Title
Predictive models of auditory perception in human electrophysiology

Author
Holdgraf, Christopher

Publication Date
2017

Peer reviewed|Thesis/dissertation

Predictive models of auditory perception in human electrophysiology

By

Christopher R. Holdgraf

A dissertation submitted in partial satisfaction of

the requirement for the degree of

Doctor of Philosophy

in

Neuroscience

and the Designated Emphasis

in

Computational and Data Science and Engineering

In the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Robert T. Knight, Chair

Professor Frédéric E. Theunissen

Professor Tom Griffiths

Professor Andrew Szeri

Summer 2017

Abstract

Encoding and Decoding models of speech in human electrophysiology

by

Christopher R Holdgraf

Doctor of Philosophy in Neuroscience

Designated Emphasis in Computational and Data Science and Engineering

University of California, Berkeley

Professor Robert T. Knight, Chair

It has long been thought that sensory systems operate by representing information in a hierarchy of sensory features, and that these features build upon one another. From low-level information such as spectral content, to high-level information such as word content, the sensory system must rapidly extract all of these features from the world. However, the precise nature of these levels of representation, as well as how they interact with one another, is not well-understood. In audition, intermediate sensory representations are often studied in animals, using techniques that treat neurons as a linear filter for incoming sensory inputs. If those inputs are spectro-temporal features (e.g., a spectrogram), then the result is a Spectro Temporal Receptive Field (STRF). This describes how the neural unit in question (e.g., a neuron) will respond to patterns in spectro-temporal space. It has been a crucial tool in understanding sensory processing in low-level neural activity. Using this approach it is also possible to study how this neural representation changes under different experimental conditions. STRF plasticity has been shown in both reward- and context-modulated experiments in animals.

In recent years, it has been suggested that similar techniques may work in modeling the activity of neural signals recorded from humans. As we cannot generally record from single unit activity in humans, this approach relies on proxies for neural activity – specifically in the high-frequency activity (HFA) of electrocorticography electrodes. This poses a unique opportunity for two reasons: First, human language is a natural stimulus set for studying hierarchical feature representations in the brain. There are many ways to decompose speech into both auditory and linguistic components, and each of these could serve as inputs to the modeling technique described above. Second, humans are especially skilled at using high-level context such as their experience and assumptions about the world in order to change their behavior. This poses a unique opportunity to study the plasticity of speech representations in the brain.

This thesis reports several new approaches towards studying the sensory representation of speech in the human brain, as well as how these representations may change due to experience. It aims to bridge the literature in rodents and songbirds with ideas in human electrophysiology in order to pursue new approaches to studying perception in humans.

Dedication

Dedicated to my parents, Michael and Mary K. Holdgraf, who have taught me to question everything but to try and be reasonable about it.

Acknowledgements are difficult because they tend to focus on the parts rather than the whole. There are a number of people to whom I owe much gratitude over the past several years, but when I look back on graduate school it will be the whole picture that I remember most. In no particular order, these are some important parts of that picture.

First thanks go to my mother and my father, who continue to entertain my love of asking questions reflexively and the devil's advocate come hell or high water. You have taught me to be empathic and kind, and also rational and pragmatic. To avoid getting my head too far up in the clouds. To focus on the people around me. To remember that there are more important things than proving yourself right (even though it's pretty satisfying). I can't begin to describe how much your guidance, wisdom, patience, and unwavering support mean to me. I love you both.

Next, I wish to express my thanks to Claire Oldfield, who has been my partner in both life and science of the last several years. I'm lucky to be with someone who has been through the same academic world, and Claire has always been there to support me through challenging times in getting this degree. She continues to be a model of grace, calm, wittiness, prettiness, and intelligence, and I cherish our lives together. I love you Claire.

I also wish to thank all of my partners, mentors, and collaborators here at UC Berkeley. I still find it amazing that I landed in a PhD program here at all, and consider it one of the great fortunes of my life to have had this opportunity. For that I owe a deep debt of gratitude to Tom Griffiths and Tania Lombrozo, who gave me the opportunity to do research at Berkeley, and who were thoughtful and insightful mentors at the very beginning of my career. I also wish to thank Frederic Theunissen, who has become a second mentor to me as well as a productive and enjoyable collaborator. My thesis work, as well as my development as a computational scientist, would not have been possible without you, Frederic.

During my time here at UC Berkeley, I've had the opportunity to be involved with a number of projects off of the typical PhD track. I owe a debt to all of the teammates that I've worked with on these projects, including Sahar Yousef and the Beyond Academia team, Cathryn Carson and the data education initiative, Kevin Koy and the Berkeley Institute for Data Science, Greg Wilson and the Software Carpentry project, Fernando Perez and the Jupyter team, and finally the broader open science community at Berkeley.

Finally, I want to thank Bob Knight for being a wonderful advisor. You have always made the time and effort to guide me through my PhD, and to give me advice as I figure out what comes next. I have no doubt that you'd be willing to do anything to help your students, which I am truly grateful for. And I couldn't mention Bob without mentioning our extended family – the Knight Lab. We are a rag-tag bunch of people with a collection of interests as diverse as our nationalities. This has led to wonderful conversations, science, and friendships, and I'll never forget my first scientific family.

The studies in this dissertation were supported in part by the NDSEG Graduate Fellowship, the Berkeley Institute for Data Science, and the Nielsen Corporation

Chapter 1 – Introduction and background

While the act of understanding speech seems simple, it is the product of an extremely complex system in the brain. From the initial vibrations on the ear's tympanic membrane, one must infer a rich collection of information such as speaker identity, location, emotion, and semantic content. In addition, the auditory information that we receive from the world is often noisy, making it extremely difficult to extract features important for comprehension. The process of speech perception has been likened to guessing the class of a boat simply by observing the waves of water as they lap upon the shore.

There are many ways to study how sound is represented in the brain. From single-unit recordings in songbirds to electrical signals recorded from the human brain, these approaches provide different advantages in studying this complex process. This document represents an attempt to join two of these fields. It adapts the machinery of receptive field modeling – a technique that has a long history of studying the tuning properties of neurons in animal cortex – with speech processing in the *human* brain. Establishing methodological connections between receptive field modeling in animals and human speech processing opens new avenues for asking questions about sensory representation in the human brain and allows the knowledge gained in the animal literature to inform new experiments in humans.

Hierarchical representations of auditory information

Auditory pathways

While speech comprehension seems to occur effortlessly and automatically, it presents a demanding computational problem. Sound enters through the ears, causing a vibration on the tympanic membrane, a tiny sheet of tissue that transforms the vibrations in the air into firing patterns of the nearly 30,000 nerve fibers in the cochlear nerve. This signal is then passed through the superior olive and then the inferior colliculus of the midbrain, through the medial geniculate nucleus of the thalamus, and ultimately to auditory cortex. Remarkably, all of this processing occurs in 10-40 milliseconds (Nelken, Chechik, Mrsic-Flogel, King, & Schnupp, 2005).

At each step of this pathway, the sound is transformed. For example, the cochlea decomposes the incoming air vibrations into a collection of sine waves. To the extent that vibrations in the sound occur with a particular frequency component, this information will be relayed out of the cochlear nerve and into the midbrain. At the next step, inputs are in the form of frequency components of sound (the spectrogram), and further processing is performed on this representation.

As this signal passes through the auditory pathway, more complex features are built on top of simpler ones, and then passed along to the next step. As such, the auditory pathway is said to be hierarchically organized and can be thought of as a sequence of feature extraction steps, with increasingly complex acoustic features extracted at each stage of neural processing (Eggermont, 2001; Theunissen & Elie, 2014). This cascade of activity allows for complex features to iteratively build upon one another, making it possible to create rich representations of the acoustic world using the relatively impoverished information coming in through the ears.

Attempts at studying auditory representations

A key question in perceptual neuroscience asks what particular form describes these intermediate representations of sound. Presumably the brain performs complex, linear and non-linear transformations on the original raw sound wave, and it is possible to probe the nature of this representation at various parts of the auditory pathway. Some studies have investigated this by creating stimuli that vary along one particular dimension of interest. For example, words vs. nonsense words (Vouloumanos, Kiehl, Werker, & Liddle, 2001). The brain activity between these two types of stimuli is contrasted, and any differences are attributed to the stimulus manipulation (in this case, whether a word has semantic meaning).

