pooling operation and in the case of the S2-C2 operation it is a more global, non-linear max-pooling operation (Riesenhuber and Poggio, 1999). Furthermore, this is why MCPA could not be compared to the transformation between the C1 and S2 layers of the HMAX model because the transformation between those layers is a passive filter operation, e.g. a trivial, constant linear function relating the between layer transformation to the stimulus condition. This example suggests one mechanism by which a network with fixed structural connectivity can give rise to adaptive communication, namely through a non-linear transformation operation that are adaptive in a linear sense. In addition to testing specific hypothesis-driven transformation operations, such as the ones in HMAX, more data-driven models of the transformation operations, such as ones in deep neural network models (Yamins et al., 2014), could also be tested using the MCPA-based RSA approach.

5.4.3 Relationship between MCPA and other functional connectivity/multivariate methods

These two properties of MCPA, 1) being able to assess distributed computational processing rather than just whether or not areas are communicating and 2) being able to determine the representational structure of the information being processed, set MCPA apart from previously proposed functional connectivity methods. In these previous methods the functional connectivity calculation is performed separately from the classification calculation. Specifically, either functional connectivity is first calculated using standard methods, then a model is built on the population of connectivity values and this model is tested using classification approaches (Finn et al., 2015; Richiardi et al., 2011; Rosenberg et al., 2016; Shirer et al., 2012; Wang et al., 2015) or the model is first built on the activity in each region and tested using classification approaches and the classification performance is correlated (Coutanche and Thompson-Schill, 2013; Kriegeskorte and Kievit, 2013). These methods are very useful for assessing how differences in large-scale patterns of connectivity relate to individual subject characteristics (e.g. connectome fingerprinting) in the first case and comparing the representational structure between regions in the second case. In contrast, in MCPA the model is the connectivity map and classification is done to directly test the information contained in these maps. The separation of the connectivity and classification calculations in other approaches precludes being able to assess distributed computational processes because these methods are sensitive to passive information exchange between encapsulated modules, as described above, and thus conflate passive and adaptive communication. Critically, they do not specifically probe how connectivity patterns change as a factor of condition or state, as is required to efficiently perform the representational similarity analysis in a practical manner and decode how the information processed in the interaction is encoded and organized. As a concrete example, these previous methods would not be able to compare the representational structure of the neural interaction between regions to the structure from a computational model, as was done here with fMRI.

MCPA can be roughly considered a multivariate extension of PPI with the addition of a prediction and classification framework. Compared to PPI, which is univariate, MCPA allows one to exploit the multivariate space of interaction patterns. As a result, MCPA is sensitive to aspects of information coded in interregional interactions that PPI may not be able to detect (Norman et al., 2006), for example in event-related fMRI designs where PPI is known to lack statistical power (O'Reilly et al., 2012). Indeed, in the fMRI data presented here, PPI was no better than chance in detecting interregional interactions at the visual category level, whereas MCPA was significantly better than chance. Much the way MVPA allows one to go beyond ANOVAs/t-tests in a single area/population (e.g. single trial classification, RSA, complex model testing), MCPA allows one to go beyond PPI and do these types of analyses at the level of the shared activity between regions.

The specific instantiation of MCPA presented here treats connectivity as a bi-directional linear mapping between two populations. However, the MCPA framework could be easily generalized into more complicated cases. For example, instead of using correlation-based methods like CCA, other directed functional connectivity algorithms, such as Granger causality based on an autoregressive framework, potentially using partial CCA for the time-lagged autoregressive step, could be used to examine directional interactions. This would allow one to examine time-

lagged multivariate connectivity patterns to infer directionality. Additionally, kernel methods, such as kernel CCA (Hardoon et al., 2004), or deep learning methods, such as deep CCA (Andrew et al., 2013), could be applied to account for non-linear interactions. Another possible and more general framework would be to use non-parametric functional regression method to build a functional mapping between the two multidimensional spaces in the two populations. MCPA can also be expanded to look at network-level representation by implementing the multiset canonical correlation analysis, wherein the cross-correlation among multiple sets of activity patterns from different brain areas is calculated (Kettenring, 1971). MCPA could be used with a dual searchlight approach to examine whole brain communication (Kriegeskorte et al., 2006). Also, MCPA could be adapted by optimizing the CCA to find the connectivity maps that uniquely describe, or at least best separate, the conditions of interest. Furthermore, both with and without these modification, the framework of MCPA may have a number of applications outside of assessing the representational content of functional interactions in the brain, such as detecting the presence of distributed processing on a computer network, or examining genetic or proteomic interactions. MCPA is used here with fMRI BOLD signals and iEEG signal, but it can be applied to nearly any neural recording modality, including scalp electroencephalography, magnetoencephalography, multiunit firing patterns, single unit firing patterns, spike-field coherence patterns, to assess the information processed by cross-frequency coupling, etc.

5.4.4 Limitations and implication from MCPA results

One caveat with the MCPA results with real data presented here is that many of the effect sizes are small. One likely reason for this is that for the decoding of individual images in fMRI and faces in iEEG the number of trials per image was very small (13 for individual images in fMRI and 15 for individual faces in iEEG). Despite the small number of trials, the classification accuracy is roughly on a par with previous exemplar-level individuation classification results using fMRI and iEEG (Ghuman et al., 2014; Nestor et al., 2011; Said et al., 2010; Skerry and Saxe, 2014). Furthermore, the HMAX-MCPA correlation is roughly on par with previously reported correlations between HMAX and single unit activity from non-human primates (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2013). Given a larger number of trials, MCPA classification performance should improve. The classification performance seen here can be considered a ”worst case scenario” to some extent given the low number of trials and yet performance still was not far below what has been previously reported using multivariate classification on these types of data. Nonetheless, the low effect size and small number of subjects reported here is a strong caveat to the potential neuroscientific interpretation of the fMRI and iEEG data.

The MCPA results from visual cortex show that the representational space derived from MCPA was negatively correlated to the representational space derived from MVPA from either of the local populations. This inverse relationship is consistent with the idea that the communication between regions represents information that has not been explained by local computational processes. With the strong caveat that these results require replication in more subjects and assessment with paradigms designed to directly test these hypotheses, this negative correlation is consistent with the hypothesis that neural interactions code for information not resolved in local computational processes (Friston, 2010; Lee and Mumford, 2003; Rumelhart et al., 1988).

The current prevalent view is that face perception is mediated by a distributed network with

multiple brain areas including the OFA and FFA. Structural and functional connectivity analysis for the core network has shown that FFA is strongly connected to OFA (Gschwind et al., 2011; Ishai, 2008; Pyles et al., 2013). While these results suggest the hypothesis that face individuation may involve the interaction between these populations (and likely other face processing regions), direct evidence for this hypothesis has been lacking. Our results here support the hypothesis that individual-level facial information is not only encoded by the activity within certain brain populations, but also represented through recurrent interactions between multiple populations at a network level. This interaction was biased towards frequencies in the Beta and low Gamma bands and exhibited a degree of cross-frequency coupling. This analysis indicates that assessing cross-frequency interactions between regions is another potential application of MCPA. In addition, MCPA showed significant face individuation in approximately the 200-500 ms time window after stimulus onset, but did not show any significant face individuation in the early time window (50-200 ms after stimulus onset), which is consistent with a previous MVPA study based on iEEG recording from FFA only (Ghuman et al., 2014). More broadly, the fMRI and iEEG MCPA results suggest that the computational work done in service of visual processing occurs not only on the local level, but also at the level of distributed brain circuits.

Previously, multivariate pattern analysis methods have been used to analyze the sensitivity to information within a certain area and functional connectivity methods have been used to assess whether or not brain networks participate in a particular process. With MCPA, the two perspectives are merged into one algorithm, which extends multivariate pattern analysis to enable the detailed examination of information sensitivity at the network level. Thus, the introduction of MCPA provides a platform for examining how computation is carried out through the interactions between different brain areas, allowing us to directly test hypotheses regarding circuit-level information processing.

5.5 Appendix: Supplement Figures

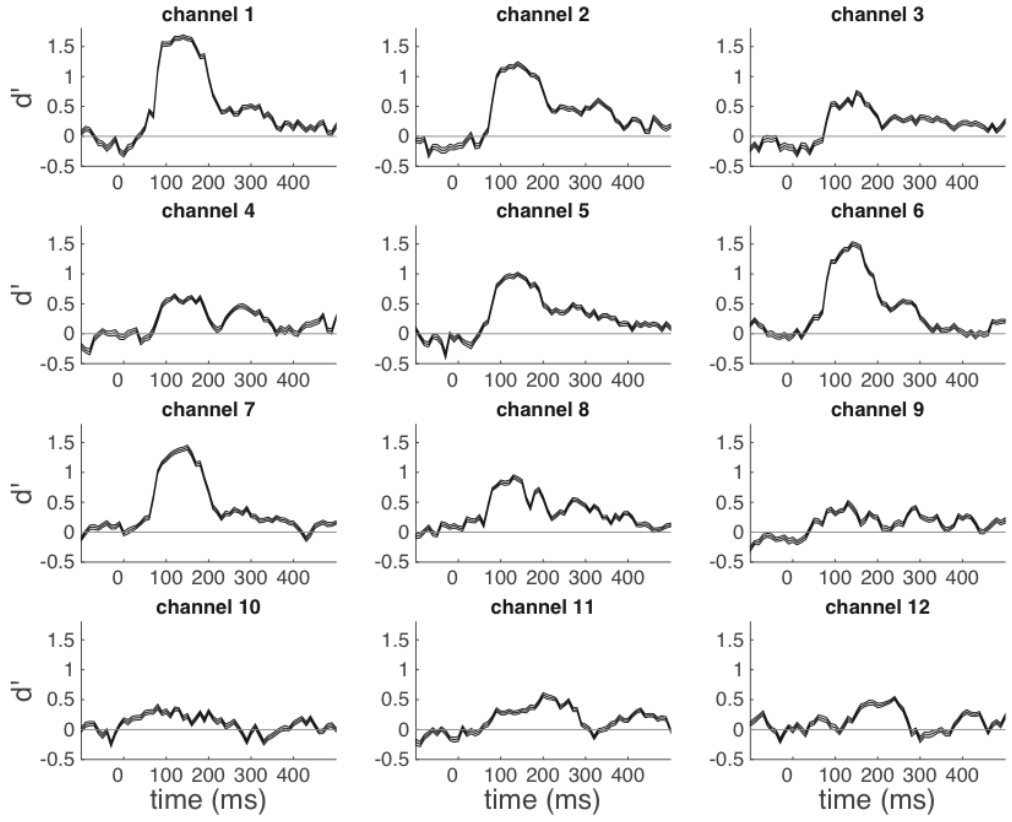


Figure 5.6: **iEEG Single electrode face sensitivity.** Time course of face categorical sensitivity in each single electrode measured by sensitivity index d' (mean d' plotted against the beginning of the 100 ms sliding window). The classifier uses time-windowed ERP signal from a single electrode (window length = 100 ms) as input features (See Methods for details). Horizontal grey line indicates chance level ($d' = 0$). The channels are labeled 1-12 from anterior to posterior. Electrodes were chosen based on the criteria that peak d' be above 1 ($p < 0.001$, channels 1, 6, and 7). Channel number 1 was used as the FFA electrode and channels 6 and 7 were used for the OFA electrodes based on their anatomical locations.

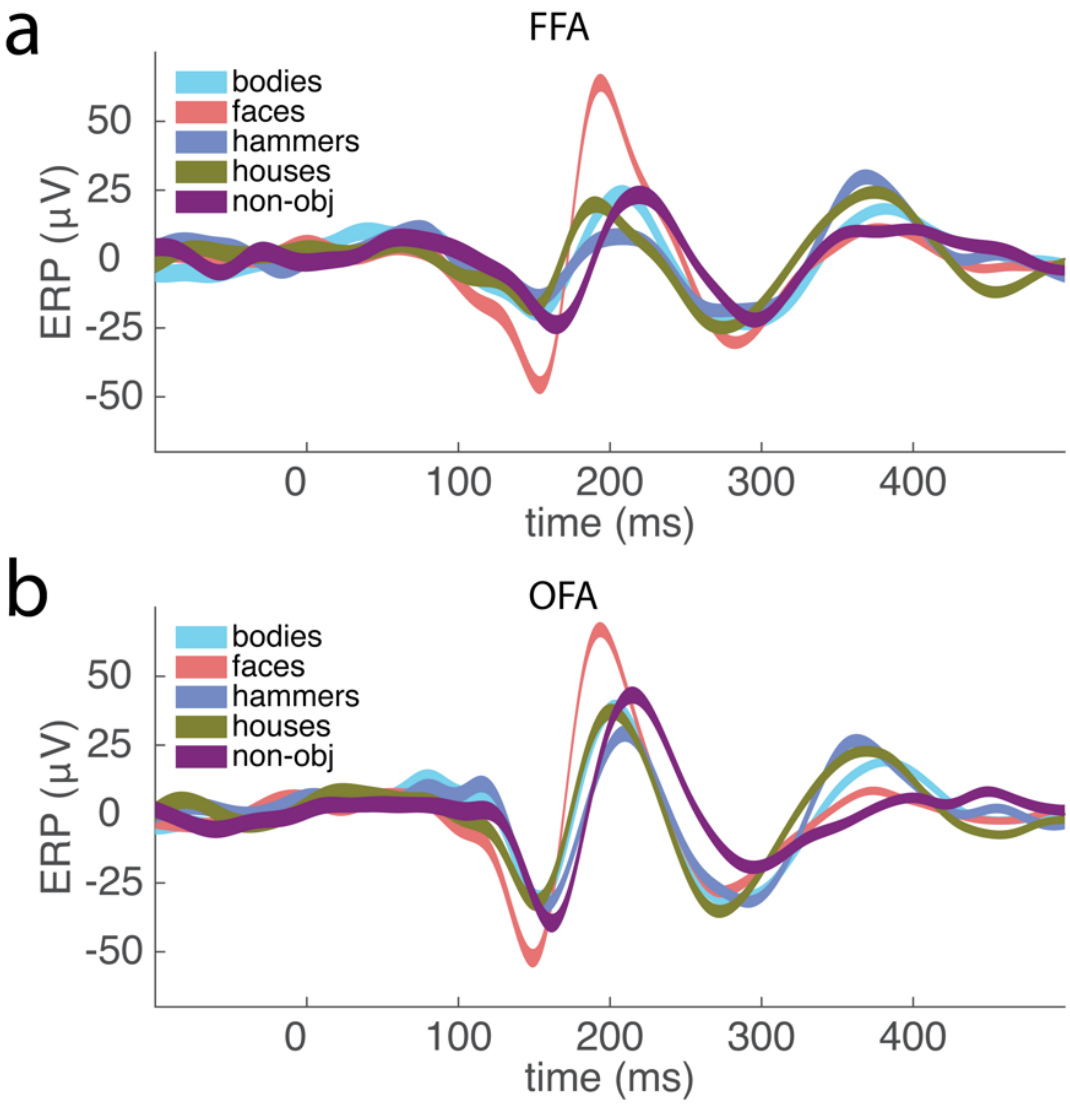


Figure 5.7: **Face selectivity in FFA and OFA.** Averaged ERP response recorded from FFA and OFA contacts for each category during the localizer task. The colored area corresponds to the standard error.

Chapter 6

Pre-stimulus modulation of the post-stimulus temporal dynamics

In previous chapters, we studies the representation structure of the event-related neural activity within regions of interest, as well as the interaction patterns between areas. In this last part of the thesis, we extend the scope of our analysis to the neural activity before the onset of the stimuli. Perception reflects not only on processing sensations, but also the endogenous neural state when sensory inputs enter the brain. However, the neural mechanism by which endogenous neural states influence perception remains largely unknown. Results from 246 electrodes implanted in visual processing regions in the brains of 30 humans shows that endogenous activity modulates the sharpness of trial-by-trial neural tuning to visual inputs, a property of the brain related to the quality of the visual representation. Furthermore, the same aspect of the endogenous activity that influences visual tuning also predicts trial-by-trial reaction time in a perceptual task. These results provide evidence for a neural mechanism that links endogenous neural states to the quality of the neural representation, which in turn influences perception.

6.1 Introduction

Neural responses in perceptual circuits with respect to sensory inputs give rise to cognitive perception of the stimuli. However, even with the same stimulus input, the corresponding neural responses can be extremely variable. Therefore, perception depends on not only sensory input, but also the neural and cognitive state when a stimulus is presented. Traditionally, this endogenous activity has been treated as noise to be discarded and averaged over. However, it is becoming increasingly clear that endogenous fluctuations in neural activity influence both the neural response to sensory input and behavior. Previous studies has linked the endogenous activity to global cognitive states, such as attention or arousal, as well as infra-slow resting-state fluctuations (Fox et al., 2006; Kastner et al., 1999; Klimesch, 1999; Seeley et al., 2007). There have been a number of studies showing the correlation between endogenous activity and post-stimulus neural response. It has been shown that the amplitude, variance, and patterns in the post-stimulus neural response are shaped by the endogenous ongoing activity (Arieli et al., 1996; Başar, 1980; Fox et al., 2006; Henriksson et al., 2015; Kisley and Gerstein, 1999). Furthermore, it has also

been shown that endogenous activity can also explain the variance in behavior across a number of perceptual domains including vision and audition (Busch et al., 2009; Kayser et al., 2016; Mathewson et al., 2009; Ress et al., 2000; Thut et al., 2006; VanRullen et al., 2011). Taken together, these results suggest a possible link between the endogenous activity and sensory perception of stimuli, in that the spontaneous activity can potentiate category recognition via modulating the post-stimulus activity.

However, how endogenous fluctuations of neural activity influence perception remains unknown because little empirical evidence exists linking these fluctuations to an aspect of the neural response associated with the quality of the perceptual representation. Specifically, it remains unknown if there is a link between endogenous activity and the sharpness of neural tuning that ultimately influences behavioral performance. This link would provide a mechanistic bridge between endogenous neural states and perception.