While these attempts are important in that they describe the raw auditory stimuli with more complex features, they suffer from many drawbacks. Performing contrast-based studies often requires using auditory stimuli that are abnormal and non-representative of the sounds that are often heard in day-to-day life. They also often require presenting these stimuli in a manner that is not reflective of our natural experience with speech (e.g., stimuli are often required to have discrete onsets and offsets). In addition, the auditory system is adapted to extract these hierarchical features from sound, and studying this system by using stimuli that differ from the sounds one hears in everyday life means that we are observing the system *outside* of its core functionality. It is important to consider the environment in which the nervous system has evolved to function because studying a neural system in an atypical environment might lead to results that do not generalize to more natural conditions (Theunissen & Elie, 2014).

In order to study the auditory system under more natural conditions, researchers have begun to use recordings from the natural world. Because these stimuli are continuously-varying and are not easily split into groups for the purposes of a contrast analysis, traditional contrast-based tests are no longer appropriate. In this case, the statistical technique of regression makes it possible to describe the stimulus as a collection of continuously-varying features, and then to subsequently determine which of those features tends to drive increases in neural activity. A complex stimulus can then be decomposed into many collections of features – perhaps representing different levels of the auditory pathway – making it possible to simultaneously probe the hierarchical nature of stimulus processing at many levels.

Following this line of research, we now have a more nuanced picture of the hierarchical representation of auditory information. For example, at the level of auditory cortex, sounds are decomposed not only in frequency channels (as in the auditory periphery) but also in terms of joint spectral and temporal modulations. The filters in this modulation filter bank are referred to as the neurons' spectro-temporal receptive field or STRF (Depireux, Simon, Klein, & Shamma, 2001; L. M. Miller, Escabí, Read, & Schreiner, 2002; Theunissen, Sen, & Doupe, 2000). This decomposition of sounds into a modulation filter bank facilitates many tasks, including the discrimination of speech from non-speech(Mesgarani, Slaney, & Shamma, 2006) and the extraction of communication signals from noise (Moore, Lee, & Theunissen, 2013).

Interactions between hierarchical information

We have thus far characterized perception in a one-directional manner. Information entering through the ears is decomposed into a collection of auditory features, and then propagates throughout the auditory pathway as more complex features are extracted. However, this implies that the way in which the brain extracts features is fixed. In reality we know that this is likely not true, as one of the primary features of any neural system (especially those in the cortex) is plasticity.

Single neuron plasticity improves signal to noise

There is much research suggesting that auditory neurons can change their behavior under different conditions. A simple example is gain control, in which a neuron will gradually adapt to the amplitude of an input stimulus such that a stimulus that once elicited a large response will now elicit a smaller one. This has been shown in a number of sensory systems, and is generally believed to improve the signal-to-noise ratio of the neural system, allowing the neuron to extract the relative changes in stimulus features rather than being overpowered by the each feature's absolute value (Rabinowitz, Willmore, Schnupp, & King, 2011). Other studies have shown that this ability is useful in extracting a signal of interest from a noisy background (Ding & Simon, 2013).

Other types of neuronal plasticity are due to the degree of high-level or contextual information present. For example, it has been shown that high-level object activations will change the firing patterns of neurons in V1 (Gilbert & Sigman, 2007). This behavior suggests that the representation of information in neural systems is dynamic and dependent on context. Put simply, high-level information (e.g. context) may influence the way in which low-level information is processed. This capacity is often referred to as "top-down", rather than bottom-up.

Receptive field plasticity

Several studies have examined STRF-based feature representations at different levels of the auditory hierarchy (Atencio, Sharpee, & Schreiner, 2012; L. M. Miller et al., 2002; Woolley, Fremouw, Hsu, & Theunissen, 2005). However it is not understood if and how these representations interact with each other. For example, the presence of a higher-level response (such as the recognition of task-relevant stimuli) may alter the way that stimulus features are represented at lower levels in the auditory processing stream (Gilbert & Sigman, 2007). It has been shown that the tuning of auditory neurons changes during behavioral tasks (Fritz, Shamma, Elhilali, & Klein, 2003; Rabinowitz, Willmore, King, & Schnupp, 2013; Rabinowitz et al., 2011; Shamma & Fritz, 2014), revealing that the STRFs describing this tuning are plastic. Further, neuroanatomical (Atiani et al., 2014; Coull, Frith, Büchel, & Nobre, 2000; Davis & Johnsrude, 2007) and neurophysiological (David, Fritz, & Shamma, 2012; Yin, Fritz, & Shamma, 2014) research has highlighted the importance of top-down mechanisms for inducing this task-dependent STRF plasticity.

This research suggests that plasticity can be found at the level of intermediate feature representations, and that this may be modulated by top-down processes in the brain. However,

these studies are largely performed in single units recorded from animals. While this represents a crucial first step in understanding the representation of auditory information in the brain, humans offer an intriguing possibility for studying hierarchical speech processing that may relate to the unique speech capacity of those reading this thesis.

Encoding and decoding models in humans

Receptive field modeling in humans

Studying the hierarchical representation of information in humans has the natural benefit that we have an intuition for how language information is hierarchically organized. A rich history of linguistics gives us many different tools for decomposing a spoken sentence into linguistic features (both low- and high-level). Moreover, it is relatively simple to create an experiment in which high-level features are experimentally manipulated: natural language does this all the time.

While the animal literature has a rich history of using encoding models to study perception, it has taken time for these methods to be adopted to the study of the human auditory system. This is partially a problem of signal-to-noise – the methods we have for recording human brain activity are significantly noisier than a single-unit recording. However, advances in our ability to record brain activity in humans, as well as improvements in our understanding of these statistical techniques and efforts to reduce noise in the signal, make it possible to use the same regression framework for studying natural speech processing in humans.

In the last decade, we have seen the emergence of a new technique for recording brain activity in awake humans. Called *electrocorticography* (ECoG), this technique uses electrodes that are placed directly on the surface of the cortex in awake humans. While this is done for clinical purposes (usually in the treatment of epilepsy), it allows for a wide range of experiments to be conducted. Putting electrodes on the cortical surface (as opposed to the scalp, as in *electroencephalography*) has the benefit of increasing the signal-to-noise of the neural signal. Importantly, this makes it possible to resolve power differences in the "high-frequency" region of the spectral content of a signal, usually between 70 and 150Hz. This high frequency activity has been shown to reflect neural firing of neurons near the electrode with relatively high spatial (millimeter) and temporal (millisecond) resolution (Ray & Maunsell, 2011).

Recent research has shown that STRF modeling may be applied to human ECoG to characterize the spectrotemporal tuning of electrodes in response to speech (Hullett, Hamilton, Mesgarani, Schreiner, & Chang, 2016; Martin et al., 2014; Pasley et al., 2012) and to investigate plasticity in the auditory cortical response (Mesgarani & Chang, 2012). In particular, the high-frequency broadband (HFB; 70-150 Hz) neural activity recorded with ECoG has both the spatial resolution to localize activity to discrete regions of the brain, and the temporal resolution to resolve the fine-grained spatio-temporal pattern of acoustic features in the human cortex. HFB is believed to reflect local cortical activity typically obtained with 4-10 mm electrode spacing (Wodlinger, Degenhart, Collinger, Tyler-Kabara, & Wang, 2011). HFB activity represent a broadband increase in power, most readily detected in frequencies centered from 70-150Hz (K. J. Miller, Zanos, Fetz, den Nijs, & Ojemann, 2009) and also provides a metric of local cortical single unit

activity (Ray, Crone, Niebur, Franaszczuk, & Hsiao, 2008). This permits using HFB activity obtained with ECoG recording to study the representation of spectro-temporal speech features in human auditory cortex and investigate how this representation changes during language processing.