In this study, we used direct recording from category selective regions in a large cohort of human subjects using intracranial electroencephalography (iEEG) to probe the relationship between endogenous activity and category neural tuning, as well as behavioral perception. Specifically, we directly tested and verified two main hypotheses. First, the endogenous activity modulates the degree of category tuning in response to visual stimuli; second, the same aspect in endogenous activity that modulates tuning also correlates with behavioral perception. We further evaluated the spatial and temporal specificity of the endogenous modulation signal. The results suggest that the endogenous modulation effect is a reflection of local processes, and the majority of the endogenous modulation effect are transient, with a small fraction of all the channels showing trial-by-trial auto-correlation in endogenous modulation of tuning.

6.2 Methods

6.2.1 Subjects

The experimental protocols were approved by the Institutional Review Board of the University of Pittsburgh. Written informed consent was obtained from all participants. 30 human subjects (11 male, 19 female) underwent surgical placement of subdural electrocorticographic electrodes or stereoelectroencephalography (together electrocorticography and stereoelectroencephalography are referred to here as iEEG) as standard of care for seizure onset zone localization. The ages of the subjects ranged from 19 to 64 years old (mean $= 38.2$, S.D. $= 11.9$). None of the participants showed evidence of epileptic activity on the fusiform electrodes used in this study nor any ictal events during experimental sessions.

6.2.2 Stimuli

In each session, 180 images of faces (50% male), bodies (50% male), words, hammers, houses, and phase scrambled faces were used as visual stimuli. Each of the six categories contained 30 images, and each image was presented twice. At random, 1/3 of the time an image would be repeated, which yielded 480 independent trials in each.

6.2.3 Paradigms

In the experiment, each image was presented for 900 ms with 900 ms inter-trial interval during which a fixation cross was presented at the center of the screen ($\sim 10^\circ \times 10^\circ$ of visual angle). At random, 1/3 of the time an image would be repeated, which yielded 480 independent trials in each session. Participants were instructed to press a button on a button box when an image was repeated (1-back), and their reaction time between stimulus onset and button press was recorded. Paradigms were programmed in MATLAB using Psychtoolbox and custom written code. All stimuli were presented on an LCD computer screen placed approximately 150 cm from participants' heads.

6.2.4 Data analysis

Data preprocessing

The electrophysiological activity was recorded using iEEG electrodes at 1000 Hz. Common reference and ground electrodes were placed subdurally at a location distant from any recording electrodes, with contacts oriented toward the dura. Single-trial potential signal was extracted by band-passing filtering the raw data between 0.2-115 Hz using a fourth order Butterworth filter to remove slow and linear drift, and high frequency noise. The 60 Hz line noise was removed using a fourth order Butterworth filter with 55-65 Hz stop-band. Power spectrum density (PSD) at 2-100 Hz with bin size of 2 Hz and time-step size of 10 ms was estimated for each trial using multi-taper power spectrum analysis with Hann tapers, using FieldTrip toolbox (Oostenveld et al., 2011). We define the neural activity within the [-500, -100] ms interval relative to the stimulus onset as the pre-stimulus activity, and the neural activity within the [100, 500] ms interval relative to the stimulus onset as the post-stimulus activity. For each channel, the PSD at each frequency was z-scored with respect to the mean and variance of the baseline activity to correct for the power scaling over frequency at each channel. The broadband gamma signal was extracted as mean z-scored PSD across 40-100 Hz. Event-related potential (ERP) and event-related broadband gamma signal (ERBB), both time-locked to the onset of stimulus from each trial, were used in the following data analysis. Specifically, the ERP signal is sampled at 1000 Hz and the ERBB is sampled at 100 Hz. In addition to the potential signal and PSD, the pre-stimulus phase information was also extracted from each trial. Specifically, discrete time Fourier transform was applied to the raw signal in the [-500, -100] ms time interval, which had a total length of 400 points sampled at 1000 Hz. As a result, the phase information between 0-1000 Hz was extracted with a step-size of 2.5 Hz. Finally, we used the phases from 0 to 150 Hz as the pre-stimulus phase features.

To reduce potential artifacts in the data, raw data were inspected for ictal events, and none were found during experimental recordings. Trials with maximum amplitude 5 standard deviations above the mean across all the trials were eliminated. In addition, trials with a change of more than 25 μV between consecutive sampling points were eliminated. These criteria resulted in the elimination of less than 1% of trials.

Electrode localization

Coregistration of grid electrodes and electrode strips was adapted from the method of Hermes et al. (2010). Electrode contacts were segmented from high resolution post-operative CT scans of patients coregistered with anatomical MRI scans before neurosurgery and electrode implantation. The Hermes method accounts for shifts in electrode location due to the deformation of the cortex by utilizing reconstructions of the cortical surface with FreeSurfer™ software and co-registering these reconstructions with a high-resolution post-operative CT scan. SEEG electrodes were localized with Brainstorm software (Tadel et al., 2011) using post-operative MRI co-registered with pre-operative MRI images.

Electrode selection

Category-selective electrodes were selected based on a 6-way classifier. Specifically, we train a multinomial logistic regression model to classify the post-stimulus neural activity with respect to the 6 different categories from each other. The sensitivity index (d') for each category was then computed as $d' = Z(\text{true positive rate}) - Z(\text{false positive rate})$, where $Z(x)$ is the inverse function of the cumulative density function of standard normal distribution. An electrode is selected as category-selective if the maximum d' across all categories is greater than 0.5 ($p < 0.01$, permutation test). The selected electrode is then assigned to the category with maximum d'.

Two-stage generalized linear model (GLM)

We considered the neural activity within the $[-500, -100]$ ms pre-stimulus time interval as proxy for the endogenous activity, noted as $X_{pre} \in \mathbb{R}^{N \times T_1}$, where N is the number of trials and T_1 is the number of features in the pre-stimulus time window; and we used neural activity from the $[100, 500]$ ms time interval relative to stimulus onset as the post-stimulus evoked activity that encodes category information, noted as $X_{evk} \in \mathbb{R}^{N \times T_2}$, where T_2 is the number of features in the post-stimulus time window. A GLM (6.1) was used to represent category tuning. Specifically,

$$p(y|X_{evk}, X_{pre}) = f(X_{evk}\beta_{evk}, X_{pre}\beta_{pre}) \tag{6.1}$$

where

$$p(y|X_{evk}, X_{pre}) = \frac{1}{1 + \exp\left(-(X_{evk}\beta_{evk} - X_{pre}\beta_{pre})\right)} \tag{6.2}$$

is sigmoid function, and $\beta_{pre} \in \mathbb{R}^{T_1+1}, \beta_{evk} \in \mathbb{R}^{T_2+1}$ are weight vectors for the pre- and post-stimulus features and the intercepts.

The model decodes the category y from the post-stimulus neural activity X_{evk}, conditioning on the pre-stimulus activity X_{pre} in each of the category-selective electrode. Considering the high dimensional settings and strong temporal correlation in the features, elastic-net penalty was used to regularize model (6.1) (Hastie et al., 2015). Therefore, fitting the model requires solving the following optimization problem:

$$\underset{\beta_{evk}, \beta_{pre}}{\text{argmin}} \quad -\ell(\beta_{evk}, \beta_{pre}) + \lambda_1 P_\alpha^{evk}(\beta_{evk}) + \lambda_2 P_\alpha^{pre}(\beta_{pre}) \tag{6.3}$$

where

$$\ell(\beta_{evk}, \beta_{pre}) = \frac{1}{N}\{y^T(X_{evk}\beta_{evk} - X_{pre}\beta_{pre}) - \mathbf{1}^T \log(1 + \exp(X_{evk}\beta_{evk} - X_{pre}\beta_{pre}))\} \quad (6.4)$$

is the log likelihood of the GLM.

$$P_\alpha^{evk}(\beta_{evk}) = \frac{1-\alpha}{2}\|\beta_{evk}\|_2^2 + \alpha\|\beta_{evk}\|_1 \quad (6.5)$$

is the standard elastic-net penalty (Zou and Hastie, 2005), and $P_\alpha^{evk}(\beta_{evk})$ is a similar elastic-net penalty but with group structure to account for the phase features (see below for a detailed description of the penalty structure).

The model was fitted in a two-stage manner (Algorithm 2). In the first step, we forced $\beta_{pre} = 0$ and only searched for optimal dimension β_{evk}^* in the post-stimulus activity that best discriminates between categories. In other words, we solve the following standard elastic-net problem, which can be solved using coordinate descent (Friedman et al., 2010) (see Appendix for details):

$$\beta_{evk}^* = \underset{\beta_{evk}}{\text{argmin}} \quad -\ell(\beta_{evk}, \beta_{pre} = 0) + \lambda_1 P_\alpha^{evk}(\beta_{evk}) \quad (6.6)$$

Solving (6.6) would result in a trial-by-trial neural metric, $X_{evk}\beta_{evk}$, which quantified the post-stimulus category selectivity. In the second step, we fixed the optimal dimension β_{evk}^* and optimized the model with respect to β_{pre}. Specifically, this is equivalent to solving the following group elastic-net problem, which can be solved using block coordinate descent, which is similar to the approaches described in Meier et al. (2008); Simon and Tibshirani (2012) (see Appendix for details):

$$\beta_{pre}^* = \underset{\beta_{pre}}{\text{argmin}} \quad -\ell(\beta_{evk}^*, \beta_{pre}) + \lambda_2 P_\alpha^{pre}(\beta_{pre}) \quad (6.7)$$

This allowed the category classification to be made conditioning on the pre-stimulus activity, and critically, provided a neural metric $X_{pre}\beta_{pre}$ in pre-stimulus activity that quantifies the amount of influence from pre-stimulus activity on the post-stimulus category selectivity on a trial-by-trial basis. We defined $MI = X_{pre}\beta_{pre}$ as the pre-stimulus modulation index (MI).

Path algorithm with decreasing λ is often used in solving such regularized optimization problem, and techniques such as warm starts, active sets, etc. are also used to speed up the algorithm. The optimal regularization parameters λ_1 and λ_2 were selected using cross-validation based on minimizing the deviance along the path. And the performance of the model can be evaluated using a held-out testing set or another level of cross-validation.

The group elastic-net penalty

For the post-stimulus part, we only consider the ERP and ERBB features, noted as $x_{evk} = [x_{evk}^{ERP}, x_{evk}^{ERBB}]$, with the corresponding weights $\beta_{evk} = [\beta_0^{evk}, \beta_{evk}^{ERP}, \beta_{evk}^{ERBB}]$, and we applied regularization term

$$P_\alpha^{evk}(\beta_{evk}) = \frac{1-\alpha}{2}\|\beta_{evk}\|_2^2 + \alpha\|\beta_{evk}^{ERP}\|_1 + \alpha\|\beta_{evk}^{ERBB}\|_1 \quad (6.8)$$

in (6.3). For the pre-stimulus part, we use ERP, ERBB and phase features, noted as $x_{pre} = [x_{pre}^{ERP}, x_{pre}^{ERBB}, x_{pre}^{phase}]$, and the corresponding weights $\beta_{pre} = [\beta_0^{pre}, \beta_{pre}^{ERP}, \beta_{pre}^{ERBB}, \beta_{pre}^{phase}]$. Assume that we have phase $[\theta_1, , \theta_K]$, where $\theta \in [-\pi, \pi)$, corresponding to frequencies of interest $[f_1, , f_K]$. To transfer the circular phase value onto the real axis in order to facilitate the ℓ_1-norm penalty, we consider feature vector $x_{pre}^{phase} = [\sin\theta_1, \cos\theta_1, ..., \sin\theta_K, \cos\theta_K]$, where $\sin\theta, \cos\theta \in [-1, 1]$, and group lasso penalty term

$$\mathcal{G}(\beta_{pre}^{phase}) = \sqrt{2}\sum_{i=1}^{K}\sqrt{{\beta_{pre,(i,1)}^{phase}}^2 + {\beta_{pre,(i,2)}^{phase}}^2} \tag{6.9}$$

where $[\beta_{pre,(i,1)}^{phase}, \beta_{pre,(i,2)}^{phase}]$ are the pair of weights corresponding to phase feature pair $[\sin\theta_1, \cos\theta_1]$. Here group-lasso penalty is applied to the sin-cos pair to ensure a uniform penalty on all $\theta \in [-\pi, \pi)$. As a result, the group elastic-net penalty for the pre-stimulus weights can be written as

$$P_\alpha^{pre}(\beta_{pre}) = \frac{1-\alpha}{2}\|\beta_{pre}\|_2^2 + \alpha\|\beta_{pre}^{ERP}\|_1 + \alpha\|\beta_{pre}^{ERBB}\|_1 + \alpha\mathcal{G}(\beta_{pre}^{phase}) \tag{6.10}$$

Algorithm 2: Training the two-stage GLM

Data: data matrices $X_{pre} \in \mathbb{R}^{N\times T_1}, X_{evk} \in \mathbb{R}^{N\times T_2}$ for pre-stimulus and post-stimulus data, data label $y \in \mathbb{R}^N$;
where N is the number of samples, $T_1 = t_{pre}^{ERP} + t_{pre}^{ERBB} + t_{pre}^{phase}$, $T_2 = t_{evk}^{ERP} + t_{evk}^{ERBB}$, and $X_{pre} = [X_{pre}^{ERP}, X_{pre}^{ERBB}, X_{pre}^{phase}]$, $X_{evk} = [X_{evk}^{ERP}, X_{evk}^{ERBB}]$;
Parameters: the elastic-net hyper-parameter α, maximum regularization parameter λ_{max} and minimum regularization parameter $\epsilon\lambda_{max}$.
Result: Weight vectors for pre- and post-stimulus features
$\beta_{pre}^* = [\beta_0^{pre}, \beta_{pre}^{ERP}, \beta_{pre}^{ERBB}, \beta_{pre}^{phase}]$, $\beta_{evk}^* = [\beta_0^{evk}, \beta_{evk}^{ERP}, \beta_{evk}^{ERBB}]$

Fit the elastic-net problem for post-stimulus features:
for *the i-th cross-validation split* $\{X_{evk,train}^{(i)}, X_{evk,test}^{(i)}\}$ **do**
  for $\lambda \leftarrow \lambda_{max}$ **to** $\epsilon\lambda_{max}$ *(decrement λ)* **do**
    solve elastic-net problem (6.6) using coordinate descent ;
    (see Appendix for detailed derivation of the coordinate descent)
    estimate the deviance of the solution on $X_{evk,test}^{(i)}$
find optimal β_{evk}^* that minimizes deviance;
Fix the post-stimulus part and fit the group elastic-net problem for pre-stimulus features:
for *the i-th cross-validation split* $\{X_{pre,train}^{(i)}, X_{pre,test}^{(i)}\}$ **do**
  for $\lambda \leftarrow \lambda_{max}$ **to** $\epsilon\lambda_{max}$ *(decrement λ)* **do**
    solve group elastic-net problem (6.7) using block coordinate descent ;
    (see Appendix for detailed derivation of the block coordinate descent)
    estimate the deviance of the solution on $X_{pre,test}^{(i)}$
find optimal β_{pre}^* that minimizes deviance.

Cross-electrode correlation in pre-stimulus MI

To evaluate the spatial properties of the pre-stimulus modulation effect, we computed the correlation of the single trial pre-stimulus MI between category-selective electrodes in each subject. For the i-th category-selective electrode, we got $MI_i = X_{pre,i}\beta_{pre,i}$ from the GLM. The cross-electrode correlation between two category-selective electrodes i and j was estimated by computing the correlation coefficient between MI_i and MI_j across all trials. To avoid confounding effect from local spatial correlation between two nearby electrodes, we only considered a pair of electrodes that were $>$ 2cm apart from each other. For each subject, the mean cross-electrode correlation was estimated by averaging the pairwise correlation coefficients across all such pairs of category-selective electrodes.

Autocorrelation in pre-stimulus MI

To evaluate the temporal properties of the pre-stimulus modulation effect, we computed the autocorrelation of the single trial pre-stimulus MI between consecutive trials with lags ranging from 1 to 20 in each category-selective electrodes. Specifically, for any given electrode, the autocorrelation with lag k is

$$r_k = \frac{\sum_{t=1}^{N-k}(MI^{(t)} - \overline{MI})(MI^{(t+k)} - \overline{MI})}{\sum_{t=1}^{N}(MI^{(t)} - \overline{MI})(MI^{(t)} - \overline{MI})}$$

To evaluate the temporal property, we tested for the significance of the first-order autocorrelation, since it is essential for any temporal dependencies caused by slow-fluctuation in the signal. Specifically, the upper bound of the 95% confidence interval was approximately estimated as $2/\sqrt{N}$ where N is the total number of trials.

Permutation test

Permutation tests were used to test for significance of the effects in this study. In order to construct a surrogated distribution of the pre-stimulus MI, in each permutation we generated random projection weight vectors $\beta_{rand}^{(i)} \in \mathbb{R}^{T_1}$, such that they all had the same number of non-zero elements as the actual solution from the real data, i.e. $\|\beta_{rand}^{(i)}\|_0 = \|\beta_{pre}^{*}\|_0$. We repeated this process 1000 times for each electrode, and the histogram of $X_{pre}\beta_{rand}^{(i)}$, $i = 1, ..., 1000$, was used as the null distribution of the pre-stimulus MI.

6.3 Results

6.3.1 Category-selective electrodes

From the 30 patients, we located 246 channels that demonstrated category-selectivity to one of the six categories (Figure 6.1, Table 6.1). Among all the channels, 230 were located in the ventral temporal cortex. As shown in Figure 6.1, the category selective electrodes covered bilateral ventral temporal cortex.

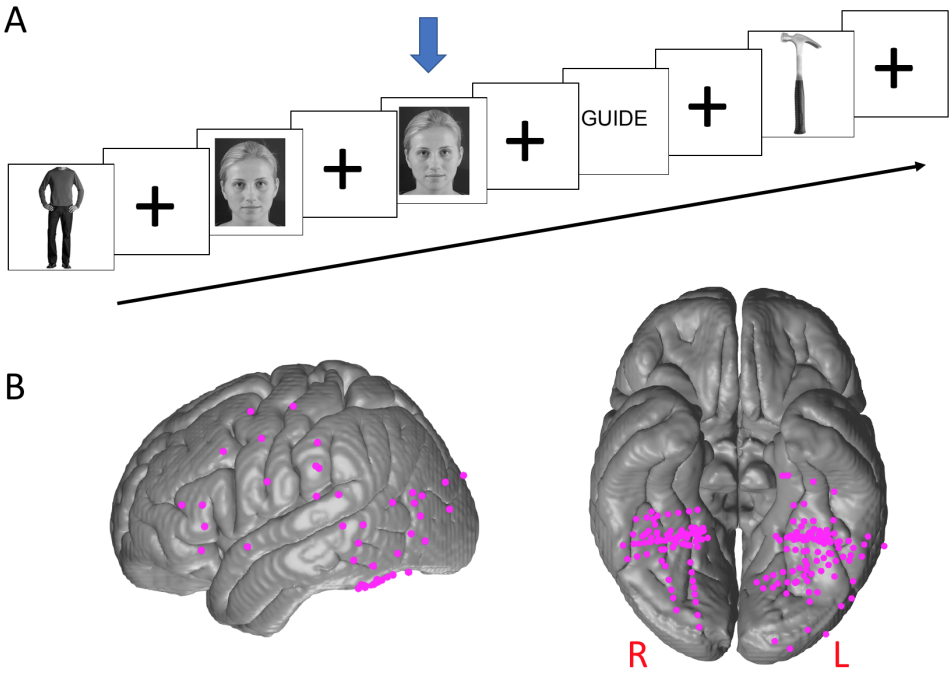


Figure 6.1: **Experiment paradigm and electrode locations. A)** Experiment paradigm in which the subject is shown a series of images and performs a 1-back repeat detection task. **B)** The lateral and ventral views of the locations of the 246 category-selective electrodes mapped onto a common brain surface.

6.3.2 Endogenous activity modulates category tuning

First, we examined whether conditioning on the pre-stimulus endogenous activity changes the classification accuracy. In each category selective electrode, we applied the two-stage GLM to the iEEG data to decode the preferred category from the others. For each of the six categories, we collected all the category-sensitive electrodes corresponding to the category, and computed the averaged classification sensitivity index (d') across all these electrodes. The two-stage GLM provides two different classifiers: the first step of the optimization results in a classifier that only extracts discriminant features from the post-stimulus evoked response; the second step of the optimization effectively trains a classifier with the post-stimulus discriminant features conditioning on pre-stimulus activities. The averaged d's from the two classifiers in the GLM were then compared against each other to test the first hypothesis. As shown in Figure 6.2A and Table 6.1, we found that the inclusion of pre-stimulus activity improved the classification accuracy for all visual categories. Specifically, across all category-selective electrodes, the mean sensitivity index $d' = 1.04$ with only post-stimulus activity. Conditioning on pre-stimulus activity, the mean d' improved to 1.17 ($p < 1 \times 10^{-5}$, paired t-test). This result confirmed that, at a functional level, conditioning on pre-stimulus endogenous activity improves trial-by-trial category tuning.

Table 6.1: Number of electrodes showing significant category sensitivity for each of the stimulus categories, and the comparisons of classification results from the two-stage GLM

Category	Bodies	Faces	Words	Tools	Houses	Scrambled non-objects
# of electrodes	9	56	92	16	37	36
d' (evoked only)	1.1822	1.3957	0.9252	0.7289	1.0585	0.8219
d' (evoked + endogenous)	1.3093	1.5072	1.0628	0.8334	1.2046	1.0105
p-value	0.0264	$< 10^{-5}$	$< 10^{-5}$	0.0024	$< 10^{-5}$	$< 10^{-5}$

6.3.3 Endogenous activity influences perceptual behavior

Using the two-stage GLM, we extracted pre-stimulus modulation index (MI) for each single trial, which is a neurological metric that quantifies the influence of the endogenous activity on the post-stimulus discriminant features. Next, we tested if the same aspect of pre-stimulus activity that modulates post-stimulus category tuning also correlates to behavioral perception. Trials were divided into whether they came from the preferred condition for a particular electrode (face trials for electrodes recording from face sensitive regions, word trials for electrodes recording from word sensitive regions, etc.) or the non-preferred condition for that electrode (non-face trials for electrodes recording from face sensitive regions, etc.). To assess this, we further divided the trials from each condition into two groups purely by the magnitude of the trial-by-trial pre-stimulus MI derived from the two-stage GLM in each category-sensitive electrode, and compared the averaged reaction time in the 1-back task in the low MI group and the high MI group across all category-sensitive electrodes. Trial-by-trial reaction time was significantly correlated to the MI for the preferred conditions across electrodes ($R^2 = 0.067$, $p = 0.0059$), but not the non-preferred conditions ($R^2 = 0.015$, $p = 0.71$). As shown in Figure 6.2B, for the preferred conditions, the mean reaction time for the high pre-stimulus modulation trials was 663.2 ms, while the mean reaction time for the low pre-stimulus modulation trials was 681.9 ms. Indeed, there is a significant difference in reaction time between high and low pre-stimulus MI trials ($p < 0.05$, permutation test). Moreover, using single-trial pre-stimulus MI to predict the perceptual behavior performance of trials with non-preferred stimulus condition of the electrode, we found no significant difference between the reaction time of high MI trials and low MI trials (high MI non-preferred trials mean $RT = 669.9$ ms, low MI non-preferred trials mean $RT = 669.6$ ms, n.s.).

6.3.4 Contribution of temporal and spectral features

We also evaluated the contribution of different kind of pre-stimulus features. Specifically, we compared the overall mean d' from the first step using only the post-stimulus features to the overall mean d' from the second step using (1) all pre-stimulus features, (2) only pre-stimulus phase features, (3) only pre-stimulus broadband features, (4) only pre-stimulus ERP features. As shown in Figure 6.3A, all 4 cases showed significant improvements in the classification accuracy

compared to the results of post-stimulus features only ($p < 10^{-5}$, paired t-test). This suggests that all kinds of features, including phase, ERP and broadband activity, are contributing to the pre-stimulus modulation of category tuning.

In addition, we are particularly interested in whether there are specific frequencies in the pre-stimulus activity that contribute to the modulation. The sparse model that we employed is often used for feature selection (Tibshirani, 1996). Therefore, by reviewing the nonzero features selected in model (6.3), we can probe the contribution of phases from different frequencies to the modulation of the post-stimulus category tuning. We computed the empirical frequency of having non-zero weight for each phase feature in model (6.3) across all electrodes and subjects. As shown in Figure 6.3, only phases of 15-30 Hz showed significant higher chance of being selected in the sparse GLM. The probability peaked at 15 Hz ($p < 0.05$, permutation test).

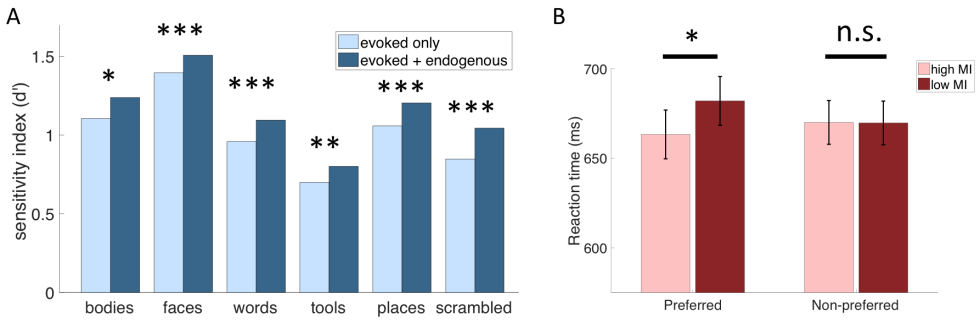


Figure 6.2: **Pre-stimulus activity contributes to category decoding and predicts reaction time. A)** Category classification accuracy before and after conditioning on pre-stimulus activity. (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, paired t-test) **B)** (Left) the averaged reaction time for low MI and high MI trials in the preferred condition of the electrode; (right) the averaged reaction time for low MI and high MI trials in the non-preferred conditions of the electrode. (error bar: standard error, *$p < 0.05$, permutation test)

6.3.5 Spatial and temporal specificity of the endogenous modulation effect

An emerging question is what is the nature of the pre-stimulus modulation effect. First we looked at the spatial property and ask whether it is a global or local effect. To evaluate the spatial specificity of the signal, we computed the trial-by-trial cross-electrode correlation in the pre-stimulus MI from category-selective electrodes in each subject. We found significant cross-electrode correlation in pre-stimulus MI between pairs of electrodes that share the same category-selectivity (mean absolute Spearman's $\rho = 0.111$, $p < 0.05$, permutation test), but not between pairs of electrodes that have different category-selectivity (mean absolute Spearman's $\rho = 0.0797$, $p > 0.05$, permutation test) (Figure 6.4A). In general, we observed larger correlation between electrodes of the same category-selectivity than electrodes of different category-selectivity ($p < 0.05$, MannWhitney U test) (Figure 6.4A).

The next question we ask about the pre-stimulus modulation is whether it is a transient or long-term effect. To evaluate the temporal specificity of the modulation effect, we computed

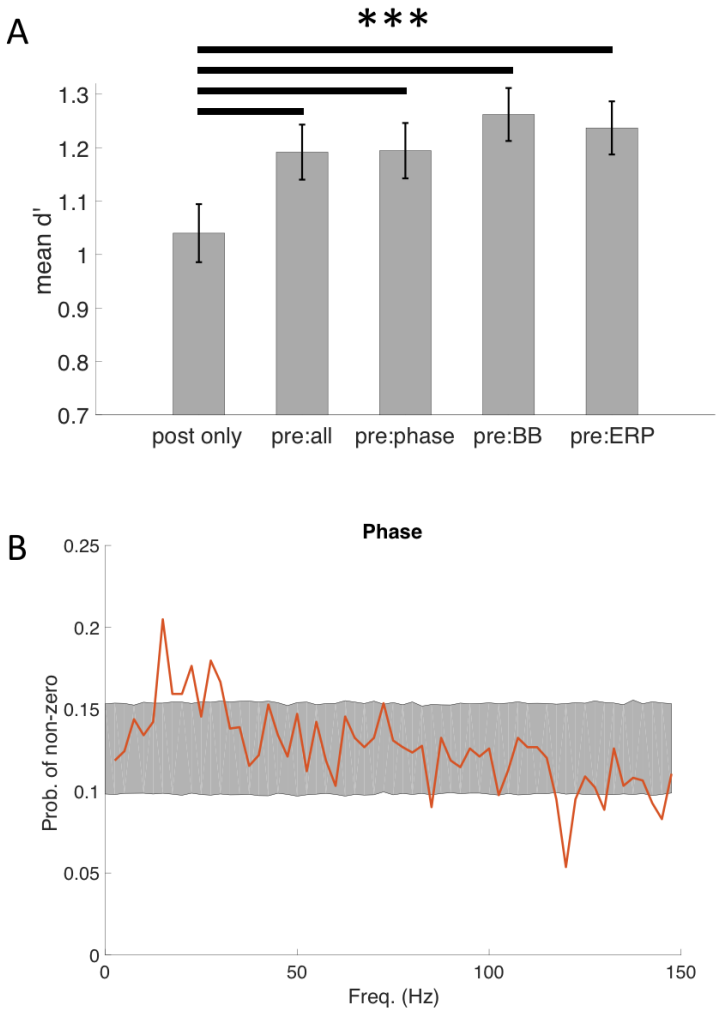


Figure 6.3: **Contributions of different pre-stimulus features in the model. A)** From left to right, the averaged classification d' across all electrodes for: (post:only) post-stimulus features only, (pre:all) all pre-stimulus features, (pre:phase) pre-stimulus phase features only, (pre:BB) pre-stimulus broadband features only, (pre-ERP) pre-stimulus ERP features only (*** $p < 10^{-5}$, paired t-test); **B)** The averaged empirical probability of having non-zero weights in the sparse GLM model for different pre-stimulus phase features of different frequency (shaded area: bootstrapped 95% confidence interval of being selected in the sparse GLM with random feature selection that has the same ℓ_0-norm as the current solutions).

the auto-correlation in the pre-stimulus modulation index in consecutive trials for each category-sensitive electrode. As shown in Figure 6.4B, we found that $\sim 15\%$ of the channels showed significant auto-correlation at $p < 0.05$ uncorrected level. Therefore while in the large majority of the channels the effect is transient, in a number of channels, where the number is higher than expected by chance, the pre-stimulus effect may be related to infra-slow fluctuations often seen in resting state studies in some cortical regions.

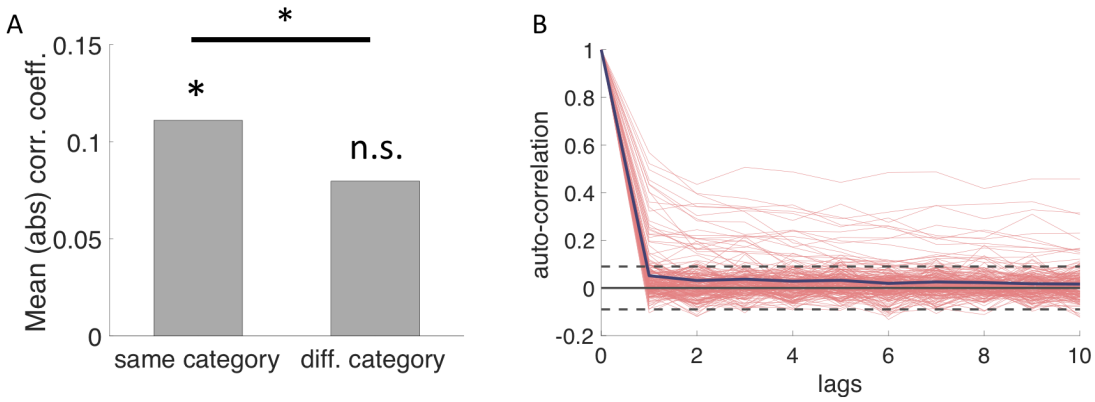


Figure 6.4: **The spatial and temporal specificity of the pre-stimulus modulation effect. A)** The mean absolute correlation coefficient (Spearman's rho) for cross-electrode correlation in the pre-stimulus MI between a pair of electrodes with the same category selectivity (left bar) versus a pair of electrodes with different category selectivity (right bar). (*$p < 0.05$, MannWhitney U test). **B)** The auto-correlation function for pre-stimulus MI across consecutive trials in each category-selective channel. The solid line indicates the average auto-correlation across all electrodes. The dotted lines correspond to $p = 0.05$ threshold, uncorrected.

6.4 Discussion

In this study, we analyzed the relationship between endogenous activity and category tuning, as well as behavioral visual perception, using iEEG data from a large cohort of 30 patients. We found that endogenous activity influences the degree of category tuning in response to visual stimuli. The same aspects in endogenous activity that influences post-stimulus category tuning also correlates with perceptual behavior performance. This modulation provides a potential neural basis for perceptual variation arising from shifts in endogenous ongoing activity. Furthermore, the endogenous modulation effect is a reflection of local processes within category-selective networks. The majority of the endogenous modulation effect are transient, but a small fraction of the category-selective channels show trial-by-trial auto-correlation in endogenous modulation of tuning.

6.4.1 The endogenous activity modulates category tuning

There have been a number of studies analyzing the relationship between the pre-stimulus activity and the post-stimulus evoked response. However, as we have pointed out, most of the previous studies have mainly focused on the overall unsupervised correlation between the endogenous activity and the evoked response in features including phase and oscillatory power of the event-related response (Becker et al., 2008; Fellinger et al., 2011; Rajagovindan and Ding, 2011) or blood oxygen-level dependent (BOLD) signal (Scheeringa et al., 2011). While these studies show that endogenous activity can affect the stimulus evoked response, they do not establish that it can modulate the quality of the neural representation for stimuli in ways that are related to perception. When we study perception, a central question is how different stimuli are represented

and encoded differentially in the corresponding brain areas (Averbeck et al., 2006; Haxby et al., 2001; Kriegeskorte et al., 2006). As a result, it is crucial to identify the relationship between endogenous activity and the critical discriminant neural representation/encoding that accounts for perception. In other words, it is important to know not only whether endogenous activity correlates to the post-stimulus activity overall, but also whether endogenous activity selectively modulates the post-stimulus activity along the critical dimensions that directly link to the tuning property of the category selective areas. A common way to study neural tuning in a category-selective area is to use a discriminant model to extract important features in the evoked response that discriminate the preferred category from the others (Norman et al., 2006). Therefore, what remains to be figured out here is whether endogenous activity modulate the post-stimulus evoked response along these critical discriminant features.

To evaluate the modulation effect along the critical discriminant dimensions, pre-stimulus activity was used as a proxy for the endogenous neural state of the brain when stimuli were presented. Specifically, changes in classification accuracy as a factor of conditioning on the pre-stimulus activity were examined. Because the pre-stimulus activity contains no information about the conditions the only way classification accuracy can improve using this model is if the pre-stimulus activity contains information about how sharply tuned the stimulus response along the critical dimension will be on a particular trial. The algorithm is designed to use this information, if it is present, to change the category boundary in the discriminant dimension on each trial to optimize classification accuracy. As we seen in our results, a significant improvement in the classification accuracy suggests that conditioning on pre-stimulus activity adds extra information to the category tuning model. In addition, as shown in the supplement results, the pre-stimulus activity predicts the distance to classification boundary on a trial by trial basis. This provides further mechanistic evidence about how pre-stimulus activity actually modulates post-stimulus category tuning. These result show that critical features of the pre-stimulus activity relate to the sharpness of the neural tuning and modifying the discriminant model based on this relationship improves classification accuracy. Therefore, these results support first hypothesis that pre-stimulus activity modulates the degree of category tuning in category-selective areas over the cortex.