Context and speech plasticity in humans

Human speech perception is an area in which top-down and bottom-up mechanisms are in constant interplay (Block & Siegel, 2013; Cusack, Deeks, Aikman, & Carlyon, 2004; Schroeder, Wilson, Radman, Scharfman, & Lakatos, 2010). The act of understanding speech requires that auditory information entering the auditory periphery is interpreted through the lens of previous experience with natural sounds and language. It is assumed that this experience plays a role in shaping the response to speech in the cortex. Recent research using human electrophysiology has shown that experience with sound or contextual information about its content corresponds to differing patterns of low-frequency activity in both auditory and premotor cortex. For example, activity in the theta band (4-8 Hz) of neural signals is reported to track the temporal structure in the speech envelope (Fontolan, Morillon, Liegeois-Chauvel, & Giraud, 2014; Giraud & Poeppel, 2012; Gross et al., 2013) and this tracking increases as noise levels are decreased in the speech stimulus (Peelle, Gross, & Davis, 2013). In addition, power in theta and beta (12-20 Hz) frequency bands has been implicated in top-down processing during speech perception (Fontolan et al., 2014). It has been suggested that these signals reflect the brain's attempt to extract relevant information in the speech signal, and to filter out noise or competing auditory streams (Lakatos et al., 2013). While these approaches delineate differing patterns of neural activity that reflect top-down processes, they do not quantify changes in the spectro-temporal tuning of cortical activity, a feature representation that is believed to be encoded in auditory cortical neurons.

This thesis describes recent attempts to utilize regression modeling in studying the sensory representation of speech in humans, as well as to investigate how this representation changes due to one's experience with the world. The document is organized in the following sections: Chapter 2 discusses the methodology around using predictive models (regression and classification) for studying relationships between sensory features and brain activity. Chapter 3 describes using this regression framework to predict the underlying linguistic content contained within recorded neural activity. Chapter 4 focuses on the topic of sensory plasticity, using receptive field models in human auditory cortex to study how experience allows us to perceive degraded speech. Finally, Chapter 5 discusses open questions and next steps in this line of questioning.

Chapter 2 – Methods for predictive modeling in cognitive electrophysiology

Introduction

While the modeling techniques of encoding and decoding have been used extensively in animal models of the brain, they have only recently been used to model neural activity in humans. This chapter is a comprehensive overview of the considerations and details necessary in order to effectively construct encoding and decoding models of the human brain. It serves as a methodological foundation for the remaining chapters, and a practical guide for researchers interested in pursuing this line of research.

Citation

Holdgraf, C.R., Martin, S., Micheli, C., Knight, R.T., Rieger, J., Theunissen, F.E. (2017). Encoding and decoding models in cognitive electrophysiology. Front. in Systems Neuroscience. In submission.

Abstract

Cognitive neuroscience has seen rapid growth in the size and complexity of data recorded from the human brain as well as in the computational tools available to analyze this data. This data explosion has resulted in an increased use of multivariate, model-based methods for asking neuroscience questions, allowing scientists to investigate multiple hypotheses with a single dataset, to use complex, time-varying stimuli, and to study the human brain under more naturalistic conditions. These tools come in the form of "Encoding" models, in which stimulus features are used to model brain activity, and "Decoding" models, in which neural features are used to generated a stimulus output. Here we review the current state of encoding and decoding models in cognitive electrophysiology and provide a practical guide towards conducting experiments and analyses in this emerging field. Our examples focus on using linear models in the study of human language and audition. We show how to calculate auditory receptive fields from natural sounds as well as how to decode neural recordings to predict speech. The paper aims to be a useful tutorial to these approaches, and a practical introduction to using machine learning and applied statistics to build models of neural activity. The data analytic approaches we discuss may also be applied to other sensory modalities, motor systems, and cognitive systems, and we cover some examples in these areas. In addition, a collection of *Jupyter* notebooks is publicly available as a complement to the material covered in this paper, providing code examples and tutorials for predictive modeling in python. The aim is to provide a practical understanding of predictive modeling of human brain data and to propose best-practices in conducting these analyses.

Background

A fundamental goal of sensory neuroscience is linking patterns of sensory inputs from the world to patterns of signals in the brain, and to relate those sensory neural representations to perception. Widely used feedforward models assume that neural processing for perception utilizes a hierarchy of stimulus representations in which more abstract stimulus features are extracted from lower-level representations, and passed along to subsequent steps in the neural processing pipeline. Much of perceptual neuroscience attempts to uncover intermediate stimulus representations in the brain and to determine how more complex representations can arise from these levels of representation. For example, human speech enters the ears as air pressure waveform, but these are quickly transformed into a set of narrow band neural signals centered on the best frequency of auditory nerve fibers. From these narrow-band filters arise a set of spectro-temporal features characterized by the spectro-temporal receptive fields (STRFs) of auditory neurons in the inferior colliculus, thalamus, and primary auditory cortex (Eggermont, 2001). STRFs refer to the patterns of stimulus power across spectral frequency and time (spectro-temporal features). Complex patterns of spectro-temporal features can be used to detect phonemes, and ultimately abstract semantic concepts (DeWitt & Rauschecker, 2012; Poeppel, Emmorey, Hickok, & Pylkkänen, 2012). It should also be noted that there are considerable feedback pathways that may influence this process (Fritz, Shamma, Elhilali, & Klein, 2003; Yin, Fritz, & Shamma, 2014).

Cognitive neuroscience has traditionally studied hierarchical brain responses by crafting stimuli that differ along a single dimension of interest (e.g., high- vs. low-frequency, or words vs nonsense words). This method dates back to Donders, who introduced mental chronometry to psychological research (1969, orig 1868). Donders suggested crafting tasks such that they differ in exactly one cognitive process to isolate the differential mental cost of two processes. Following Donders, the researcher contrasts the averaged brain activity evoked by two sets of stimuli assuming that the neural response to these two stimuli/tasks is well-characterized by averaging out the trial-to-trial variability (Pulvermüller, Lutzenberger, & Preissl, 1999). One then performs inferential statistical testing to assess whether the two mean activations differ. While much has been learned about perception using these methods, they have intrinsic shortcomings. Using tightly-controlled stimuli focuses the experiment and its interpretation on a restricted set of questions, inherently limiting the independent variables one may investigate with a single task. This approach is time-consuming, often requiring separate stimuli or experiments in order to study many feature representations and may cause investigators to miss important brain-behavior findings. Moreover, it can lead to artificial task designs in which the experimental manipulation renders the stimulus unlike those encountered in everyday life. For example, contrasting brain activity between two types of stimuli requires many trials with a discrete stimulus onset and offset (e.g. segmented speech) so that evoked neural activity can be calculated, though natural auditory stimuli (e.g. conversational speech) rarely come in this time-segregated manner (Felsen & Dan, 2005; Theunissen & Elie, 2014). In addition, this approach requires a priori hypotheses about the architecture of the cognitive processes in the brain to guide the experimental design. Since these hypotheses are often based on simplified experiments, the results do not readily transfer to more realistic everyday situations.

There has been an increase in techniques that use computationally-heavy analysis in order to increase the complexity or scope of questions that researchers may ask. For example, in cognitive neuroscience the "Multi-voxel pattern analysis" (MVPA) framework utilizes a machine learning technique known as classification to detect condition-dependent differences in patterns of activity across multiple voxels in the fMRI scan (usually within a Region of Interest, or ROI: Hanke et al., 2009; Norman, Polyn, Detre, & Haxby, 2006; Varoquaux et al., 2016). MVPA has proven useful in expanding the sensitivity and flexibility of methods for detecting condition-based differences in brain activity. However, it is generally used in conjunction with single-condition based block design that is common in cognitive neuroscience.

An alternative approach studies sensory processes using multivariate methods that allow the researcher to study multiple feature representations using complex, naturalistic stimuli. This approach entails modeling the activity of a neural signal while presenting stimuli varying along multiple continuous stimulus features as seen in the natural world. In this sense, it can be seen as an extension of the MVPA approach that utilizes complex stimuli and provides a more direct model of the relationship between stimulus features and neural activity. Using statistical methods such as regression, one may create an optimal model that represents the combination of elementary stimulus features that are present in the activity of the recorded neural signal. These techniques have become more tractable in recent years with the increase in computing power and the improvement of methods to extract statistical models from empirical data. The benefits over a traditional stimulus-contrast approach include the ability to make predictions about new datasets (Nishimoto et al., 2011), to take a multivariate approach to fitting model weights (Huth, Nishimoto, Vu, & Gallant, 2012), and to use multiple feature representations within a single, complex stimulus set (Di Liberto, O'Sullivan, & Lalor, 2015; Hullett, Hamilton, Mesgarani, Schreiner, & Chang, 2016).