6.4.2 Endogenous activity correlates to perceptual behavior

The sharpness of tuning is believed to reflect the quality of the neural representation, which in turn influences the quality of perception. To assess the relationship between the endogenous neural state and perception, we determined whether the critical aspects of the endogenous activity that modulate tuning corresponds to the reaction time on a simple perceptual task. Previous studies have established that endogenous neural activity correlates to perceptual behavior. A number of previous studies have shown connections between pre-stimulus endogenous activity and perceptual behavior across several sensory modalities, including vision, audition and somatosensory. Similar to the modulation effect, different aspects in the pre-stimulus features, including phase and amplitude of the event-related response/field (Bompas et al., 2015; Linkenkaer-Hansen et al., 2004; Mathewson et al., 2009), as well as blood oxygen-level dependent (BOLD) signal (Boly et al., 2007; Hesselmann et al., 2008; Sadaghiani et al., 2009; Schölvinck et al., 2012), are correlated to the variations in the perceptual behavior. However, the mechanism by which perception

is modulated by the endogenous activity is unknown because evidence that the same aspect of the endogenous activity that influences the quality of the neural representation also influences perception has not been established. A recent study based on scalp EEG and an auditory perception task (Kayser et al., 2016) suggested that the endogenous modulation of perception and the correlation to behavioral performance are found in different neural circuits. However, here we presented direct evidence that the two processes can be attributed to the same aspects of pre-stimulus endogenous activity in the same local category-sensitive circuit. Our results demonstrated a significant relationship between the pre-stimulus MI and the reaction time in detecting repetitions in the category that the electrode is sensitive to. Furthermore, no significant correlation was found between the pre-stimulus MI and the reaction time with respect to categories that the electrode is insensitive to. Effectively, the amount of influence on the post-stimulus discriminant dimension from the pre-stimulus endogenous activity predicted the reaction time in the one-back task with regard to stimuli in the preferred categories. Taken together, these results show that the same aspect of the endogenous activity that influences the trial-by-trial tuning in a region also correlates with the trial-by-trial response time on a perceptual task.

6.4.3 Concerns and possible confounding factor

One concern with regard to the behavior performance is that the task demands discrimination of individual images, while the category classifier requires discriminant information at category level. However, as shown in supplement results, the exemplar-level coding and category-level coding are often correlated in the category selective regions. Previous studies have also suggested that dynamic neural activity in the same category selective area contribute to both category-level encoding and exemplar-level encoding (Ghuman et al., 2014; Hirshorn et al., 2016; Li et al., 2018). As a result, although we used category classifier to define the pre-stimulus MI, it is reasonable to extend the usage of it to the exemplar case. Indeed, our results also confirmed this point that pre-stimulus MI derived from a category-level model can actually predict the task performance that requires individual exemplar level discriminant information.

Another possible confounding factor is the long-lasting broadband activity induced by the one-back task, which has been demonstrated in previous studies (Ghuman et al., 2014). This could become problematic when two consecutive trials shared the same category conditions but did not exactly repeat at the exemplar level. However, as shown in Supplement results, with category-level repetitions completely removed from the trials, similar modulation effects were still found when comparing the classification accuracy with and without conditioning on the pre-stimulus activity.

6.4.4 Spatial and temporal properties of the endogenous modulational signal

Our analysis on the pre-stimulus features showed that the pre-stimulus ERP, which is dominated by the low frequency component, the pre-stimulus ERBB, which reflects the power of high frequency broadband activity, and the pre-stimulus phases all contributed to the modulation of category tuning. Specifically, the alpha/beta phases, peaked at 15 Hz, showed a consistent

patterns of modulation on the post-stimulus category tuning.

One possible mechanism for this pre-stimulus modulation is that it is a reflection of fluctuations in the global cognitive state. For example, fluctuations in how much attention the patients' were paying on each trial or fluctuations in arousal. Indeed, previous studies have also tied variance in the endogenous activity to attention (Bauer et al., 2014; Kastner et al., 1999; Worden et al., 2000), and showed that attention can modulate neural tuning (Luck et al., 1997; Saproo and Serences, 2010). If it is true that endogenous modulation is a global effect, significant cross-electrode correlation in the pre-stimulus MI components should be expected in each subject, regardless of category-selectivity of the electrodes. However, as the cross-electrode correlation analysis revealed, cross-electrode correlation in pre-stimulus MI is only significant between electrodes that share the same category-selectivity, but not for electrodes of different category-selectivity. It would require some local fluctuation at circuit/cellular level to facilitate such an selective effect (Lee et al., 2018). The differentiation in the cross-electrode correlations could be a result of stronger anatomical connection between regions in the same cortical network that show similar category-selectivity, and the local fluctuation is propagated heterogeneously due to higher level of anatomical connection between these regions (Pyles et al., 2013; Saygin et al., 2012). As a result, the endogenous modulation is partially a reflection of fluctuations within category-specific networks (the effect size is weak), but it does not seem to be a reflection of the global cognitive state, such as attention or arousal.

The other possibility mechanism is that this is a reflection of infra-slow fluctuations, which has been described in resting state studies (Becker et al., 2011; Fox et al., 2006; Henriksson et al., 2015). If it is a reflection of intra-slow fluctuations as seen in the resting-state activity, we should expect significant auto-correlation within each channel between consecutive trials. However, as demonstrated in the auto-correlation analysis, only in a small fraction of the electrodes, where the number is higher than expected by chance, the pre-stimulus effect demonstrated long-range temporal correlation which is often seen in resting state studies (Linkenkaer-Hansen et al., 2001; Smit et al., 2013), while in the large majority of the channels the effect is transient. The endogenous activity that influences neural tuning and perception is primarily a reflection of transient fluctuations, though a significant subset of the effect does reflect infra-slow fluctuations, such as those previously reported in resting state

Prior studies have shown that endogenous activity can influence task related neural activity and perception. However, a link between the endogenous activity and the quality of the neural representation that could provide a neural mechanism by which endogenous activity can influence perception has been lacking. The results here show that the phase in the alpha and beta frequency bands (12-30 Hz) influences how sharply tuned a region will be to sensory input on a trial-by-trial basis. Furthermore, the same aspect of the endogenous activity that influences tuning is also correlated to perception, providing an empirical evidence for a mechanistic link between endogenous activity, neural tuning, and perception. This aspect of the endogenous activity was not a reflection of global neural state, such as global arousal or attention, but rather is reflection of a mix of transient and infra-slow local fluctuations in endogenous activity. This is suggestive of local neural fluctuations of endogenous processes, such a local fluctuations of neurotransmitter levels (Lee et al., 2018) or fluctuations in stimulus-specific attention or preference (Kastner et al., 1999). Future studies will be required to determine the precise nature of the aspect of the endogenous activity that influences neural tuning. Taken together, these results provide

empirical support for a mechanism in which the present neural state influences the perception of sensory input by modulating the tuning properties of local neural populations.

6.5 Appendix: Supplement methods and results

6.5.1 Solving the two-stage GLM using coordinate descent

As shown in (6.3), the overall optimization problem can be written as

$$\underset{\beta_{evk},\beta_{pre}}{\text{argmin}} \quad -\ell(\beta_{evk},\beta_{pre}) + \lambda_1 P_\alpha^{evk}(\beta_{evk}) + \lambda_2 P_\alpha^{pre}(\beta_{pre}) \tag{6.11}$$

where

$$\ell(\beta_{evk},\beta_{pre}) = \frac{1}{N}\{y^T(X_{evk}\beta_{evk} - X_{pre}\beta_{pre}) - \mathbf{1}^T\log(1+\exp(X_{evk}\beta_{evk} - X_{pre}\beta_{pre}))\} \tag{6.12}$$

Solve the elastic-net problem in the first step

In the first step, we set $\beta_{pre} = 0$, and solve the following elastic-net problem:

$$\underset{\beta_{evk}}{\text{argmin}} \quad -\ell(\beta_{evk}) + \lambda_1 P_\alpha^{evk}(\beta_{evk}) \tag{6.13}$$

where β_{evk} is a vector that contains the intercept β_0^{evk} and the feature weights β^{evk}

$$\begin{aligned} -\ell(\beta_{evk}) &= -\frac{1}{N}\{y^T(X_{evk}\beta_{evk}) - \mathbf{1}^T\log(1+\exp(X_{evk}\beta_{evk}))\} \\ &= -\frac{1}{N}\sum_{i=1}^{N}\{y_i(\beta_0^{evk} + x_i^T\beta^{evk}) - \log[1+\exp(\beta_0^{evk} + x_i^T\beta^{evk})]\} \end{aligned} \tag{6.14}$$

and

$$\begin{aligned} P_\alpha^{evk}(\beta_{evk}) &= \frac{1-\alpha}{2}\|\beta_{evk}\|_2^2 + \alpha\|\beta_{evk}\|_1 \\ &= \sum_{i=1}^{P}[\frac{1}{2}(1-\alpha)\|\beta_j^{evk}\|_2^2 + \alpha\|\beta_j^{evk}\|_1] \end{aligned} \tag{6.15}$$

For simplicity of the notation, we are omitting the superscript 'evk' in the following part, and we assume that X has been standardized such that each dimension $x_{:j}$ has 0 mean and unit variance.

During the optimizing iterations, assume that the current solution is $[\tilde{\beta}, \tilde{\beta}_0]$, we are solving the updated solution $[\beta_0, \beta]$ following problem:

$$\underset{\beta,\beta_0}{\text{argmin}} \quad -\ell_{\tilde{\beta},\tilde{\beta}_0}(\beta,\beta_0) + \lambda_1 P_\alpha(\beta) \tag{6.16}$$

we can use quadratic approximation around $[\tilde{\beta}, \tilde{\beta}_0]$ for the negative log likelihood term in (6.16)

$$-\ell_{\tilde{\beta},\tilde{\beta}_0}(\beta, \beta_0) = -\ell(\tilde{\beta}_0, \tilde{\beta}) - \nabla\ell(\tilde{\beta}_0, \tilde{\beta})^T \Delta\beta - \frac{1}{2}\Delta\beta^T H(\tilde{\beta}_0, \tilde{\beta})\Delta\beta + R(\|\Delta\beta\|^2) \tag{6.17}$$

$$\approx -\frac{1}{2}\sum_{i=1}^{N} w_i(z_i - \beta_0 - x_i^T\beta)^2 + C(\tilde{\beta}_0, \tilde{\beta}) \tag{6.18}$$

Note that the last term does not depend on $[\beta_0, \beta]$, and we have gradient and Hessian at $[\tilde{\beta}, \tilde{\beta}_0]$ as

$$\nabla\ell(\beta) = \frac{1}{N}\sum_{i=1}^{N}[y_i - P_{\tilde{\beta}_0,\tilde{\beta}}(x_i)]x_i \tag{6.19}$$

$$H(\tilde{\beta}_0, \tilde{\beta}) = -\frac{1}{N}\sum_{i=1}^{N} P_{\tilde{\beta}_0,\tilde{\beta}}(x_i)(1 - P_{\tilde{\beta}_0,\tilde{\beta}}(x_i))x_i x_i^T \tag{6.20}$$

where $\Delta\beta = [\beta_0 - \tilde{\beta}_0, \beta^T - \tilde{\beta}^T]^T$ is the update difference, and $P_{\tilde{\beta}_0,\tilde{\beta}}(x_i) = \frac{1}{1+\exp(-\tilde{\beta}_0 - x_i^T\tilde{\beta})}$ is the estimated likelihood at $[\tilde{\beta}, \tilde{\beta}_0]$.

Plugging (6.19),(6.20) into (6.17) and the comparing with (6.18), we get

$$z_i = \tilde{\beta}_0 + x_i^T\tilde{\beta} + \frac{y_i - P_{\tilde{\beta}_0,\tilde{\beta}}(x_i)}{P_{\tilde{\beta}_0,\tilde{\beta}}(x_i)(1 - P_{\tilde{\beta}_0,\tilde{\beta}}(x_i))} \tag{6.21}$$

$$w_i = \frac{1}{N}P_{\tilde{\beta}_0,\tilde{\beta}}(x_i)(1 - P_{\tilde{\beta}_0,\tilde{\beta}}(x_i)) \tag{6.22}$$

As a result, solving (6.16) becomes solving the following regularized weighted least-squares problem:

$$\underset{\beta,\beta_0}{\text{argmin}} \quad -\frac{1}{2}\sum_{i=1}^{N} w_i(z_i - \beta_0 - x_i^T\beta)^2 + \lambda_1\sum_{i=1}^{T_2}[\frac{1}{2}(1-\alpha)\|\beta_j\|_2^2 + \alpha\|\beta_j\|_1] \tag{6.23}$$

We use coordinate descent to solve (6.23). Taking subgradient and set it to 0, through some calculus we get coordinate-wise update

$$\tilde{\beta}_j \leftarrow \frac{S(\sum_{i=1}^{N} w_i x_{ij}(z_i - \tilde{z}_i^{(-j)}), \lambda\alpha)}{\sum_{i=1}^{N} w_i x_{ij}^2 + \lambda(1-\alpha)} \tag{6.24}$$

where $\tilde{z}_i^{(-j)} = \tilde{\beta}_0 + \sum_{k\neq j} x_{ik}\tilde{\beta}_k$ is the fitted value excluding the contribution from x_{ij}, and $S(z, \gamma) = \text{sign}(z)(|z| - \gamma)_+$ is the soft-thresholding operator, where

$$S(z, \gamma) = \text{sign}(z)(|z| - \gamma)_+ = \begin{cases} z - \gamma & \text{if } z > 0 \text{ and } \gamma < |z| \\ z + \gamma & \text{if } z < 0 \text{ and } \gamma < |z| \\ 0 & \text{if } \gamma \geq |z| \end{cases} \tag{6.25}$$

To sum up, in the first step, we solve the elastic-net regularized GLM using coordinate descent, as shown in Algorithm 3.

Algorithm 3: Solve the elastic-net regularized GLM using coordinate descent

Data: data matrix $X_{evk} \in \mathbb{R}^{N \times T_2}$ for post-stimulus part of the data, data label $y \in \mathbb{R}^N$; where N is the number of samples, $T_2 = t_{evk}^{ERP} + t_{evk}^{ERBB}$, and $X_{evk} = [X_{evk}^{ERP}, X_{evk}^{ERBB}]$;
Parameters: the elastic-net hyper-parameter α, maximum regularization parameter λ_{max} and minimum regularization parameter $\epsilon\lambda_{max}$.
Result: Weight vectors for post-stimulus features $\beta_{evk}^* = [\beta_0^{evk}, \beta_{evk}^{ERP}, \beta_{evk}^{ERBB}]$

```
Fit the elastic-net problem for post-stimulus features:
for the i-th cross-validation split {X^(i)_{evk,train}, X^(i)_{evk,test}} do
    for λ ← λ_max to ελ_max (decrement λ) do
        while not converge do
            update the current quadratic approximation (6.23)by computing (6.21),(6.22);
            for j ← 1 to T_2 (cyclic coordinate descent) do
                update the weight of each coordinate β̃_j using (6.24);
        estimate the deviance of the solution for current λ on X^(i)_{evk,test};
find optimal λ* and the corresponding β*_{evk} that minimizes deviance.
```

Solve the group elastic-net GLM problem in the second step

The second step of fitting the two-stage GLM requires fixing the contribution from post-stimulus features and optimize the model with group elastic-net penalty on the pre-stimulus features. By fixing the weights from post-stimulus features, for each sample x_i, we get a fixed offset $b_i = \beta_0^{evk} + x_i^T \beta_{evk}^*$. Therefore, as in (6.7), we have

$$\beta_{pre}^* = \underset{\beta_{pre}}{\operatorname{argmin}} \quad -\ell(\beta_{evk}^*, \beta_{pre}) + \lambda_2 P_\alpha^{pre}(\beta_{pre}) \tag{6.26}$$

where

$$\begin{aligned} -\ell(\beta_{evk}^*, \beta_{pre}) &= -\frac{1}{N}\{y^T(b - X_{pre}\beta_{pre}) - \mathbf{1}^T \log(1 + \exp(b - X_{pre}\beta_{pre}))\} \\ &= -\frac{1}{N}\sum_{i=1}^{N}\{y_i(b_i - \beta_0^{pre} - (x_i^{pre})^T\beta^{pre}) - \log(1 + \exp(b_i - \beta_0^{pre} - (x_i^{pre})^T\beta^{pre}))\} \end{aligned} \tag{6.27}$$

and

$$\begin{aligned} P_\alpha^{pre}(\beta_{pre}) &= \frac{1-\alpha}{2}\|\beta_{pre}\|_2^2 + \alpha\|\beta_{pre}^{ERP}\|_1 + \alpha\|\beta_{pre}^{ERBB}\|_1 + \alpha\mathcal{G}(\beta_{pre}^{phase}) \\ &= \frac{1-\alpha}{2}\|\beta_{pre}\|_2^2 + \alpha\sum_{g=1}^{G}\sqrt{p_g}\|\beta_{pre}^{(g)}\|_2 \end{aligned} \tag{6.28}$$

where the second term is the group-lasso penalty on the pre-stimulus features. Similarly to the previous part, from now on we omit the 'pre' in superscript and subscript for simplicity.

Similar to previous part, we first take the quadratic approximation of the negative log likelihood at the current iteration step around $[\tilde{\beta}_0, \tilde{\beta}]$ as

$$-\ell_{\tilde{\beta},\tilde{\beta}_0}(\beta,\beta_0) = -\ell(\tilde{\beta}_0,\tilde{\beta}) - \nabla\ell(\tilde{\beta}_0,\tilde{\beta})^T\Delta\beta - \frac{1}{2}\Delta\beta^T H(\tilde{\beta}_0,\tilde{\beta})\Delta\beta + R(\|\Delta\beta\|^2) \tag{6.29}$$

$$\approx -\frac{1}{2}\sum_{i=1}^{N} w_i(z_i - \beta_0 - x_i^T\beta)^2 + C(\tilde{\beta}_0,\tilde{\beta})$$

$$= -\frac{1}{2}(z - \sum_{g=1}^{G} X^{(g)}\beta^{(g)})^T W(z - \sum_{g=1}^{G} X^{(g)}\beta^{(g)}) + C(\tilde{\beta}_0,\tilde{\beta}) \tag{6.30}$$

where $z = [z_1, ..., z_N]^T$, $W = diag\{w_1, ..., w_N\}$, and $X = [X^{(1)}, ..., X^{(G)}]$ is the blocks in X that corresponding to each group $\beta^{(g)}$ and

$$z_i = b_i + \tilde{\beta}_0 + x_i^T\tilde{\beta} + \frac{y_i - P_{\tilde{\beta}_0,\tilde{\beta}}(x_i)}{P_{\tilde{\beta}_0,\tilde{\beta}}(x_i)(1 - P_{\tilde{\beta}_0,\tilde{\beta}}(x_i))} \tag{6.31}$$

$$w_i = \frac{1}{N}P_{\tilde{\beta}_0,\tilde{\beta}}(x_i)(1 - P_{\tilde{\beta}_0,\tilde{\beta}}(x_i)) \tag{6.32}$$

and

$$P_{\tilde{\beta}_0,\tilde{\beta}}(x_i) = \frac{1}{1 + \exp(-b_i - \tilde{\beta}_0 - x_i^T\tilde{\beta})} \tag{6.33}$$

Analogously, solving (6.26) becomes iteratively solving the following regularized weighted least-squares problem:

$$\underset{\beta,\beta_0}{\text{argmin}} \quad -\frac{1}{2}(z - \sum_{g=1}^{G} X^{(g)}\beta^{(g)})^T W(z - \sum_{g=1}^{G} X^{(g)}\beta^{(g)}) + \lambda_2\frac{1-\alpha}{2}\|\beta\|_2^2 + \lambda_2\alpha\sum_{g=1}^{G}\sqrt{p_g}\|\beta^{(g)}\|_2 \tag{6.34}$$

Let $r^{(-g)} = z - \sum_{j\neq g} X^{(j)}\beta^{(j)}$ be the residual excluding the contribution of $\beta^{(g)}$. The first-order optimality condition gives

$$(X^{(g)})^T W r^{(-g)} + \big[\lambda_2(1-\alpha)I^{(g)} - (X^{(g)})^T W X^{(g)}\big]\beta^{(g)} + \lambda_2\alpha\sqrt{p_g}\nu^{(g)} = 0 \tag{6.35}$$

where subgradient

$$\nu^{(g)} \in \begin{cases} \{\frac{\beta^{(g)}}{\|\beta^{(g)}\|_2}\} & \text{if } \beta^{(g)} \neq 0 \\ \{u \mid \|u\| \leq 1\} & \text{if } \beta^{(g)} = 0 \end{cases} \tag{6.36}$$

The optimal solution for each group is given as

$$\tilde{\beta}^{(g)} = \begin{cases} \Big((X^{(g)})^T W X^{(g)} + \lambda_2\big[(1-\alpha) + \frac{\alpha\sqrt{p_g}}{\|\tilde{\beta}^{(g)}\|}\big]I^{(g)}\Big)^{-1}(X^{(g)})^T W r^{(-g)} & \text{if } \|(X^{(g)})^T W r^{(-g)}\|_2 > \lambda_2\alpha\sqrt{p_g} \\ 0 & \text{if } \|(X^{(g)})^T W r^{(-g)}\|_2 \leq \lambda_2\alpha\sqrt{p_g} \end{cases} \tag{6.37}$$

An assumption that is often made in group lasso problems is the within-group orthonormality, where $(X^{(g)})^T X^{(g)} = I$, so that (6.37) has closed form solution (Hastie et al., 2015; Simon and Tibshirani, 2012). For our case, this orthonormality does not necessarily hold. Therefore we solve for the general case. Let $Q^{(g)} = W^{1/2}X^{(g)} = (diag\{\sqrt{w_1}, ..., \sqrt{w_N}\})X^{(g)}$, and $\mu^{(-g)} = W^{1/2}r^{-g}$, then we rewrite (6.37) as

$$\tilde{\beta}^{(g)} = \begin{cases} \Big((Q^{(g)})^T Q^{(g)} + \lambda_2\big[(1-\alpha) + \frac{\alpha\sqrt{p_g}}{\|\tilde{\beta}^{(g)}\|}\big] I^{(g)}\Big)^{-1} (Q^{(g)})^T \mu^{(-g)} & \text{if } \|(Q^{(g)})^T\mu^{(-g)}\|_2 > \lambda_2\alpha\sqrt{p_g} \\ 0 & \text{if } \|(Q^{(g)})^T\mu^{(-g)}\|_2 \leq \lambda_2\alpha\sqrt{p_g} \end{cases} \tag{6.38}$$

For the case of $\|(Q^{(g)})^T\mu^{(-g)}\|_2 \leq \lambda_2\alpha\sqrt{p_g}$, we have explicit solution that $\tilde{\beta}^{(g)} = 0$. Therefore we focus on the case of $\|(Q^{(g)})^T\mu^{(-g)}\|_2 > \lambda_2\alpha\sqrt{p_g}$. Notice that if we know the ℓ_2-norm $\|\beta^{(g)}\|$, then (6.38) becomes closed form solution. Therefore, we first find the norm $\|\beta^{(g)}\|$. We take the singular value decomposition $Q^{(g)} = U^{(g)}D^{(g)}(V^{(g)})^T$, where $U^{(g)}$ and $V^{(g)}$ has orthonormal columns, and $D^{(g)} = diag\{d_1^{(g)}, ..., d_{p_g}^{(g)}\}$ is diagonal matrix. Let $\eta^{(-g)} = [\eta_1^{(-g)}, ..., \eta_{p_g}^{(-g)}]^T = (U^{(g)})^T\mu^{(-g)}$. For simplicity of notations, we are omitting superscript '(g)' in U,D and V, but keep in mind that we are solving for an individual group g. As a result, for (6.38), we have

$$\tilde{\beta}^{(g)} = \Big((Q^{(g)})^T Q^{(g)} + \lambda_2\big[(1-\alpha) + \frac{\alpha\sqrt{p_g}}{\|\tilde{\beta}^{(g)}\|}\big] I^{(g)}\Big)^{-1} (Q^{(g)})^T \mu^{(-g)} \tag{6.39}$$

$$\iff \quad \tilde{\beta}^{(g)} = \Big(VD^2V^T + \lambda_2 V\big[(1-\alpha) + \frac{\alpha\sqrt{p_g}}{\|\tilde{\beta}^{(g)}\|}\big] I^{(g)}V^T\Big)^{-1} VDU^T\mu^{(-g)} \tag{6.40}$$

$$\iff \quad V^T\tilde{\beta}^{(g)} = \Big(D^2 + \lambda_2\big((1-\alpha) + \frac{\alpha\sqrt{p_g}}{\|\tilde{\beta}^{(g)}\|}\big) I\Big)^{-1} D\eta^{(-g)} \tag{6.41}$$

Note that LHS and RHS of (6.41) are two vectors, and take ℓ_2-norm on both sides, we get

$$\|\tilde{\beta}^{(g)}\|_2^2 = \sum_{i=1}^{p_g} \frac{d_i^2\eta_i^2}{\big(d_i^2 + \lambda_2(1-\alpha) + \lambda_2\frac{\alpha\sqrt{p_g}}{\|\tilde{\beta}^{(g)}\|}\big)^2} \tag{6.42}$$

$$\iff \quad \sum_{i=1}^{p_g} \frac{d_i^2\eta_i^2}{\big([d_i^2 + \lambda_2(1-\alpha)]\|\tilde{\beta}^{(g)}\|_2 + \lambda_2\alpha\sqrt{p_g}\big)^2} = 1 \tag{6.43}$$

Therefore, the ℓ_2-norm of the optimal solution, $\|\tilde{\beta}^{(g)}\|_2$, is the solution to equation $f(\gamma) = 0$, which is

$$f(\gamma) = \sum_{i=1}^{p_g} \frac{d_i^2\eta_i^2}{\big([d_i^2 + \lambda_2(1-\alpha)]\gamma + \lambda_2\alpha\sqrt{p_g}\big)^2} - 1 = 0 \tag{6.44}$$

It is easy to check that f is convex, $f(0) > 0$, and f is monotonically decreasing as γ increases, and $\lim_{\gamma\to\infty} f(\gamma) = -1$. Therefore, $\|\tilde{\beta}^{(g)}\|_2$ is the only solution to $f(\gamma) = 0$. And $\|\tilde{\beta}^{(g)}\|_2$ can be efficiently found by Newton's method.

Once we find the optimal $\|\tilde{\beta}^{(g)}\|_2$, (6.38) becomes the closed form solution, and we can solve (6.26) using block coordinate descent. The algorithm can be summarized as in Algorithm 4.

Note that Algorithms 3 and 4 are effectively second-order methods with a fixed step size of 1. It is also possible to adapt it into a backtrack line search to find an optimal step length adaptively.

Algorithm 4: Solve the group elastic-net regularized GLM using coordinate descent

Data: data matrix $X_{pre} \in \mathbb{R}^{N \times T_1}$ for pre-stimulus part of the data, and the corresponding group partition into G groups, with group length $\{p_1, ..., p_G\}$, data label $y \in \mathbb{R}^N$ and fixed post-stimulus solution β^*_{post}; where N is the number of samples, $T_1 = t^{ERP}_{pre} + t^{ERBB}_{pre} + t^{phase}_{pre}$, and $X_{pre} = [X^{ERP}_{pre}, X^{ERBB}_{pre}, X^{phase}_{pre}]$;

Parameters: the elastic-net hyper-parameter α, maximum regularization parameter λ_{max} and minimum regularization parameter $\epsilon\lambda_{max}$.

Result: Weight vectors for post-stimulus features $\beta^*_{pre} = [\beta^{pre}_0, \beta^{ERP}_{pre}, \beta^{ERBB}_{pre}, \beta^{phase}_{pre}]$

Fit the group elastic-net problem for pre-stimulus features:
for *the i-th cross-validation split* $\{X^{(i)}_{pre,train}, X^{(i)}_{pre,test}\}$ **do**
  for $\lambda \leftarrow \lambda_{max}$ **to** $\epsilon\lambda_{max}$ *(decrement* λ*)* **do**
    while *not converge* **do**
      update the current quadratic approximation (6.34) by computing (6.31),(6.32);
      for $j \leftarrow 1$ **to** G *(cyclic block coordinate descent)* **do**
        use Newton's method to solve for the norm $\|\tilde{\beta}^{(g)}\|$ in equation (6.44);
        update the weight of each coordinate group $\tilde{\beta}^{(g)}$ using $\|\tilde{\beta}^{(g)}\|$ and (6.38) ;
    estimate the deviance of the solution for current λ on $X^{(i)}_{pre,test}$;
find optimal λ^* and the corresponding β^*_{pre} that minimizes deviance.

6.5.2 Category-level decoding and exemplar-level decoding

One concern with regard to the behavior performance is that the task demands discrimination of individual images, while the category classifier requires discriminant information at category level. However, as shown in supplement results, the exemplar-level coding and category-level coding are often correlated in the category selective regions. Previous studies have also suggested that dynamic neural activity in the same category selective area contribute to both category-level encoding and exemplar-level encoding. Specifically, a significant positive correlation between the decoding accuracy (d') for face category and the decoding accuracy (d') for facial expressions was seen (Pearson correlation $r = 0.57$, $N = 21$, $P = 0.007$) (Li et al., 2018); a positive correlation between the decoding accuracy for face category and the decoding accuracy for facial identity was seen in another case ($r = 0.47$, $N = 13$, $P = 0.10$) (Ghuman et al., 2014; Li et al., 2018). As a result, although we used category classifier to define the pre-stimulus MI, it is reasonable to extend the usage of it to the exemplar case. Indeed, our results also confirmed this point that pre-stimulus MI derived from a category-level model can actually predict the task performance that requires individual exemplar level discriminant information.

6.5.3 Classification results excluding all categorically repeated trials

Another possible confounding factor is the long-lasting broadband activity induced by the one-back task, which has been demonstrated in previous studies(Ghuman et al., 2014). This could become problematic when two consecutive trials shared the same category conditions but did

not exactly repeat at the exemplar level. However, as shown in Table 6.2, with category-level repetitions completely removed from the trials, similar modulation effects were still found when comparing the classification accuracy with and without conditioning on the pre-stimulus activity.

Table 6.2: The comparisons of classification results from the two-stage GLM when excluding all repeated trials with the same category as the 1-back trial

Category	Bodies	Faces	Words	Tools	Houses	Scrambled non-objects
# of electrodes	9	56	92	16	37	36
d' (evoked only)	1.1018	1.5301	1.0847	0.7881	1.0594	0.8677
d' (evoked + endogenous)	1.1936	1.6091	1.1904	0.8990	1.1948	1.0651
p-value	0.0908	1.6×10^{-5}	$< 10^{-5}$	9.7×10^{-4}	$< 10^{-5}$	$< 10^{-5}$

6.5.4 Predicting distance to post-stimulus decision boundary

In addition to the two-stage GLM presented in the main text, a linear regression model was directly applied to evaluate the relationship between pre-stimulus activity and the absolute distance to the decision boundary in the post-stimulus discriminant model. Specifically, we solved the following linear regression problem:

$$|X_{evk}\beta_{evk}| = X_{pre}\beta_{pre}$$

Similar to the main results presented in Figure 6.2 and Table 6.1, we found significant correlation between pre-stimulus activity and absolute distance to the decision boundary in all categories (Table 6.3).

Table 6.3: The R^2 of the linear regression model between pre-stimulus activity and the absolute distance to the decision boundary in the post-stimulus discriminant model. (p-value estimated using the Fisher Z-transformation).

Category	Bodies	Faces	Words	Tools	Houses	Scrambled non-objects
# of electrodes	9	56	92	16	37	36
R^2	0.0717	0.0507	0.0377	0.0275	0.0361	0.0221
p-value	0.0327	$< 10^{-5}$	2.8×10^{-4}	0.0150	0.0017	0.0678

Chapter 7

Conclusion and future directions

7.1 Main conclusion

In this thesis, we elaborate around the multivariate representational space in population neural activity and explore along different methodological dimensions in order to address gaps in the prevalent hierarchical model of the visual cortex, and attain a comprehensive understanding the spatiotemporal dynamics and interactions underlying visual perception.

The first part of the thesis (Chapters 2,3,4) mainly focuses on multivariate analysis of the representation dynamics in local areas. In Chapters 2 and 3 We extend the classical multivariate functional mapping framework and applied it to analyze the spatiotemporal dynamics in the category-selective patches in fusiform. These areas are critical for visual recognition with damage to these patches leading to category-selective impairments in object recognition, such as acquired alexia and prosopagnosia. However, many gaps remain in our understanding of the dynamic role the fusiform plays in contributing to multiple stages of category-specific information processing. The results show strong decoding accuracy for faces and words in the FFA and VWFA respectively, first becoming statistically significant between 50-100 ms and peaking between 150-200 ms. Next we examined the dynamics of within category decoding. For words significant decoding was seen in both subjects between approximately 100-250 ms wherein visually similar words could not be decoded from one another, but dissimilar words could be decoded (organized by orthographic similarity). There was a later phase between approximately 250-500 ms where even orthographically similar words could be significantly decoded from one another (individual-level representation). For faces significant expression-invariant decoding was seen in each subject in the same 250-500 ms time frame. The neural response for faces was organized by facial feature similarity, emphasizing the role of the eyes and mouth in face individuation. In addition, results in Chapter 4 show that expression sensitivity display a spatiotemporal division between early and late processing. Specifically, facial expressions could be decoded from the fusiform. Significant expression decoding was seen in the 100-250 ms in posterior face sensitive fusiform and from 250-500 ms in mid-fusiform. Taken together, these results suggest a multistage information processing dynamic wherein the representation in category-selective fusiform gyrus evolves from a coarse category-level representation to an invariant and highly detailed individual representation over the course of 500 ms.

In addition to analyzing local representation dynamics, in the second part of the thesis (Chapter 5) we turn our attention to the representation through population interactions in neural circuits and we introduce a novel analysis method: MCPA. Previously, multivariate pattern analysis methods have been used to analyze the sensitivity to information within a certain area and functional connectivity methods have been used to assess whether or not brain networks participate in a particular process. With MCPA, the two perspectives are merged into one algorithm, which extends multivariate pattern analysis to enable the detailed examination of information sensitivity at the network level. Thus, the introduction of MCPA provides a platform for examining how computation is carried out through the interactions between different brain areas, allowing us to directly test hypotheses regarding circuit-level information processing. As proof of concept, we show examples of the applications of MCPA to different neural data to probe the representations through neural interactions. We showed that the interactions between areas in the visual hierarchy encode both categorical and exemplar level information about the stimuli, and that within area (MVPA) and between area representations (MCPA) encode complimentary information.

The third part of the thesis (Chapter 6), we move along a different dimension of the methodological space, and study the relationship between visual category tuning and the pre-stimulus ongoing brain activity. We found that pre-stimulus activity influences the degree of category tuning in response to visual stimuli in category-selective cortical regions. The same aspects in pre-stimulus activity that influences post-stimulus category tuning also correlates with perceptual behavior performance. In addition, the pre-stimulus modulation effect is a reflection of local processes within the network of regions with the same category-selectivity. The majority of the pre-stimulus modulation effect are transient, but $\sim 15\%$ of the channels show trial-by-trial auto-correlation in endogenous modulation of tuning.

To sum up, we study the extended information processing dynamics, the interactive and adaptive information communication, as well as the state-dependence of evoked response in the brain network underlying visual perception. By addressing these gaps in the classical hierarchical model of the visual system, we establish a dynamic model of the visual perception network, where the early stage (50-150 ms) is dominated by feedforward sweep for coarse visual representation and the later stage (200-500 ms) reflects recurrent refinement for detailed visual representation, and the local activity is influenced by ongoing neural state.

7.2 Limitations, Challenges and Future directions

Toward optimal representational basis: from local dynamics to full-brain analysis

With the advances in recording techniques, such as the application of high-density grids, it becomes possible to record population dynamics with larger and denser coverage. This allows us to study the representation dynamics at larger scale. The framework that we adopt in this thesis is mainly ROI-oriented. In principle, our multivariate representation framework can be directly applied to larger scale neural recordings. However, more care should be taken as we move towards higher-dimensional settings. The high spatial and temporal correlation structure in the iEEG data makes many of the classical high-dimensional techniques undesirable, such as lasso and other ℓ_1-based regularization methods. To address this problem, future statistical models

should be able to find the appropriate representational basis from the high-dimensional iEEG data. In addition, as shown in this thesis, the different aspects of the recorded neural activity from iEEG, including the ERP, the broadband activity, and phases, all contributing to different tasks and analysis. Therefore, it suggests that instead of arbitrarily defined features like ERP, broadband, and phases, we may be able to use a data-driven way to design statistical model that picks the optimal representation basis for specific types of analysis.