These models come in two complementary flavors. The first are called "encoding" models, in which stimulus features are used to predict patterns of brain activity. Encoding models have grown in popularity in fMRI (Naselaris, Kay, Nishimoto, & Gallant, 2011), electrocorticography (Mesgarani, Cheung, Johnson, & Chang, 2014), and EEG/MEG (Di Liberto et al., 2015). The second are called "decoding" models, which predict stimulus features using patterns of brain activity (Martin et al., 2014; Mesgarani & Chang, 2012; Pasley et al., 2012). Note that in the case of decoding, "stimulus features" does not necessarily mean a sensory stimulus – it could be an experimental condition or an internal state, though in this paper we use the term "stimulus" or "stimulus features". Both "encoding" and "decoding" approaches fall under the general approach of predictive modeling, and can often be represented mathematically as either a regression or classification problem.

We begin with a general description of predictive modeling and how it has been used to answer questions about the brain. Next we discuss the major steps in using predictive models to ask questions about the brain, including practical considerations for both encoding and decoding and associated experimental design and stimulus choice considerations. We then highlight areas of research that have proven to be particularly insightful, with the goal of guiding the reader to better understand and implement these tools for testing particular hypotheses in cognitive neuroscience. To facilitate using these methods, we have included a small sample

dataset, along with several scripts in the form of *jupyter* notebooks that illustrate how one may construct predictive models of the brain with widely-used packages in Python. These techniques can be run interactively in the cloud as a GitHub repository[1].

The predictive modeling framework.

Predictive models allow one to study the relationship between brain activity and combinations of stimulus features using complex, often naturalistic stimulus sets. They have been described with varying terminology and approaches (Santoro et al., 2014; Wu, David, & Gallant, 2006; Yamins & DiCarlo, 2016), but generally involve the following steps which are outlined below (see **Figure 1**).

1. **Input feature extraction**: In an encoding model, features of a stimulus (or experimental condition) are used as inputs. These features are computed or derived from "real world" parameters describing the stimulus (e.g. sound pressure waveform in auditory stimuli, contrast at each pixel in visual stimuli). The choice of input features is a key step in the analysis: features must be adapted to the level in the sensory processing stream being studied and multiple feature-spaces can be tried to test different hypotheses. This is generally paired with the assumption that the neural representation of stimulus features becomes increasingly non-linear as one moves along the sensory pathway. For example, if one is fitting a linear model, a feature space based on the raw sound pressure waveform could be used to predict the responses of auditory nerve fibers (Kiang, 1984), but would perform significantly worse in predicting activity of neurons in the inferior colliculus (Andoni & Pollak, 2011) or for ECoG signals recorded from auditory cortex (Pasley et al., 2012). This is because the neural representation of the stimulus is rapidly transformed such that neural activity no longer has a linear relationship with the original raw signal. While a linear model may capture some of this relationship, it will be a poor approximation of the more complex stimulus-response function. At the level of secondary auditory areas, the prediction obtained from higher-level features such as word representations could be contrasted to that based on spectral features (as the alternative feature space) to test the hypothesis that these higher-level features (words) are particularly well-represented in this brain region (de Heer, Huth, Griffiths, Gallant, & Theunissen, 2017). Other examples of feature spaces for natural auditory signals are modulation frequencies (Mesgarani, Slaney, & Shamma, 2006; Pasley et al., 2012; Santoro et al., 2014), phonemes (Khalighinejad, Cruzatto da Silva, & Mesgarani, 2017; Mesgarani et al., 2014), or words (Huth, Heer, Griffiths, Theunissen, & Gallant, 2016; Huth et al., 2012). For stimulus features that are not continuously-varying, but are either "present" or not, one uses a binary vector indicating that feature's state at each moment in time. It may also be possible to combine multiple feature representations with a single model, though care must be

[1] https://github.com/choldgraf/paper-encoding_decoding_electrophysiology

taken account for the increased complexity of the model and for dependencies between features (de Heer et al., 2017; Lescroart, Stansbury, & Gallant, 2015).

2. **Output feature extraction:** Similarly, a representation of the neural signal is chosen as an output of the encoding model. This output feature is often a derivation of the "raw" signal recorded from the brain, such as amplitude in a frequency band of the time-varying voltage of an ECoG signal (Holdgraf et al., 2016; Mesgarani et al., 2014; Pasley et al., 2012), pixel intensity in fMRI (Naselaris et al., 2011), and spike rates in a given window or spike patterns from single unit recordings (Fritz et al., 2003; Theunissen & Elie, 2014). Choosing a particular *region* of the brain from which to record can also be considered a kind of "feature selection" step. In either case, the choice of features underlies assumptions about how information is represented in the neural responses. In combination with the choice of derivations of the raw signal to use, as well as which brain regions to use in the modeling process, the predictive framework approach can be used to test how and where a given stimulus feature is represented. For example, the assumption that sensory representations are hierarchically organized in the brain (Felleman & Van Essen, 1991) can be tested directly.
3. **Model architecture and estimation:** A model is chosen to map input stimulus features to patterns of activity in a neural signal. The structure and complexity of the model will determine the kind of relationships that can be represented between input and output features. For example, a linear encoding model can only find a linear relationship between input feature values and brain activity, and as such it is necessary to choose features that are carefully selected. A non-linear model may be able to uncover a more complex relationship between the raw stimulus and the brain activity, though it may be more difficult to interpret, will require more data, and still may not adequately capture the actual non-linear relationship between inputs and outputs (Ahrens, Paninski, & Sahani, 2008; Eggermont, Johannesma, & Aertsen, 1983; Paninski, 2003; Sahani & Linden, 2003). In cognitive neuroscience it is common to use a linear model architecture in which outputs are a weighted sum of input features. Non-linear relationships between the brain and the raw stimulus are explicitly incorporated into the model in the choice of input and output feature representations (e.g., performing a Gabor wavelet decomposition followed by calculating the envelope of each output is a non-linear expansion of the input signal). Once the inputs / outputs as well we the model architecture have been specified, the model is fit (in the linear case, the input weights are calculated) by minimizing a metric of error between the model prediction and the data used to fit the model. The metric of error can be rigorously determined based on statistical theory (such as maximum likelihood) and a probability model for the non-deterministic fraction of the response (the noise). For example, if one assumes the response noise is normally distributed, a maximum likelihood approach yields the sum of squared errors as an error metric. Various analytical and numerical methods are then used to minimize the error metric and, by doing so, estimate the model parameters (Hastie, Tibshirani, & Friedman, 2009; Naselaris et al., 2011; Wu et al., 2006).
4. **Validation:** Once model parameters have been estimated, the model is validated with data which were not used in the fit: in order to draw conclusions from the model, it

must generalize to new data. This means that it must be able to predict new patterns of data that have never been used in the original model estimation. This may be done on a "held-out" set of data that was collected using the same experimental task, or on a new kind of task that is hypothesized to drive the neural system in a similar manner. In the case of regression with normally distributed noise, the variance explained by the model on cross-validated data can be compared to the variance that could be explained based on differences between single data trials and the average response across multiple repetitions of the same trial. This ratio fully quantifies the goodness of fit of the model. While this can be difficult to estimate, it allows one to calculate an "upper bound" on the expected model performance and can be used to more accurately gauge the quality of a model, see section What *is a "good" model score?* (Hsu, Borst, & Theunissen, 2004; Sahani & Linden, 2003).

5. **Inspection and Interpretation:** If an encoding model is able to predict novel patterns of data, then one may further inspect the model parameters to gain insight into the relationship between brain activity and stimulus features. In the case of linear models, model parameters have a relatively straightforward definition – each parameter's weight is the amount the output would be expected to change given a unit increase in that parameter's value. Model parameters can then be compared across brain regions or across subjects (Hullett et al., 2016; Huth et al., 2016). It is also possible to inspect models by assessing their ability to generalize their predictions to new kinds of data. See section 6, *Interpreting the model*.