The other challenge that would emerge when shifting to full-brain scale is to combine representations from multiple subjects. Because of the heterogeneity in the electrode coverage, it is not as straight-forward as the ROI-based approach to directly combine results from multiple subjects with different electrode coverage. How to find a common latent space from the subjects is another direction that needs further investigation.

From a methodological point of view, these two points above both requires dimensionality reduction and representation learning. Recent years have witnessed great advances in deep models (Goodfellow et al., 2016), and the application of supervised models with deep neural nets have been successful in directly decoding neural signals (Bashivan et al., 2015; Lawhern et al., 2016; Manor and Geva, 2015; Schirrmeister et al., 2017; Stober et al., 2014; Wang et al., 2013), as well as representational similarity analysis between the model feature spaces and the neural spaces at different brain regions (Cadieu et al., 2014; Yamins et al., 2013; Yamins and DiCarlo, 2016; Yamins et al., 2014). However, few works have explored unsupervised models that extract and characterize the intrinsic representation space of the spatiotemporal dynamics in the intracranial EEG signal. Therefore, another future direction is to design deep generative model for nonlinear dimensionality reduction to discover the low-dimensional nonlinear manifold that characterizes the spatiotemporal dynamics of the neural activity in ventral visual stream. This low-dimensional structure can be used to identify the "codebook" used by the neural circuits for encoding visual information in the task-evoked response.

Toward dynamic causal analysis

In Chapter 2, we disrupt the neural activity, using electrical stimulation, to build a causal link between local neural activity and cognitive functions. Due to the limitation of clinical mapping session, at this point we do not have the temporal resolution to precisely stimulate the region at different stages. Therefore, we only focus on building a overall causal link between the region and cognitive function. However, to get a precise understanding of the temporal dynamics and delineate the dynamic role of each area at different stages would require precise control of the stimulation timing and duration. Future works are necessary to develop and apply more precise stimulation paradigms to perform such dynamic causal analysis for individual ROIs and even multiple ROIs.

Combining local dynamics and network interactions

The approaches that we take in this thesis isolate the analysis of local dynamics and the analysis of neural interactions from each other, and we demonstrate that the two can be independent or complementary from each other. An interesting extension would be a network model that combine both feature representation within each node and the interactions between nodes. Ap-

proaches from graph neural networks could be potentially adopted toward this task (Hamilton et al., 2017).

Toward Marr's framework

Marr's three levels of analysis framework has guided cognitive neuroscience, especially visual cognitive neuroscience, for the past few decades. According to Marr's framework, there are three different levels of understanding of a cognitive system (Marr, 1982):

1. the computational theory level: what computational problem does the system solve;
2. the representation and algorithm level: from algorithmic point of view, how does the system do what it does, what representation is used and how does it manipulate the representation in order to achieve the computation goal;
3. hardware implementation level: how is the system physically realized.

The works presented in this thesis mainly focus on the first two levels. Future works should explore the algorithmic and physical space to fully understand the biological basis of the neural dynamics underlying perception, which may provide insights into building artificial intelligent systems that overcome current limitations.

Bibliography

Amal Achaibou, Eva Loth, and Sonia J Bishop. Distinct frontal and amygdala correlates of change detection for facial identity and expression. *Social Cognitive and Affective Neuroscience*, 11(2):225–233, 2015. 4.1

Truett Allison, Aina Puce, Dennis D Spencer, and Gregory McCarthy. Electrophysiological studies of human face perception. i: Potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cerebral Cortex*, 9(5):415–430, 1999. 3.1, 3.3.1, 4.3.1

Elissa M Aminoff, Yuanning Li, John A Pyles, Michael J Ward, R Mark Richardson, and Avniel S Ghuman. Associative hallucinations result from stimulating left ventromedial temporal cortex. *Cortex*, 83:139–144, 2016. 1.2.4

Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958. 5.4

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning (ICML)*, pages 1247–1255, 2013. 5.4.3

Stefano Anzellotti, Scott L Fairhall, and Alfonso Caramazza. Decoding representations of face identity that are tolerant to rotation. *Cerebral Cortex*, 24(8):1988–1995, 2013. 3.1

Amos Arieli, Alexander Sterkin, Amiram Grinvald, and AD Aertsen. Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science*, 273(5283): 1868, 1996. 1.2.3, 6.1

Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366, 2006. 1.1, 6.4.1

Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, Department of Statistics, University of California, Berkeley, 2005. 5.2.2

Annelies Baeck, Dwight Kravitz, Chris Baker, and Hans P Op de Beeck. Influence of lexical status and orthographic similarity on the multi-voxel response of the visual word form area. *NeuroImage*, 111:321–328, 2015. 2.3.1

Jason JS Barton. Structure and function in acquired prosopagnosia: lessons from a series of 10 patients with brain damage. *Journal of Neuropsychology*, 2(1):197–225, 2008. 4.1

Jason JS Barton, Daniel Z Press, Julian P Keenan, and Margaret O'Connor. Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. *Neurology*, 58

(1):71–78, 2002. 3.1, 4.1

Erol Başar. *EEG-brain dynamics: relation between EEG and brain evoked potentials*. Elsevier-North-Holland Biomedical Press, 1980. 1.2.3, 6.1

Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*, 2015. 7.2

Markus Bauer, Max-Philipp Stenner, Karl J Friston, and Raymond J Dolan. Attentional modulation of alpha/beta and gamma oscillations reflect functionally distinct processes. *Journal of Neuroscience*, 34(48):16117–16125, 2014. 6.4.4

GC Baylis, Edmund T Rolls, and CM Leonard. Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey. *Brain Research*, 342(1):91–102, 1985. 3.1, 3.3.1, 3.4

Robert Becker, Petra Ritter, and Arno Villringer. Influence of ongoing alpha rhythm on the visual evoked potential. *NeuroImage*, 39(2):707–716, 2008. 6.4.1

Robert Becker, Matthias Reinacher, Frank Freyer, Arno Villringer, and Petra Ritter. How ongoing neuronal oscillations account for evoked fmri variability. *Journal of Neuroscience*, 31(30): 11016–11027, 2011. 6.4.4

Marlene Behrmann and David C Plaut. Distributed circuits, not circumscribed centers, mediate visual recognition. *Trends in Cognitive Sciences*, 17(5):210–219, 2013. 3.4

Marlene Behrmann and Tim Shallice. Pure alexia: A nonspatial visual disorder affecting letter activation. *Cognitive Neuropsychology*, 12(4):409–454, 1995. 2.1, 2.3.1, 2.4, 2.5

Michal Ben-Shachar, Robert F Dougherty, Gayle K Deutsch, and Brian A Wandell. The development of cortical sensitivity to visual word forms. *Journal of Cognitive Neuroscience*, 23(9): 2387–2399, 2011. 2.1

Shlomo Bentin and Leon Y Deouell. Structural encoding and identification in face processing: Erp evidence for separate mechanisms. *Cognitive Neuropsychology*, 17(1-3):35–55, 2000. 4.4

Shlomo Bentin, Truett Allison, Aina Puce, Erik Perez, and Gregory McCarthy. Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, 8(6):551–565, 1996. 1.2.1

Peter J Bickel and Elizaveta Levina. Some theory for fisher's linear discriminant function,'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, pages 989–1010, 2004. 4.2.6

Jeffrey R Binder, David A Medler, Chris F Westbury, Einat Liebenthal, and Lori Buchanan. Tuning of the human left fusiform gyrus to sublexical orthographic structure. *NeuroImage*, 33 (2):739–748, 2006. 2.1, 2.3.3

Sonia J Bishop, Geoffrey K Aguirre, Anwar O Nunez-Elizalde, and Daniel Toker. Seeing the world through non rose-colored glasses: anxiety and the amygdala response to blended expressions. *Frontiers in Human Neuroscience*, 9:152, 2015. 4.1

Vera C Blau, Urs Maurer, Nim Tottenham, and Bruce D McCandliss. The face-specific n170

component is modulated by emotional facial expression. *Behavioral and Brain Functions*, 3 (1):7, 2007. 4.4

Mélanie Boly, Evelyne Balteau, Caroline Schnakers, Christian Degueldre, Gustave Moonen, André Luxen, Christophe Phillips, Philippe Peigneux, Pierre Maquet, and Steven Laureys. Baseline brain activity fluctuations predict somatosensory perception in humans. *Proceedings of the National Academy of Sciences*, 104(29):12187–12192, 2007. 6.4.2

Aline Bompas, Petroc Sumner, Suresh D Muthumumaraswamy, Krish D Singh, and Iain D Gilchrist. The contribution of pre-stimulus neural oscillatory activity to spontaneous response time variability. *NeuroImage*, 107:34–45, 2015. 6.4.2

Patricia Greig Bowers and Maryanne Wolf. Theoretical links among naming speed, precise timing mechanisms and orthographic skill in dyslexia. *Reading and Writing*, 5(1):69–85, 1993. 2.4

Michael E Brandt and Ben H Jansen. The relationship between prestimulus alpha amplitude and visual evoked potential amplitude. *International Journal of Neuroscience*, 61(3-4):261–268, 1991. 1.2.3

Hans C Breiter, Nancy L Etcoff, Paul J Whalen, William A Kennedy, Scott L Rauch, Randy L Buckner, Monica M Strauss, Steven E Hyman, and Bruce R Rosen. Response and habituation of the human amygdala during visual processing of facial expression. *Neuron*, 17(5):875–887, 1996. 4.1

Silvia Brem, Silvia Bach, Karin Kucian, Janne V Kujala, Tomi K Guttorm, Ernst Martin, Heikki Lyytinen, Daniel Brandeis, and Ulla Richardson. Brain sensitivity to print emerges when children learn letter–speech sound correspondences. *Proceedings of the National Academy of Sciences*, 107(17):7939–7944, 2010. 2.1

Vicki Bruce and Andy Young. Understanding face recognition. *British Journal of Psychology*, 77(3):305–327, 1986. 3.1, 4.1, 4.4, 4.4

Maggie Bruck. Word-recognition skills of adults with childhood diagnoses of dyslexia. *Developmental Psychology*, 26(3):439, 1990. 2.4

Daniel N Bub, Martin Arguin, and Andre Roch Lecours. Jules dejerine and his interpretation of pure alexia. *Brain and Language*, 45(4):531–559, 1993. 2.1, 2.4

Peter Burt and Edward Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983. 3.2.9

Niko A Busch, Julien Dubois, and Rufin VanRullen. The phase of ongoing eeg oscillations predicts visual perception. *Journal of Neuroscience*, 29(24):7869–7876, 2009. 1.2.3, 6.1

Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, 10(12):e1003963, 2014. 7.2

Andrew J Calder and Andrew W Young. Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, 6(8):641–651, 2005. 4.1, 4.4, 4.4

Jacob Cohen. Statistical power analysis for the behavioral sciences. *NJ: Lawrence Earlbaum*

Associates, 2, 1988. 4.2.6, 4.4

Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff, and François Michel. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2):291–307, 2000. 1.1

Laurent Cohen, Stéphane Lehéricy, Florence Chochon, Cathy Lemer, Sophie Rivaud, and Stanislas Dehaene. Language-specific tuning of visual cortex? functional properties of the visual word form area. *Brain*, 125(5):1054–1069, 2002. 2.1

Jessica A Collins and Ingrid R Olson. Beyond the ffa: the role of the ventral anterior temporal lobes in face processing. *Neuropsychologia*, 61:65–79, 2014. 3.4

Marc N Coutanche and Sharon L Thompson-Schill. Informational connectivity: identifying synchronized discriminability of multi-voxel patterns across the brain. *Frontiers in Human Neuroscience*, 7, 2013. 1.2.2, 5.4.3

Alan S Cowen, Marvin M Chun, and Brice A Kuhl. Neural portraits of perception: reconstructing face images from evoked brain activity. *NeuroImage*, 94:12–22, 2014. 3.1

David D Cox and Robert L Savoy. Functional magnetic resonance imaging (fmri) "brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. *NeuroImage*, 19(2):261–270, 2003. 5.1

Ivor Cribben, Ragnheidur Haraldsdottir, Lauren Y Atlas, Tor D Wager, and Martin A Lindquist. Dynamic connectivity regression: determining state-related changes in brain connectivity. *NeuroImage*, 61(4):907–920, 2012. 5.1

Sebastien M Crouzet, Holle Kirchner, and Simon J Thorpe. Fast saccades toward faces: face detection in just 100 ms. *Journal of Vision*, 10(4):16–16, 2010. 3.1, 3.3.1, 3.4

Ido Davidesco, Elana Zion-Golumbic, Stephan Bickel, Michal Harel, David M Groppe, Corey J Keller, Catherine A Schevon, Guy M McKhann, Robert R Goodman, Gadi Goelman, et al. Exemplar selectivity reflects perceptual similarities in the human fusiform cortex. *Cerebral Cortex*, 24(7):1879–1893, 2013. 3.1, 3.3.3, 3.3.4

Peter Dayan and Laurence F Abbott. *Theoretical neuroscience*, volume 806. Cambridge, MA: MIT Press, 2001. 1.1

Joseph DeGutis, Jeremy Wilmer, Rogelio J Mercado, and Sarah Cohan. Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition*, 126(1):87–100, 2013. 3.1, 3.4

Stanislas Dehaene and Laurent Cohen. The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6):254–262, 2011. 2.1, 2.3.3, 2.4, 4.1, 4.4

Stanislas Dehaene, Gurvan Le Clec'H, Jean-Baptiste Poline, Denis Le Bihan, and Laurent Cohen. The visual word form area: a prelexical representation of visual words in the fusiform gyrus. *Neuroreport*, 13(3):321–325, 2002. 2.1

Stanislas Dehaene, Felipe Pegado, Lucia W Braga, Paulo Ventura, Gilberto Nunes Filho, Antoinette Jobert, Ghislaine Dehaene-Lambertz, Régine Kolinsky, José Morais, and Laurent Cohen. How learning to read changes the cortical networks for vision and language. *Science*,

330(6009):1359–1364, 2010. 2.1

Robert Desimone, Thomas D Albright, Charles G Gross, and Charles Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4(8):2051–2062, 1984. 5.1

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. 1.1

Paul E Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001. 1.1

Brad Duchaine and Galit Yovel. A revised neural framework for face processing. *Annual Review of Vision Science*, 1:393–416, 2015. 4.1, 4.4, 4.4

Keith J Duncan, Chotiga Pattamadilok, and Joseph T Devlin. Investigating occipito-temporal contributions to reading with tms. *Journal of Cognitive Neuroscience*, 22(4):739–750, 2010. 2.3.3

Shimon Edelman, Kalanit Grill-Spector, Tammar Kushnir, and Rafael Malach. Toward direct visualization of the internal shape representation space by fmri. *Psychobiology*, 26(4):309–321, 1998. 5.1, 5.4.2

Simon B Eickhoff, Angela R Laird, Christian Grefkes, Ling E Wang, Karl Zilles, and Peter T Fox. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping*, 30(9):2907–2926, 2009. 4.2.12

Simon B Eickhoff, Danilo Bzdok, Angela R Laird, Florian Kurth, and Peter T Fox. Activation likelihood estimation meta-analysis revisited. *NeuroImage*, 59(3):2349–2361, 2012. 4.2.12

Martin Eimer. Effects of face inversion on the structural encoding and recognition of faces: Evidence from event-related brain potentials. *Cognitive Brain Research*, 10(1):145–158, 2000a. 4.4

Martin Eimer. Event-related brain potentials distinguish processing stages involved in face perception and recognition. *Clinical Neurophysiology*, 111(4):694–705, 2000b. 4.4

Martin Eimer. The face-specific n170 component reflects late stages in the structural encoding of faces. *Neuroreport*, 11(10):2319–2324, 2000c. 4.3.1, 4.4

Martin Eimer. The face-sensitive n170 component of the event-related brain potential. *The Oxford handbook of face perception*, 28, 2011. 4.3.1, 4.4

Martin Eimer, Amanda Holmes, and Francis P McGlone. The role of spatial attention in the processing of facial expression: an erp study of rapid brain responses to six basic emotions. *Cognitive, Affective, & Behavioral Neuroscience*, 3(2):97–110, 2003. 4.4

Andrew D Engell and Gregory McCarthy. The relationship of gamma oscillations and face-specific erps recorded subdurally from occipitotemporal cortex. *Cerebral Cortex*, 21(5):1213–1221, 2010. 3.1, 3.3.1, 3.3.4, 3.4, 4.4

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998. 1.1

Tolgay Ergenoglu, Tamer Demiralp, Zubeyir Bayraktaroglu, Mehmet Ergen, Huseyin Beydagi, and Yagiz Uresin. Alpha rhythm of the eeg modulates visual detection performance in humans. *Cognitive Brain Research*, 20(3):376–383, 2004. 1.2.3

Martha J Farah. *Visual agnosia*. MIT press, 2004. 1.1

Martha J Farah and Marcie A Wallace. Pure alexia as a visual impairment: A reconsideration. *Cognitive Neuropsychology*, 8(3-4):313–334, 1991. 2.1

Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex (New York, NY: 1991)*, 1(1):1–47, 1991. 1.1

Robert Fellinger, Wolfgang Klimesch, Walter Gruber, Roman Freunberger, and Michael Doppelmayr. Pre-stimulus alpha phase-alignment predicts p1-amplitude. *Brain Research Bulletin*, 85 (6):417–423, 2011. 6.4.1

Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11): 1664–1671, 2015. 1.2.2, 5.1, 5.4.3

Jerry A Fodor. *The modularity of mind: An essay on faculty psychology*. MIT press, 1983. 5.1, 5.4.1

Elaine Foley, Gina Rippon, Ngoc Jade Thai, Olivia Longe, and Carl Senior. Dynamic facial expressions evoke distinct activation in the face perception network: a connectivity analysis study. *Journal of Cognitive Neuroscience*, 24(2):507–520, 2012. 4.1

Christopher J Fox, So Young Moon, Giuseppe Iaria, and Jason JS Barton. The correlates of subjective perception of identity and expression in the face network: an fmri adaptation study. *NeuroImage*, 44(2):569–580, 2009. 4.1