This predictive modeling framework affords many benefits, making it possible to study brain activity in response to complex "natural" stimuli, reducing the need for separate experiments for each stimulus feature of interest, and loosening the requirement that stimuli have clear-cut onsets and offsets. Moreover, naturalistic stimuli are better-matched to the sensory statistics of the environment in which the target organism of study has evolved, leading to more generalizable and behaviorally-relevant conclusions.

In addition, because a formal model describes a quantifiable means of transforming input values into output values, it can be "tested" in order to confirm that the relationship found between inputs / outputs generalizes to new data. Given a set of weights that have been previously fit to data, it is possible to calculate the "predictive power" for a given set of features and model weights. This is a reflection of the error in predictions of the model, that is, the difference between predicted outputs and actual outputs (also called the "prediction score").

While the underlying math is the same between encoding and decoding models when using regression, the interpretation and nature of model fitting differs between the two. The next section describes the unique properties of each approach to modeling neural activity.

Encoding models

Encoding models are useful for exploring multiple levels of abstraction within a complex stimulus, and investigating how each affects activity in the brain. For example, natural speech is a continuous stream of sound with a hierarchy of complex information embedded within it (Hickok & Small, 2015). A single speech utterance contains many representations of information, such as spectrotemporal features, phonemes, prosody, words, and semantics. The

neural signal is a continuous response to this input with multiple embedded streams of information in it due to recording the activity from many neurons spread across a relatively large region of cortex. The components of the neural signal operate on many timescales (e.g., responding to the slow fluctuations of the speech envelope vs. fast fluctuations of spectral content of speech (David & Shamma, 2013)) as information propagates throughout auditory cortex, and are not well-described by a single event-related response to a stimulus onset (Khalighinejad et al., 2017). Naturalistic stimuli pose a challenge for event-related analysis, but are naturally handled in a predictive modeling framework. In the predictive modeling approach, the solution takes the form of a linear regression problem. (Hastie et al., 2009)

$$activity(t) = \sum_{i}^{N_{features}} feature_i(t) * weight_i + error(t)$$

Where the neural activity at time t is modeled as a weighted sum of *N* stimulus features. Note that it becomes clear from this equation that features that have never been presented will not enter the model and contribute to the sum. Thus, both the choice of stimuli and input feature space are critical and have a strong influence on the interpretation of the encoding model. It is also common to include several time-lagged versions of each feature as well, accounting for the fact that the neural signal may respond to particular feature patterns in time. In this case, the model formulation becomes:

$$activity(t) = \sum_{j}^{N_{lags}} \sum_{i}^{N_{features}} feature_i(t-j) * weight_{i,j} + error(t)$$

In other words, this model describes how dynamic stimulus features are *encoded* into patterns of neural activity. It is convenient to write this in linear algebra terms:

$$\boldsymbol{activity} = \boldsymbol{Sw} + \boldsymbol{\epsilon}$$

In this case $\boldsymbol{S}$ is the stimulus matrix where each row corresponds to a timepoint of the response, and the columns are the feature values at that timepoint and time-lag (there are $N_{lags} * N_{features}$ columns). $\boldsymbol{w}$ is a vector of model weights (one for each feature * time lag), and $\boldsymbol{\epsilon}$ is a vector of random noise at each timepoint (most often to be Gaussian for continuous signals or Poisson for discrete signals). The observed output activity can then be written as a single dot product assumed between feature values and their weights plus additive noise. This dot product operation is identical to explicitly looping over features and time lags separately (each "iteration" over lag/feature combinations becomes a column in $\boldsymbol{S}$ and an single value in $\boldsymbol{w}$, thus the dot-product achieves the same result).

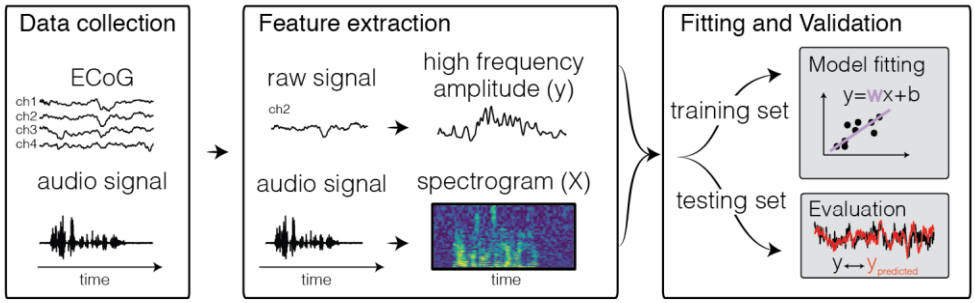


Figure 1 – Predictive modeling overview. *The general framework of predictive models consists of three steps. First, input and output data are collected, for example during passive listening to natural speech sentences. Next, features are extracted; Traditional features for the neural activity can be the time-varying power of various frequency bins, such as high frequency range (70-150 Hz, shown above). For auditory stimuli, the audio envelope or spectrogram are often used. Finally, the data are split into a training and test set. The training set is used to fit the model, and the test set is used to calculate predictive score of the model.*

As mentioned above, the details of neural activity under study (the output features), as well as the input features used to predict that activity, can be flexibly changed, often using the same experimental data. In this manner, one may construct and test many hypotheses about the kinds of features that elicit brain activity. For example to explore the neural response to spectro-temporal features, one may use a spectrogram of audio as input to the model (Eggermont et al., 1983; Sen, Theunissen, & Doupe, 2001). To explore the relationship between the overall energy of the incoming auditory signal (regardless of spectral content) and neural activity, one may probe the correlation between neural activity and the speech envelope (Zion Golumbic et al., 2013). To explore the response to speech features such as phonemes, audio may be converted into a collection of binary phoneme features, with each feature representing the presence of a single phoneme (de Heer et al., 2017; Leonard, Bouchard, Tang, & Chang, 2015). Each of these stimulus feature representations may predict activity in a different region of the brain. Researchers have also used non-linearities to explore different hypotheses about more complex relationships between inputs and neural activity, see section Choosing *a modeling framework*.

In summary, encoding models of sensory cortex attempt to model cortical activity as a function of stimulus features. These features may be complex and applied to “naturalistic” stimuli allowing one to study the brain under conditions observed in the real world. This provides a flexible framework for estimating the neural tuning to particular features, and assessing the quality of a feature set for predicting brain activity.

Decoding models

Conversely, decoding models allow the researcher to use brain activity to infer the stimulus and/or experimental properties that were most likely present at each moment in time.

$$feature(t) = \sum_{j}^{N_{lags}} \sum_{i}^{N_{channels}} activity_i(t+j) * weight_{i,j} + error(t)$$

which, in vector notation, is represented as the following:

$$\boldsymbol{s} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

where $\boldsymbol{s}$ is a vector of stimulus feature values recorded over time, and $\boldsymbol{X}$ is the channel activity matrix where each row is a timepoint and each column is a neural feature (with time-lags being treated as a separate column each). $\boldsymbol{w}$ is a vector of model weights (one for each neural feature * time lag), and $\boldsymbol{\epsilon}$ is a vector of random noise at each timepoint (often assumed to be Gaussian noise). Note that here the time lags are negative ("$+j$" in the equation above) reflecting the fact that neural activity in the present is being used to predict stimulus values in the past. This is known as an *acausal* relationship because the inputs to the model are not assumed to causally influence the outputs. If the model output corresponds to discrete event types (e.g. different phonemes), then the model is performing *classification*. If the output is a continuously-varying stimulus property such as the power in one frequency band of a spectrogram, the model performs regression and can be used, for example, in *stimulus reconstruction*.

In linear decoding, the weights can operate on a multi-dimensional neural signal, allowing the researcher to consider the joint activity across multiple channels (e.g. electrodes or voxels) around the same time (See Figure 3). By fitting a weight to each neural signal, it is possible to infer the stimulus or experiment properties that gave rise to the distributed patterns of neural activity.