Michael D Fox, Abraham Z Snyder, Jeffrey M Zacks, and Marcus E Raichle. Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. *Nature Neuroscience*, 9(1):23–25, 2006. 1.2.3, 6.1, 6.4.4

David J Freedman, Maximilian Riesenhuber, Tomaso Poggio, and Earl K Miller. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, 23(12):5235–5246, 2003. 3.1, 3.3.4, 3.4

Winrich A Freiwald and Doris Y Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005):845–851, 2010. 4.4

Winrich A Freiwald, Doris Y Tsao, and Margaret S Livingstone. A face feature space in the macaque temporal lobe. *Nature Neuroscience*, 12(9):1187–1196, 2009. 3.1, 3.3.3, 5.1

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010. 6.2.4

Pascal Fries, Sergio Neuenschwander, Andreas K Engel, Rainer Goebel, and Wolf Singer. Rapid feature selective neuronal synchronization through correlated latency shifting. *Nature Neuroscience*, 4(2), 2001. 1.2.3

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. 5.4.4

Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *NeuroImage*, 19(4): 1273–1302, 2003. 5.1

KJ Friston, CD Frith, PF Liddle, and RSJ Frackowiak. Functional connectivity: the principal-component analysis of large (pet) data sets. *Journal of Cerebral Blood Flow & Metabolism*, 13(1):5–14, 1993. 1.1, 1.1

KJ Friston, C Buechel, GR Fink, J Morris, E Rolls, and RJ Dolan. Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage*, 6(3):218–229, 1997. 5.1, 5.2.1, 5.2.5, 5.4.1

Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982. 1.1

Maura L Furey, Topi Tanskanen, Michael S Beauchamp, Sari Avikainen, Kimmo Uutela, Riitta Hari, and James V Haxby. Dissociation of face-selective cortical responses by attention. *Proceedings of the National Academy of Sciences*, 103(4):1065–1070, 2006. 4.4

Nicholas Furl, Nicola J van Rijsbergen, Alessandro Treves, Karl J Friston, and Raymond J Dolan. Experience-dependent coding of facial expression in superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 104(33):13485–13489, 2007. 4.4

Nicholas Furl, Michael Lohse, and Francesca Pizzorni-Ferrarese. Low-frequency oscillations employ a general coding of the spatio-temporal similarity of dynamic faces. *NeuroImage*, 157:486–499, 2017. 4.2.6, 4.3.4

Raphaël Gaillard, Lionel Naccache, Philippe Pinel, Stéphane Clémenceau, Emmanuelle Volle, Dominique Hasboun, Sophie Dupont, Michel Baulac, Stanislas Dehaene, Claude Adam, et al. Direct intracranial, fmri, and lesion evidence for the causal role of left inferotemporal cortex in reading. *Neuron*, 50(2):191–204, 2006. 2.1, 2.3.1

Tzvi Ganel, Kenneth F Valyear, Yonatan Goshen-Gottstein, and Melvyn A Goodale. The involvement of the fusiform face area in processing facial expression. *Neuropsychologia*, 43 (11):1645–1654, 2005. 4.1

George L Gerstein and Donald H Perkel. Simultaneously recorded trains of action potentials: analysis and functional interpretation. *Science*, 164(3881):828–830, 1969. 1.1

George L Gerstein, Purvis Bedenbaugh, and Ad MHJ Aertsen. Neuronal assemblies. *IEEE Transactions on Biomedical Engineering*, 36(1):4–14, 1989. 1.1

Avniel Singh Ghuman, Nicolas M Brunet, Yuanning Li, Roma O Konecky, John A Pyles, Shawn A Walls, Vincent Destefino, Wei Wang, and R Mark Richardson. Dynamic encoding of face information in the human fusiform gyrus. *Nature Communications*, 5:5672, 2014. 1.2.4, 2.2.2, 2.4, 4.1, 4.2.2, 4.2.6, 4.2.10, 4.3.1, 4.3.7, 4.4, 4.4, 5.1, 5.2.6, 5.3.6, 5.4.4, 6.4.3, 6.5.2, 6.5.3

Charles D Gilbert and Wu Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, 2013. 1.2.2, 5.4.1

Robert Gittins. *Canonical analysis: a review with applications in ecology*, volume 12. Springer Science & Business Media, 2012. 5.4

Laurie S Glezer, Xiong Jiang, and Maximilian Riesenhuber. Evidence for highly selective neuronal tuning to whole words in the "visual word form area". *Neuron*, 62(2):199–204, 2009. 2.1

Laurie S Glezer, Judy Kim, Josh Rule, Xiong Jiang, and Maximilian Riesenhuber. Adding words to the brain's visual dictionary: novel word learning selectively sharpens orthographic representations in the vwfa. *Journal of Neuroscience*, 35(12):4965–4972, 2015. 2.1

Elfi Goesaert and Hans P Op de Beeck. Representations of facial identity information in the ventral visual stream investigated with multivoxel pattern analyses. *Journal of Neuroscience*, 33(19):8549–8558, 2013. 3.1, 4.1

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 5.4.1, 7.2

Stephen Grossberg. How does a brain build a cognitive code? In *Studies of mind and brain*, pages 1–52. Springer, 1982. 5.1

Markus Gschwind, Gilles Pourtois, Sophie Schwartz, Dimitri Van De Ville, and Patrik Vuilleumier. White-matter connectivity between face-responsive regions in the human brain. *Cerebral Cortex*, 22(7):1564–1576, 2011. 5.4.4

Nigel D Haig. Exploring recognition with interchanged facial features. *Perception*, 15(3):235–247, 1986. 3.1, 3.4

Carlos M Hamamé, Marcin Szwed, Michael Sharman, Juan R Vidal, Marcella Perrone-Bertolotti, Philippe Kahane, Olivier Bertrand, and Jean-Philippe Lachaux. Dejerine's reading area revisited with intracranial eeg selective responses to letter strings. *Neurology*, 80(6):602–603, 2013. 2.3.1, 2.3.2

Carlos M Hamamé, Juan R Vidal, Marcela Perrone-Bertolotti, Tomas Ossandón, Karim Jerbi, Philippe Kahane, Olivier Bertrand, and Jean-Philippe Lachaux. Functional selectivity in the human occipitotemporal cortex during natural vision: Evidence from combined intracranial eeg and eye-tracking. *NeuroImage*, 95:276–286, 2014. 2.3.1

William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017. 7.2

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004. 5.2.2, 5.4.3

Richard J Harris, Andrew W Young, and Timothy J Andrews. Brain regions involved in processing facial identity and expression are differentially selective for surface and edge information. *NeuroImage*, 97:217–223, 2014. 4.4

Bronson Harry, Mark A Williams, Chris Davis, and Jeesun Kim. Emotional expressions evoke a differential response in the fusiform face area. *Frontiers in Human Neuroscience*, 7, 2013. 4.4

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015. 6.2.4, 9

James V Haxby, Elizabeth A Hoffman, and M Ida Gobbini. The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4(6):223–233, 2000. 1.2.1, 3.1, 3.4, 4.1, 4.4, 4.4

James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001. 1.1, 1.1, 1.2.2, 5.1, 6.4.1

James V Haxby, Andrew C Connolly, and J Swaroop Guntupalli. Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, 37:435–456, 2014. 1.2.2, 4.2.6, 5.1, 5.2.1, 5.4.2

John-Dylan Haynes and Geraint Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534, 2006. 5.1

Linda Henriksson, Seyed-Mahdi Khaligh-Razavi, Kendrick Kay, and Nikolaus Kriegeskorte. Visual representations are dominated by intrinsic fluctuations correlated between areas. *NeuroImage*, 114:275–286, 2015. 1.2.3, 5.3.8, 6.1, 6.4.4

Dora Hermes, Kai J Miller, Herke Jan Noordmans, Mariska J Vansteensel, and Nick F Ramsey. Automated electrocorticographic electrode localization on individually rendered brain surfaces. *Journal of Neuroscience Methods*, 185(2):293–298, 2010. 3.2.5, 4.2.4, 6.2.4

Guido Hesselmann, Christian A Kell, and Andreas Kleinschmidt. Ongoing activity fluctuations in hmt+ bias the perception of coherent visual motion. *Journal of Neuroscience*, 28(53): 14481–14485, 2008. 6.4.2

Elizabeth A Hirshorn, Yuanning Li, Michael J Ward, R Mark Richardson, Julie A Fiez, and Avniel Singh Ghuman. Decoding and disrupting left midfusiform gyrus activity during word reading. *Proceedings of the National Academy of Sciences*, 113(29):8162–8167, 2016. 1.2.4, 4.2.6, 5.1, 6.4.3

Elizabeth A Hoffman and James V Haxby. Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, 3(1):80–84, 2000. 4.1

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. 3.2.10

David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591, 1959. 5.1

Alumit Ishai. Let's face it: it's a cortical network. *NeuroImage*, 40(2):415–419, 2008. 1.1, 4.1, 4.4, 5.4.4

Roxane J Itier and Margot J Taylor. N170 or n1? spatiotemporal differences between object and face processing using erps. *Cerebral Cortex*, 14(2):132–142, 2004. 3.1, 3.3.1, 3.4

Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–685, 2005. 5.1

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11): 4302–4311, 1997. 1.1, 3.1

Elissaios Karageorgiou, Scott M Lewis, J Riley McCarten, Arthur C Leuthold, Laura S Hemmy, Susan E McPherson, Susan J Rottunda, David M Rubins, and Apostolos P Georgopoulos. Canonical correlation analysis of synchronous neural interactions and cognitive deficits in alzheimer's dementia. *Journal of Neural Engineering*, 9(5):056003, 2012. 5.4

Robert E Kass and Larry Wasserman. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90(431): 928–934, 1995. 4.2.9, 4.3.3

Sabine Kastner, Mark A Pinsk, Peter De Weerd, Robert Desimone, and Leslie G Ungerleider. Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, 22(4):751–761, 1999. 6.1, 6.4.4

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009. 4.2.8, 4.2.9, 4.3.3

Hiroto Kawasaki, Naotsugu Tsuchiya, Christopher K Kovach, Kirill V Nourski, Hiroyuki Oya, Matthew A Howard, and Ralph Adolphs. Processing of facial emotion in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, 24(6):1358–1370, 2012. 3.1, 3.3.4, 3.4, 4.1, 4.4

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. 5.2.5, 5.3.2, 5.3.8

KN Kay, T Naselaris, and JL Gallant. fmri of human visual areas in response to natural images. *CRCNS. org*, 2011. 5.2.5, 5.3.2

Stephanie J Kayser, Steven W McNair, and Christoph Kayser. Prestimulus influences on auditory perception from sensory representations and decision processes. *Proceedings of the National Academy of Sciences*, 113(17):4842–4847, 2016. 6.1, 6.4.2

Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971. 5.4.3

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11): e1003915, 2014. 5.4.4

Michael A Kisley and George L Gerstein. Trial-to-trial variability and state-dependent modulation of auditory-evoked responses in cortex. *Journal of Neuroscience*, 19(23):10451–10460, 1999. 1.2.3, 6.1

Wolfgang Klimesch. Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2-3):169–195, 1999. 6.1

Dwight J Kravitz, Kadharbatcha S Saleem, Chris I Baker, Leslie G Ungerleider, and Mortimer Mishkin. The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, 17(1):26–49, 2013. 1.1, 1.1, 1.2.1

Nikolaus Kriegeskorte. Pattern-information analysis: from stimulus decoding to computational-model testing. *NeuroImage*, 56(2):411–421, 2011. 5.4.2

Nikolaus Kriegeskorte and Rogier A Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412, 2013. 1.2.2, 4.2.11, 5.1, 5.4.2, 5.4.3

Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National academy of Sciences*, 103(10):3863–3868, 2006. 5.4.3, 6.4.1

Nikolaus Kriegeskorte, Elia Formisano, Bettina Sorger, and Rainer Goebel. Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, 104(51):20600–20605, 2007. 3.4

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008. 5.4.2

Kestutis Kveraga, Avniel S Ghuman, and Moshe Bar. Top-down predictions in the cognitive brain. *Brain and Cognition*, 65(2):145–168, 2007. 5.1

Angela R Laird, P Mickle Fox, Cathy J Price, David C Glahn, Angela M Uecker, Jack L Lancaster, Peter E Turkeltaub, Peter Kochunov, and Peter T Fox. Ale meta-analysis: Controlling the false discovery rate and performing statistical contrasts. *Human Brain Mapping*, 25(1): 155–164, 2005. 4.2.12

Antonio H Lara and Jonathan D Wallis. Executive control processes underlying multi-item working memory. *Nature Neuroscience*, 17(6):876–883, 2014. 3.3.4, 3.4

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: A compact convolutional network for eeg-based brain-computer interfaces. *arXiv preprint arXiv:1611.08024*, 2016. 7.2

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 1.1

Tae-Ho Lee, Steven G Greening, Taiji Ueno, David Clewett, Allison Ponzio, Michiko Sakaki, and Mara Mather. Arousal increases neural gain via the locus coeruleus–noradrenaline system in younger adults but not in older adults. *Nature Human Behaviour*, 2(5):356, 2018. 6.4.4

Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7):1434–1448, 2003. 5.1, 5.4.4

David A Leopold, Igor V Bondar, and Martin A Giese. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 442(7102):572–575, 2006. 3.1

Roger Levy, Klinton Bicknell, Tim Slattery, and Keith Rayner. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50):21086–21090, 2009. 2.4

Yuanning Li, R Mark Richardson, and Avniel Singh Ghuman. Multi-connection pattern analysis: Decoding the representational content of neural communication. *NeuroImage*, 162:32–44, 2017. 1.2.4, 4.4

Yuanning Li, R Mark Richardson, and Avniel Singh Ghuman. Posterior fusiform and mid-fusiform contribute to distinct stages of facial expression processing. *Cerebral Cortex*, 2018. 1.2.4, 6.4.3, 6.5.2

Klaus Linkenkaer-Hansen, Vadim V Nikouline, J Matias Palva, and Risto J Ilmoniemi. Long-

range temporal correlations and scaling behavior in human brain oscillations. *Journal of Neuroscience*, 21(4):1370–1377, 2001. 6.4.4

Klaus Linkenkaer-Hansen, Vadim V Nikulin, Satu Palva, Risto J Ilmoniemi, and J Matias Palva. Prestimulus oscillations enhance psychophysical performance in humans. *Journal of Neuroscience*, 24(45):10186–10190, 2004. 6.4.2

Hesheng Liu, Yigal Agam, Joseph R Madsen, and Gabriel Kreiman. Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, 62(2):281–290, 2009. 3.1, 3.3.1, 3.4

Steven J Luck, Leonardo Chelazzi, Steven A Hillyard, and Robert Desimone. Neural mechanisms of spatial selective attention in areas v1, v2, and v4 of macaque visual cortex. *Journal of Neurophysiology*, 77(1):24–42, 1997. 6.4.4

Artur Luczak, Peter Barthó, and Kenneth D Harris. Spontaneous events outline the realm of possible sensory responses in neocortical populations. *Neuron*, 62(3):413–425, 2009. 1.2.3

Daniel Lundqvist, Anders Flykt, and Arne Öhman. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, (1998), 1998. 3.2.2, 4.2.2, 5.2.6

J Mani, B Diehl, Z Piao, SS Schuele, E Lapresto, P Liu, DR Nair, DS Dinner, and HO Lüders. Evidence for a basal temporal visual language center cortical stimulation producing pure alexia. *Neurology*, 71(20):1621–1627, 2008. 2.3.2

Jeremy R Manning, Joshua Jacobs, Itzhak Fried, and Michael J Kahana. Broadband shifts in local field potential power spectra are correlated with single-neuron spiking in humans. *Journal of Neuroscience*, 29(43):13613–13620, 2009. 3.1, 3.4

Ran Manor and Amir B Geva. Convolutional neural network for multi-category rapid serial visual presentation bci. *Frontiers in Computational Neuroscience*, 9, 2015. 7.2

Eric Maris and Robert Oostenveld. Nonparametric statistical testing of eeg-and meg-data. *Journal of Neuroscience Methods*, 164(1):177–190, 2007. 2.2.10, 3.2.8, 5.2.6

David Marr. Vision: A computational investigation into the human representation and processing of visual information. mit press. *Cambridge, Massachusetts*, 1982. 7.2

Alex Martin. The representation of object concepts in the brain. *Annual Review of Psychology*, 58:25–45, 2007. 1.1

Anna Martin, Martin Kronbichler, and Fabio Richlan. Dyslexic brain activation abnormalities in deep and shallow orthographies: A meta-analysis of 28 functional neuroimaging studies. *Human Brain Mapping*, 37(7):2676–2699, 2016. 2.4

Kyle E Mathewson, Gabriele Gratton, Monica Fabiani, Diane M Beck, and Tony Ro. To see or not to see: prestimulus α phase predicts visual awareness. *Journal of Neuroscience*, 29(9): 2725–2732, 2009. 1.2.3, 6.1, 6.4.2

Urs Maurer, Daniel Brandeis, and Bruce D McCandliss. Fast, visual specialization for reading in english revealed by the topography of the n170 erp response. *Behavioral and Brain Functions*, 1(1):13, 2005. 2.3.1

Bruce D McCandliss, Laurent Cohen, and Stanislas Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, 7(7):293–299, 2003. 1.2.1, 2.1

Gregory McCarthy, Aina Puce, John C Gore, and Truett Allison. Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, 9(5):605–610, 1997. 3.1

Gregory McCarthy, Aina Puce, Aysenil Belger, and Truett Allison. Electrophysiological studies of human face perception. ii: Response properties of face-specific potentials generated in occipitotemporal cortex. *Cerebral Cortex*, 9(5):431–444, 1999. 3.4

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008. 6.2.4

Kai J Miller, Gerwin Schalk, Dora Hermes, Jeffrey G Ojemann, and Rajesh PN Rao. Spontaneous decoding of the timing and content of human object perception from cortical surface recordings reveals complementary information in the event-related potential and broadband spectral change. *PLoS Computational Biology*, 12(1):e1004660, 2016. 2.2.10, 4.2.6

Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 6:414–417, 1983. 1.1

Kathrin Müsch, Carlos M Hamamé, Marcela Perrone-Bertolotti, Lorella Minotti, Philippe Kahane, Andreas K Engel, Jean-Philippe Lachaux, and Till R Schneider. Selective attention modulates high-frequency activity in the face-processing network. *Cortex*, 60:34–51, 2014. 4.1, 4.4

Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009. 5.2.5, 5.3.2

Adrian Nestor, David C Plaut, and Marlene Behrmann. Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences*, 108(24):9998–10003, 2011. 1.1, 2.2.10, 3.1, 3.4, 4.1, 5.1, 5.4.4

Anna C Nobre, Truett Allison, Gregory McCarthy, et al. Word recognition in the human inferior temporal lobe. *Nature*, 372(6503):260–263, 1994. 2.3.1, 2.3.2

Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in Cognitive Sciences*, 10(9):424–430, 2006. 1.1, 5.1, 5.4.3, 6.4.1

Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011:1, 2011. 4.2.3, 6.2.4

Jill X O'Reilly, Mark W Woolrich, Timothy EJ Behrens, Stephen M Smith, and Heidi Johansen-Berg. Tools of the trade: psychophysiological interactions and functional connectivity. *Social Cognitive and Affective Neuroscience*, 7(5):604–609, 2012. 5.4.3

Josef Parvizi, Corentin Jacques, Brett L Foster, Nathan Withoft, Vinitha Rangarajan, Kevin S Weiner, and Kalanit Grill-Spector. Electrical stimulation of human fusiform face-selective

regions distorts face perception. *Journal of Neuroscience*, 32(43):14915–14920, 2012. 2.3.2, 3.4

Marius V Peelen and Paul E Downing. The neural basis of visual body perception. *Nature Reviews Neuroscience*, 8(8):636–648, 2007. 3.4

Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. *International Conference on Machine Learning (ICML)*, 1:727–734, 2000. 4.2.9

DI Perrett, Edmond T Rolls, and W Caan. Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research*, 47(3):329–342, 1982. 3.1, 3.3.1, 3.4

David Pitcher, Lucie Charles, Joseph T Devlin, Vincent Walsh, and Bradley Duchaine. Triple dissociation of faces, bodies, and objects in extrastriate cortex. *Current Biology*, 19(4):319–324, 2009. 3.1, 3.4

David Pitcher, Daniel D Dilks, Rebecca R Saxe, Christina Triantafyllou, and Nancy Kanwisher. Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage*, 56(4):2356–2363, 2011. 4.1

David Pitcher, Tanya Goldhaber, Bradley Duchaine, Vincent Walsh, and Nancy Kanwisher. Two critical and functionally distinct stages of face and body perception. *Journal of Neuroscience*, 32(45):15877–15885, 2012. 3.1

Russell A Poldrack. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, 72(5):692–697, 2011. 5.1

Sean M Polyn, Vaidehi S Natu, Jonathan D Cohen, and Kenneth A Norman. Category-specific cortical activity precedes retrieval during memory search. *Science*, 310(5756):1963–1966, 2005. 5.1

Gilles Pourtois, Laurent Spinelli, Margitta Seeck, and Patrik Vuilleumier. Modulation of face processing by emotional expression and gaze direction during intracranial recordings in right fusiform cortex. *Journal of Cognitive Neuroscience*, 22(9):2086–2107, 2010. 4.1, 4.4

Cathy J Price and Joseph T Devlin. The myth of the visual word form area. *NeuroImage*, 19(3): 473–481, 2003. 1.2.1, 2.1

Cathy J Price and Joseph T Devlin. The interactive account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences*, 15(6):246–253, 2011. 2.1, 2.3.3, 2.4

Aina Puce, Truett Allison, and Gregory McCarthy. Electrophysiological studies of human face perception. iii: Effects of top-down processing on face-specific potentials. *Cerebral Cortex*, 9 (5):445–458, 1999. 1.1, 1.2.1

John A Pyles, Timothy D Verstynen, Walter Schneider, and Michael J Tarr. Explicating the face perception network with white matter connectivity. *PLoS One*, 8(4):e61611, 2013. 3.4, 5.4.4, 6.4.4

Rajasimhan Rajagovindan and Mingzhou Ding. From prestimulus alpha oscillation to visual-evoked response: an inverted-u function and its attentional modulation. *Journal of Cognitive Neuroscience*, 23(6):1379–1394, 2011. 6.4.1

Supratim Ray and John HR Maunsell. Network rhythms influence the relationship between spike-triggered local field potential and functional connectivity. *Journal of Neuroscience*, 31 (35):12674–12682, 2011. 3.1, 3.4

Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. Toward a model of eye movement control in reading. *Psychological Review*, 105(1):125, 1998. 2.4

David Ress, Benjamin T Backus, and David J Heeger. Activity in primary visual cortex predicts performance in a visual detection task. *Nature Neuroscience*, 3(9), 2000. 1.2.3, 6.1

Jonas Richiardi, Hamdi Eryilmaz, Sophie Schwartz, Patrik Vuilleumier, and Dimitri Van De Ville. Decoding brain states from fmri connectivity graphs. *NeuroImage*, 56(2):616–626, 2011. 1.2.2, 5.1, 5.4.3

BARRY J Richmond, LANCE M Optican, MICHAEL Podell, and HEDVA Spitzer. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. i. response characteristics. *Journal of Neurophysiology*, 57(1):132–146, 1987. 3.4

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999. 1.1, 5.3.3, 5.4.2

Maximilian Riesenhuber and Tomaso Poggio. Models of object recognition. *Nature Neuroscience*, 3:1199–1204, 2000. 1.1

Monica D Rosenberg, Emily S Finn, Dustin Scheinost, Xenophon Papademetris, Xilin Shen, R Todd Constable, and Marvin M Chun. A neuromarker of sustained attention from whole-brain functional connectivity. *Nature Neuroscience*, 19(1):165–171, 2016. 1.2.2, 5.4.3

Bruno Rossion and Stéphanie Caharel. Erp evidence for the speed of face categorization in the human brain: Disentangling the contribution of low-level visual cues from face perception. *Vision Research*, 51(12):1297–1311, 2011. 3.1, 3.3.1

David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive Modeling*, 5(3):1, 1988. 5.4.4

Sepideh Sadaghiani, Guido Hesselmann, and Andreas Kleinschmidt. Distributed and antagonistic contributions of ongoing activity fluctuations to auditory stimulus detection. *Journal of Neuroscience*, 29(42):13410–13417, 2009. 6.4.2

Christopher P Said, Christopher D Moore, Andrew D Engell, Alexander Todorov, and James V Haxby. Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision*, 10(5):11–11, 2010. 5.4.4

Sameer Saproo and John T Serences. Spatial attention improves the quality of population codes in human visual cortex. *Journal of Neurophysiology*, 104(2):885–895, 2010. 6.4.4

Shlomo S Sawilowsky. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):597–599, 2009. 4.2.6

Zeynep M Saygin, David E Osher, Kami Koldewyn, Gretchen Reynolds, John DE Gabrieli, and Rebecca R Saxe. Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. *Nature Neuroscience*, 15(2):321, 2012. 6.4.4

René Scheeringa, Ali Mazaheri, Ingo Bojak, David G Norris, and Andreas Kleinschmidt. Mod-

ulation of visually evoked cortical fmri responses by phase of ongoing occipital alpha oscillations. *Journal of Neuroscience*, 31(10):3813–3820, 2011. 6.4.1

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. 7.2

Bradley L Schlaggar and Bruce D McCandliss. Development of neural systems for reading. *Annual Review of Neuroscience*, 30:475–503, 2007. 2.1

Marieke L Schölvinck, Karl J Friston, and Geraint Rees. The influence of spontaneous activity on stimulus processing in primary visual cortex. *NeuroImage*, 59(3):2700–2708, 2012. 6.4.2

William W Seeley, Vinod Menon, Alan F Schatzberg, Jennifer Keller, Gary H Glover, Heather Kenna, Allan L Reiss, and Michael D Greicius. Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, 27(9):2349–2356, 2007. 6.1

Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429, 2007a. 1.1

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007b. 5.2.5, 5.3.3

Michael N Shadlen and William T Newsome. Neural basis of a perceptual decision in the parietal cortex (area lip) of the rhesus monkey. *Journal of Neurophysiology*, 86(4):1916–1936, 2001. 3.1, 3.3.4, 3.4

WR Shirer, S Ryali, E Rykhlevskaia, V Menon, and MD Greicius. Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cerebral Cortex*, 22(1):158–165, 2012. 1.2.2, 5.1, 5.4.3

Jennifer Shum, Dora Hermes, Brett L Foster, Mohammad Dastjerdi, Vinitha Rangarajan, Jonathan Winawer, Kai J Miller, and Josef Parvizi. A brain area for visual numerals. *Journal of Neuroscience*, 33(16):6709–6715, 2013. 2.5

Noah Simon and Robert Tibshirani. Standardization and the group lasso penalty. *Statistica Sinica*, 22(3):983, 2012. 6.2.4, 9

Amy E Skerry and Rebecca Saxe. A common neural code for perceived and inferred emotion. *Journal of Neuroscience*, 34(48):15997–16008, 2014. 4.4, 5.4.4

Dirk JA Smit, Klaus Linkenkaer-Hansen, and Eco JC de Geus. Long-range temporal correlations in resting-state alpha oscillations predict human timing-error dynamics. *Journal of Neuroscience*, 33(27):11212–11220, 2013. 6.4.4

Stephen M Smith, Thomas E Nichols, Diego Vidaurre, Anderson M Winkler, Timothy EJ Behrens, Matthew F Glasser, Kamil Ugurbil, Deanna M Barch, David C Van Essen, and Karla L Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, 18(11):1565–1567, 2015. 5.4

Sebastian Stober, Daniel J Cameron, and Jessica A Grahn. Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1449–1457, 2014. 7.2

M Streit, AA Ioannides, L Liu, W Wölwer, J Dammers, J Gross, W Gaebel, and H-W Müller-Gärtner. Neurophysiological correlates of the recognition of facial expressions of emotion as revealed by magnetoencephalography. *Cognitive Brain Research*, 7(4):481–491, 1999. 4.1

Yasuko Sugase, Shigeru Yamane, Shoogo Ueno, and Kenji Kawano. Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, 400(6747):869–873, 1999. 1.1, 1.2.1, 3.1, 3.3.1, 3.4

François Tadel, Sylvain Baillet, John C Mosher, Dimitrios Pantazis, and Richard M Leahy. Brainstorm: a user-friendly application for meg/eeg analysis. *Computational Intelligence and Neuroscience*, 2011:8, 2011. 4.2.4, 6.2.4

Kazuyo Tanji, Kyoko Suzuki, Arnaud Delorme, Hiroshi Shamoto, and Nobukazu Nakasato. High-frequency γ-band activity in the basal temporal cortex during picture-naming and lexical-decision tasks. *Journal of Neuroscience*, 25(13):3287–3293, 2005. 2.3.3

Cibu Thomas, Galia Avidan, Kate Humphreys, Kwan-jin Jung, Fuqiang Gao, and Marlene Behrmann. Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia. *Nature Neuroscience*, 12(1):29–31, 2009. 3.4

Kathleen M Thomas, Wayne C Drevets, Paul J Whalen, Clayton H Eccard, Ronald E Dahl, Neal D Ryan, and BJ Casey. Amygdala response to facial expressions in children and adults. *Biological Psychiatry*, 49(4):309–316, 2001. 4.1

Gregor Thut, Annika Nietzel, Stephan A Brandt, and Alvaro Pascual-Leone. α-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection. *Journal of Neuroscience*, 26(37):9494–9502, 2006. 1.2.3, 6.1

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 6.3.4

Andreĭ Nikolaevich Tikhonov and Vasiliĭ Arsenin. *Solutions of ill-posed problems*, volume 14. Winston Washington, DC, 1977. 5.2.2

Joseph K Torgesen, Carol Alexander Rashotte, and Richard K Wagner. *TOWRE: Test of word reading efficiency*. Pro-ed Austin, TX, 1999. 2.5, 2.5, 2.10

Doris Y Tsao, Winrich A Freiwald, Roger BH Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006. 3.1, 3.3.1, 3.4, 5.1

M Tsodyks, Tal Kenet, Amiram Grinvald, and A Arieli. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, 286(5446):1943–1946, 1999. 1.2.3

Naotsugu Tsuchiya, Hiroto Kawasaki, Hiroyuki Oya, Matthew A Howard III, and Ralph Adolphs. Decoding face information in time, frequency and space from direct intracranial recordings of the human brain. *PLoS One*, 3(12):e3892, 2008. 4.1, 4.4

Peter E Turkeltaub, Simon B Eickhoff, Angela R Laird, Mick Fox, Martin Wiener, and Peter Fox.

Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Human Brain Mapping*, 33(1):1–13, 2012. 4.2.12

Hanneke Van Dijk, Jan-Mathijs Schoffelen, Robert Oostenveld, and Ole Jensen. Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. *Journal of Neuroscience*, 28(8):1816–1823, 2008. 1.2.3

Rufin VanRullen, NA Busch, Jan Drewes, and Julien Dubois. Ongoing eeg phase as a trial-by-trial predictor of perceptual and attentional variability. *Frontiers in Psychology*, 2, 2011. 1.2.3, 6.1

Mark D Vida, Adrian Nestor, David C Plaut, and Marlene Behrmann. Spatiotemporal dynamics of similarity-based neural representations of facial identity. *Proceedings of the National Academy of Sciences*, 114(2):388–393, 2017. 4.4

Fabien Vinckier, Stanislas Dehaene, Antoinette Jobert, Jean Philippe Dubus, Mariano Sigman, and Laurent Cohen. Hierarchical coding of letter strings in the ventral stream: dissecting the inner organization of the visual word-form system. *Neuron*, 55(1):143–156, 2007. 2.1, 2.3.3

Hrishikesh D Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4(2):147–166, 1976. 5.2.2

Patrik Vuilleumier, Jorge L Armony, Jon Driver, and Raymond J Dolan. Effects of attention and emotion on face processing in the human brain: an event-related fmri study. *Neuron*, 30(3): 829–841, 2001. 4.1

Katherine Vytal and Stephan Hamann. Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of Cognitive Neuroscience*, 22(12): 2864–2885, 2010. 4.3.6

Richard K Wagner, Joseph K Torgesen, and Carol Alexander Rashotte. *Comprehensive test of phonological processing: CTOPP*. Pro-ed, 1999. 2.5, 2.5, 2.10

Brian A Wandell. The neurobiological basis of seeing words. *Annals of the New York Academy of Sciences*, 1224(1):63–80, 2011. 1.1, 2.3.1

Yida Wang, Jonathan D Cohen, Kai Li, and Nicholas B Turk-Browne. Full correlation matrix analysis (fcma): An unbiased method for task-related functional connectivity. *Journal of Neuroscience Methods*, 251:108–119, 2015. 1.2.2, 5.1, 5.4.3

Zuoguan Wang, Siwei Lyu, Gerwin Schalk, and Qiang Ji. Deep feature learning using target priors with applications in ecog signal decoding for bci. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 1785–1791, 2013. 7.2

Elizabeth K Warrington and TIM Shallice. Word-form dyslexia. *Brain*, 103(1):99–112, 1980. 2.1

Kevin S Weiner and Kalanit Grill-Spector. Sparsely-distributed organization of face and limb activations in human ventral temporal cortex. *NeuroImage*, 52(4):1559–1573, 2010. 5.2.6

Kevin S Weiner, Golijeh Golarai, Julian Caspers, Miguel R Chuapoco, Hartmut Mohlberg, Karl Zilles, Katrin Amunts, and Kalanit Grill-Spector. The mid-fusiform sulcus: a landmark identifying both cytoarchitectonic and functional divisions of human ventral temporal cortex. *NeuroImage*, 84:453–465, 2014. 3.1, 4.3.3, 4.4

Kevin S Weiner, Michael A Barnett, Simon Lorenz, Julian Caspers, Anthony Stigliani, Katrin Amunts, Karl Zilles, Bruce Fischl, and Kalanit Grill-Spector. The cytoarchitecture of domain-specific regions in human high-level visual cortex. *Cerebral Cortex*, 27(1):146–161, 2017. 4.4

Carl Wernicke. Der aphasischer symptomenkomplex: eine psychologische studie auf anatomischer basis. In Gertrude H Eggert, editor, *Wernicke's Works on Aphasia: A Sourcebook and Review*. Mouton de Gruyter, 1977. 2.1, 2.4

Paul J Whalen, Scott L Rauch, Nancy L Etcoff, Sean C McInerney, Michael B Lee, and Michael A Jenike. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience*, 18(1):411–418, 1998. 4.1

Meagan Lee Whaley, Cihan Mehmet Kadipasaoglu, Steven James Cox, and Nitin Tandon. Modulation of orthographic decoding by frontal cortex. *Journal of Neuroscience*, 36(4):1173–1184, 2016. 2.3.1, 2.4

Michael S Worden, John J Foxe, Norman Wang, and Gregory V Simpson. Anticipatory biasing of visuospatial attention indexed by retinotopically specific-band electroencephalography increases over occipital cortex. *Journal of Neuroscience*, 20(RC63):1–6, 2000. 6.4.4

Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013. 3.2.9, 4.2.10, 4.3.6

Xiaokun Xu and Irving Biederman. Loci of the release from fmri adaptation for changes in facial expression, identity, and viewpoint. *Journal of Vision*, 10(14):36–36, 2010. 4.1

Gui Xue and Russell A Poldrack. The neural substrates of visual perceptual learning of words: implications for the visual word form area hypothesis. *Journal of Cognitive Neuroscience*, 19 (10):1643–1655, 2007. 2.1

Daniel L Yamins, Ha Hong, Charles Cadieu, and James J DiCarlo. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3093–3101, 2013. 5.4.4, 7.2

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. 7.2

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. 1.1, 5.4.2, 7.2

Malcolm P Young and Shigeru Yamane. Sparse population coding of faces in the inferotemporal cortex. *Science*, 256(5061):1327–1331, 1992. 2.4, 3.4

Hui Zhang, Shruti Japee, Rachel Nolan, Carlton Chu, Ning Liu, and Leslie G Ungerleider. Face-selective regions differ in their ability to classify facial expressions. *NeuroImage*, 130:77–90, 2016. 4.1, 4.4

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of*

the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005. 6.2.4