The decoder is a proof of concept: given a new pattern of unlabeled brain activity (that is, brain activity *without* its corresponding stimulus properties), it may be possible to reconstruct the most likely stimulus value that resulted in the activity seen in the brain (Naselaris, Prenger, Kay, Oliver, & Gallant, 2009; Pasley et al., 2012). The ability to accurately reconstruct stimulus properties relies on recording signals from the brain that are tuned to a diverse set of stimulus features. If neural signals from multiple channels show a diverse set of tuning properties (and thus if they contain independent information about the stimulus), one may combine the activity of many such channels during decoding in order to increase the accuracy and diversity of decoded stimuli, provided that they carry independent information about the stimulus (Moreno-Bote et al., 2014).

Benefits of the predictive modeling framework

As discussed above, predictive modeling using multivariate analyses is one of many techniques used in studying the brain. While the relative merits of one analysis over another is not black and white, it is worth discussing specific pros and cons of the framework described in this paper. Below are a few key benefits of the predictive modeling approach.

1. **Generalize on test set data.** Classical statistical tests compare means of measured variables, and statements about significance are based on the error of the point estimates such as the standard error of the mean. When using predictive modeling, cross-validated models are tested for their ability to generalize to new data, and thus are judged against the variability of the population of measurements. As such, classical inferential testing

makes statements of statistical significance, while cross-validated encoding/decoding models make statements about the relevance of the model. This allows for more precise statements about the relationship between inputs and outputs. In addition, encoding models offer a continuous measure of model quality, which is a more subtle and complete description of the neural signal being modeled.

2. **Jointly consider many variables.** Many statistical analyses (e.g., Statistical Parametric Mapping fMRI analysis (Friston, 2003)) employ massive parallel univariate testing in which variables are first selected if they pass some threshold (e.g., activity in response to auditory stimuli), and subsequent statistical analyses are conducted on this subset of features. This can lead to inflated family-wise error rate and is prone to "double-dipping" if the thresholding is not carried out properly. The predictive modeling approach discussed here uses a multivariate analysis that jointly considers feature values, describing the relative contributions of features as a single weight vector. Because multiple parameters are estimated simultaneously the parameters patterns should be interpreted as a whole. This gives a more complex picture of feature interaction and relative importance, and also reduces the amount of statistical comparisons being made. However, note that it is also possible to perform statistical inference on individual model parameters.
3. **Generate hypotheses with complex stimuli.** Because predictive models can flexibly handle complex inputs and outputs, they can be used as an exploratory step in generating hypotheses about the representation of stimulus features at different regions of the brain. Using the same stimulus and neural activity, researchers can explore hypotheses of stimulus representation at multiple levels of stimulus complexity. This is useful for generating new hypotheses about sensory representation in the brain, which can be confirmed with follow-up experiments.
4. **Discover multivariate structure in the data.** Because predictive models consider input features jointly, they are able to uncover structure in the input features that may not be apparent when testing using univariate methods. For example, spectro-temporal receptive fields describe complex patterns in spectro-temporal space that are not apparent with univariate testing (see Figure 5). It should be noted that any statistical technique will give misleading results if the covariance between features is not taken into consideration, though it is more straightforward to consider feature covariance using the modeling approach described here.
5. **Model subtle time-varying detail in the data.** Traditional statistical approaches tend to collapse data over dimensions such as time (e.g., when calculating a per-trial average). With predictive modeling, it is straightforward to incorporate the relationship between inputs and outputs at each timepoint without treating between-trail variability as noise. This allows one to make statements about the time-varying relationship between inputs and outputs instead of focusing only on whether activity goes up or down on average. Researchers have used this in order to investigate more subtle changes in neural activity such as those driven by subjective perception and internal brain states (Chang et al., 2011; Reichert et al., 2014).

Ultimately, predictive modeling is not a replacement of traditional univariate methods, but should be seen as a complementary tool for asking questions about complex, multivariate

inputs and outputs. The following sections describe several types of stimuli and experimental setups that are well-suited for predictive modeling. They cover the general workflow in a predictive modeling framework analysis, as well as a consideration of the differences between regression and classification in the context of encoding and decoding.

Identifying Input / Output Features

The application of linear regression or classification models requires transforming the stimulus and the neural activity such that they have a linear relationship with one another. This follows the assumption that generally there is a non-linear relationship between measures of neural responses (e.g. spike rate) and those of the raw stimulus (e.g., air pressure fluctuations in the case of speech), but that the relationship becomes *linear* after some non-linear transformation of that raw stimulus (e.g., the speech envelope of the stimulus). The nature of this non-linear transformation is used to investigate what kind of information the neural signal carries about the stimulus. As such, when using the raw stimulus values, a linear model will not be able to accurately model the neural activity, but after a non-linear transformation that matches the transformations performed in the brain, the linear model is now able to explain variance in the neural signal. This is a process called *linearizing* the model (David, 2004; David & Gallant, 2005).

As the underlying math of linear models is straightforward, picking the right set of input / output features is a crucial tool for testing hypotheses. Stimulus linearization can be thought of as a process of *feature extraction / generation*. Features are generally chosen based on previous knowledge or assumptions about a brain region under study, and have been used to investigate the progression of simple to complex feature representations along the sensory pathway.

The following sections describe common feature representations that have been used for building linearized *encoding* and *decoding* models in cognitive electrophysiology. They reflect a restricted set of questions about stimulus transformations in the brain drawn from the literature and are not an exhaustive set of possible questions. Also note that it is possible to use other neural signals as inputs to an encoding model (for example, an autoregressive model uses past timepoints of the signal being predicted as input, which is useful for finding autocorrelations, repeating patterns, and functional connectivity metrics (Bressler & Seth, 2011)). However, this article focuses on external stimuli.

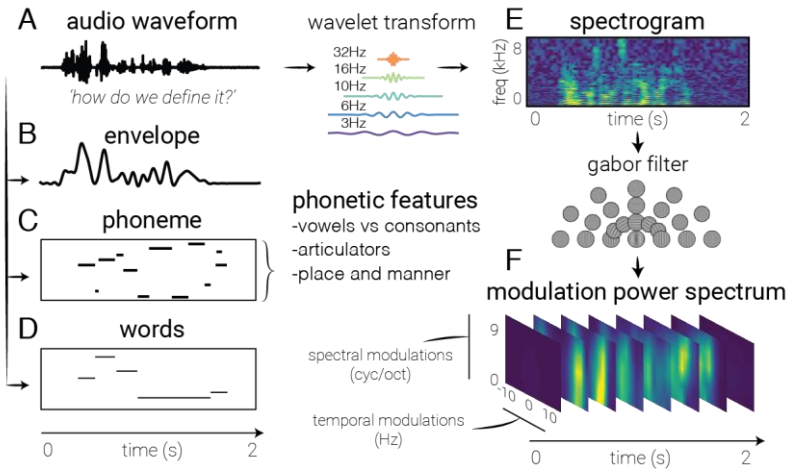


Figure 2 – Feature extraction. *Several auditory representations are shown for the same natural speech utterance. (A) Raw audio. Generally used as a starting point for feature extraction, rarely in linear models, though can be used with non-linear models (and sufficient amounts of data). (B) Speech envelope. The raw waveform can be rectified and low-pass filtered to extract the speech envelope, representing the amount of time-varying energy present in the speech utterance. (E) Spectrogram. A time-frequency decomposition of the raw auditory waveform can be used to generate a spectrogram that reflects spectro-temporal fluctuations over time, revealing spectro-temporal structure related to higher-level speech features. (F) Modulation Power Spectrum. A two-dimensional Gabor decomposition of the spectrogram itself can be used to create the MPS of the stimulus, which summarizes the presence or absence (i.e., power) of specific spectro-temporal fluctuations in the spectrogram. (C) Phonemes. In contrast to previous features which are defined acoustically, one may also use linguistic features to code the auditory stimulus, in this case with categorical variables corresponding to the presence of phonemes. (D) Words. Another higher-order feature that is not directly related to any one spectrotemporal pattern, these types of features may be used to investigate higher-level activity in the brain's response.*

Encoding models

Encoding models define model inputs by decomposing the raw stimulus (be it an image, an audio stream, etc.) into either well-defined high-level features with both a direct relationship with the physical world linked with a particular percept (e.g. spectrogram modulations, center frequencies, cepstral coefficients) or statistical descriptions of these features (e.g., principal or independent components). This is in contrast to a classic approach that builds receptive field maps using spectrograms of white noise used for stimulus generation. The classic approach works well for neural activity in low-level sensory cortex (Marmarelis & Marmarelis, 1978) but results in sub-optimal models for higher-level cortical areas, due in part to the fact that white noise contains no higher-level structure (David, 2004).

The study of sound coding in early auditory cortices commonly employs a windowed decomposition of the raw audio waveform to generate a spectrogram of sound – a description of the spectral content in the signal as it changes over time (see **Figure 2**). Using a spectrogram as input to a linear model has been used to create a *spectro-temporal receptive field* (STRF). This can be interpreted as a filter that describes the spectro-temporal properties of sound that elicit an increase in activity in the neural signal. The STRF is a feature representation used to study both single unit behavior (Aertsen & Johannesma, 1981; Depireux, Simon, Klein, &

Shamma, 2001; Escabi & Schreiner, 2002; Theunissen et al., 2001; Theunissen, Sen, & Doupe, 2000) and human electrophysiology signals (Di Liberto et al., 2015; Holdgraf et al., 2016; Hullett et al., 2016; Pasley & Knight, 2012).

It should be noted that spectrograms (or other time-frequency decompositions) are not the only way to represent auditory stimuli. Others researchers have used cepstral decompositions of the spectrogram (Hermansky & Morgan, 1994), which embed perceptual models within the definition of the stimuli features or have chosen stimulus feature representations that are thought to mimic the coding of sounds in the sensory periphery (Chi, Ru, & Shamma, 2005; Pasley et al., 2012). Just as sensory systems are believed to extract features of increasing abstraction as they continue up the sensory processing chain, researchers have used features of increasing complexity to model higher-order cortex (Sharpee, Atencio, & Schreiner, 2011). For example, while spectrograms are used to model early auditory cortices, researchers often perform a secondary non-linear decomposition on the spectrograms to implement hypothesized transformations implemented in the auditory hierarchy such as phonemic, lexical, or semantic information. These are examples of *linearizing* the relationship between brain activity and the stimulus representation.

In one approach, the energy modulations across both time and frequency are extracted from a speech spectrogram by using a filter bank of two-dimensional Gabor functions (see Sidenote on Gabors). This results extracts the Modulation Power Spectrum of the stimulus (in the context of receptive fields, also called the Modulation Transfer Function). This feature representation has been used to study higher-level regions in auditory cortex (Chi et al., 2005; Elliott & Theunissen, 2009; Pasley et al., 2012; Santoro et al., 2014; Theunissen et al., 2001). There have also been efforts to model brain activity using higher-order features that are not easily connected to low-level sensory features, such as semantic categories (Huth et al., 2016). This also opens opportunities for studying more abstract neural features such as the activity of a distributed network of neural signals.

Alternatively, one could create features that exploit the stimulus statistics, for example features that are made statistically independent from each other (Bell & Sejnowski, 1995) or by exploiting the concept of sparsity of stimulus representation bases (Olshausen & Field, 1997, 2004; Shelton, Sheikh, Bornschein, Sterne, & Lücke, 2015). Feature sparseness of can improve the predictive power and interpretability of models because the representation of stimulus features in active neural populations may be inherently sparse (Olshausen & Field, 2004). For example, researchers have used the concept of sparseness to learn model features from the stimuli set by means of an unsupervised approach that estimates the primitives related to the original stimuli (e.g. for vision: configurations of 2-D bars with different orientations). This approach is also known as “dictionary learning” and has been used to model the neural response to simple input features in neuroimaging data (Güçlü & van Gerven, 2014; Henniges & Puertas, 2010). It should be noted that more “data-driven” methods for feature extraction often discover features that are similar to those defined *a priori* by researchers. For example, Gabor functions have proven to be a useful way to describe both auditory (Lewicki, 2002) and visual (Touryan, Felsen, & Dan, 2005) structure, and are both commonly used in the neural modeling literature. In parallel, methods that attempt to define features using methods that

maximize between-feature statistical independence (such as Independent Components Analysis) also often discover features that look similar to Gabor wavelets (Olshausen & Field, 1997) (see *Sidenote on Gabors* for more detail [2]).

It is also possible to select different neural *output* features (e.g., power in a particular frequency band of the LFP) to ask different questions about neural activity. The choice of neural feature impacts the model's ability to predict patterns of activity, as well as the conclusions one may draw from interpreting the model's weights. For example, encoding models in electrocorticography are particularly useful because of "high-frequency" activity (70-200 Hz) that reflects local neural processing (Ray & Maunsell, 2011). This signal has a high signal-to-noise ratio, making it possible to fit models with more complicated features. Since it is tightly linked to ensembles of neurons, it is more straightforward to interpret how the stimulus features are encoded in the brain Hullett et al., 2016; Pasley et al. (2012) and to connect with the single-unit encoding literature (Theunissen & Elie, 2014). Researchers have also used more complex representations of neural activity to investigate the type of information they may encode. For example, in order to investigate the interaction between attention and multiple speech streams, (Zion Golumbic et al., 2013) computed a "temporal receptive field" of an auditory speech envelope for theta activity in ECoG subjects. A similar analysis has been performed with EEG (Di Liberto et al., 2015). It is also possible to describe patterns of distributed activity in neural signals (e.g., using Principle Components Analysis or network activity levels), and use this as the output being predicted (though this document treats each output (i.e. channel) as a single recording unit).

An important development in the field of linear encoding models is loosening of the assumptions of stationarity to treat the input/output relationship as a dynamic process (Meyer, Williamson, Linden, & Sahani, 2017). While a single model assumes stationarity in this relationship, fitting multiple models on different points in time or different experimental

---

[2] **SIDENOTE ON GABORS**: A Gabor function is a sinusoidal function windowed with a Gaussian density function (in either 1- or 2-D), and is commonly used to derive stimulus representations in both visual (Kay & Gallant, 2009; Lescroart, Kanwisher, & Golomb, 2016; Naselaris et al., 2009; Nishimoto et al., 2011), and auditory cortex (Qiu, Schreiner, & Escabí, 2003; Santoro et al., 2014; Theunissen et al., 2001). For example, it is possible to create a spectro-temporal representation of sounds by constructing a collection of Gabor wavelets with linearly- or logarithmically-increasing frequencies, filtering the raw sound with each one, then calculating the amplitude envelope of the output of each filter. If the nature of the stimulus is 2-D (e.g., an image, movie, or spectro-temporal representation), a collection of 2-D Gabor wavelets may be created with successive frequencies and orientations (Frye et al., 2016). Gabor functions may also be a particularly efficient means of storing stimulus information, and studies that use a sparse coding framework to model the way that neurons represent information often result in Gabor-like decompositions (Olshausen & Field, 1997).

conditions allows the researcher to make inferences about how (and why) the relationship between stimulus features and neural activity changes. For example, Fritz et al recorded activity in the primary auditory cortex of ferrets during a tone frequency detection task (Fritz, Elhilali, & Shamma, 2005). The authors showed that spectro-temporal receptive fields of neurons changed their tuning when the animal was actively attending to a frequency vs. passively listening to stimuli, suggesting that receptive fields are more plastic than classically assumed (Meyer, Diepenbrock, Ohl, & Anemüller, 2014). Further support for dynamic encoding is provided by *Holdgraf et al,* who implemented a task in which ECoG subjects listened to degraded speech sentences. A degraded speech sentence was played, followed by an "auditory context" sentence, and then the degraded speech was repeated. The context created a powerful behavioral "pop-out" effect whereby the degraded speech was rendered intelligible. The authors compared the STRF of electrodes in the auditory cortex in response to degraded speech *before* and *after* this context was given, and showed that it exhibited plasticity that was related to the perceptual "pop-out" effect (Holdgraf et al., 2016). Our understanding of the dynamic representation of low-level stimulus features continues to evolve as we learn more about the underlying computations being performed by sensory systems, and the kinds of feature representations needed to perform these computations (Thorson, Liénard, & David, 2015).

Decoding models

While decoding models typically utilize the same features as encoding models, there are special precautions to consider because inputs and outputs are reversed relative to encoding models. Speech decoding is a complex problem that can be approached with different goals, strategies, and methods. In particular, two main categories of decoding models have been employed: classification and reconstruction.

In a classification framework, the neural activity during specific events is identified as belonging to one of a finite set of possible event types. For instance, one of six words or phrases. There are many algorithms (linear and non-linear) for fitting a classification model, such as support-vector machines, Bayesian classifiers, and logistic regression (Hastie et al., 2009). All these algorithms involve weighting input features (neural signals) and outputting a discrete value (the class of a datapoint) or a value between 0 and 1 (probability estimate for the class of a datapoint). This may be used to predict many types of discrete outputs, such as the trial or stimulus "types" (e.g., consonant vs. dissonant chords), image recognition (Rieger et al., 2008), finger movements (Quandt et al., 2012), social decisions (Hollmann et al., 2011), or even subjective conscious percepts (Reichert et al., 2014). In this case, the experimental design requires a finite number of repetitions of each stimulus type (or class). In speech research, discrete speech features have been predicted above chance levels, such as vowels and consonants (Bouchard & Chang, 2014; Pei, Barbour, Leuthardt, & Schalk, 2011), phonemes (Brumberg, Wright, Andreasen, Guenther, & Kennedy, 2011; Chang et al., 2010; Mugler et al., 2014), syllables (Blakely, Miller, Rao, Holmes, & Ojemann, 2008), words (Kellis et al., 2010; Martin et al., 2016), sentences (Zhang et al., 2012), segmental features (Lotte et al., 2015) and semantic information (Degenhart, Sudre, Pomerleau, & Tyler-Kabara, 2011).

In a reconstruction approach, continuous features of the stimulus are reconstructed to match

the original feature set. For instance, upper limb movement parameters, such as position, velocity and force were successively decoded to operate a robotic arm (Hochberg et al., 2012). In speech reconstruction, features of the sound spectrum, such as formant frequencies (Brumberg, Nieto-Castanon, Kennedy, & Guenther, 2010), amplitude power and spectrotemporal modulations (Martin et al., 2014, 2016; Pasley et al., 2012), mel-frequency cepstral-coefficients (Chakrabarti, Krusienski, Schalk, & Brumberg, 2013), or the speech envelope (Kubanek, Brunner, Gunduz, Poeppel, & Schalk, 2013) have been accurately reconstructed. In a recent study, formant frequencies of intended speech were decoded in real-time directly from the activity of neurons recorded from intracortical electrodes implanted in the motor cortex, and speech sounds were synthesized from the decoded acoustic features (Brumberg et al., 2010).

While both encoding and decoding models are used to relate stimulus features and neural activity, decoding models have an added potential to be used in applications that attempt to use patterns of neural activity to control physical objects (such as robotic arms) or predict the stimulus properties underlying the neural activity (such as inner speech prediction). These are both examples of neural prosthetics, which are designed to utilize brain activity to help disabled individuals interact with the world and improve their quality of life. However, it is also possible (and preferable in some cases) to decode stimulus properties *using an encoding model*. In this case, encoding model parameters may be used to build probability distributions over the most likely stimulus properties that resulted in a (novel) pattern of brain activity (Kay, Naselaris, Prenger, & Gallant, 2008; Naselaris et al., 2011; Nishimoto et al., 2011).

In summary, linearizing stimulus features allows one to use linear models to find non-linear relationships between datasets. This approach is simpler, requires less computation, and is generally more interpretable than using non-linear models, and is flexible with respect to the kinds of features chosen (de Heer et al., 2017; Naselaris et al., 2011; Shamma, 2015). The challenge often lies in choosing these features based on previous literature and the hypothesis one wants to test, and interpreting the resulting model weights (see Interpreting Models section, as well as **Figure 2** for a description of many features used in predictive modeling).

Choosing and fitting the model

After choosing stimulus features (as inputs to an encoding model, or outputs to a decoding model) as well as the neural signal of interest, one must link these two data sets by “fitting” the model. The choice of modeling framework will influence the nature of the inputs and outputs, as well as the questions one may ask with it. This section discusses common modeling frameworks for encoding and decoding (see **Figure 3** for a general description of the components that make up each modeling framework). It focuses on the linear model, an approach that has proven to be powerful in answering complex questions about the brain. We highlight some caveats and best-practices.

Choosing a modeling framework

The choice of modeling framework affects the relationship one may find between inputs and outputs. Finding more complex relationships usually requires more data and is prone to overfitting, while finding simpler relationships can be more straightforward and efficient, but runs the risk of missing a more complex relationship between inputs and outputs.

While many model architectures have been used in neural modeling, this paper focuses on those that find linear relationships between inputs and outputs. We focus on this case because of the ubiquity and flexibility of linear models, though it should be noted that many other model structures have been used in the literature. For example, it is common to include non-linearities on the *output* of a linear model (e.g., a sigmoid that acts as a non-linear suppression of output amplitude). This can be used to transform the output into a value that corresponds to neural activity such as a Poisson firing rate (Christianson, Sahani, & Linden, 2008; Paninski, 2004), to incorporate knowledge of the biophysical properties of the nervous system (McFarland, Cui, & Butts, 2013), to incorporate the outputs of other models such as neighboring neural activity (Pillow, Ahmadian, & Paninski, 2011), or to accommodate a subsequent statistical technique (e.g., in logarithmic classification, see above). It is also possible to use summary statistics or mathematical descriptions of the receptive fields described above as inputs to a subsequent model (Thorson et al., 2015).

It is possible to fit non-linear models directly in order to find more complex relationships between inputs and outputs. These may be an extension of linear modeling, such as models that estimate input non-linearities (Ahrens et al., 2008), spike-triggered covariance (Paninski, 2003; Schwartz, Pillow, Rust, & Simoncelli, 2006), and other techniques that fit multi-component linear filters for a single neural output (Meyer et al., 2017; Sharpee, Rust, & Bialek, 2004). Note that, after projecting the stimulus into the subspace spanned by these multiple filters, the relationship between this projection and the response can be non-linear, and this approach can be used to estimate the higher-order terms of the stimulus-response function (Eggermont, 1993). While non-linear methods find a more complicated relationship between inputs and outputs, they may be hard to interpret (but see Sharpee, 2016), require significantly more data in order to generalize to test-set data, and often contain many more free-parameters that must be tweaked to optimize the model fit (Ahrens et al., 2008). In addition, optimization-based methods for fitting these models generally requires traversing a more complex error landscape, with multiple local minima that do not guarantee that the model will converge upon a global minimum (Hastie et al., 2009).

As described in Section 3, generalized linear models provide the complexity of non-linear feature transformations (in the form of feature extraction steps) with the simplicity and tractability of a linear model. For this reason linear modeling has a strong presence in neuroscience literature, and will be the focus of this manuscript. See (Meyer et al., 2017) for an in-depth review of many (linear and non-linear) modeling frameworks that have been used in neural encoding and decoding.

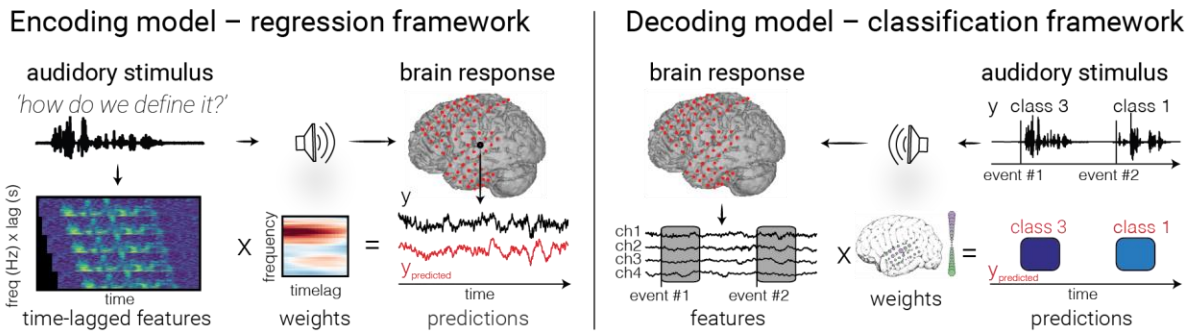


Figure 3 – Model fitting. *An example of encoding (left panel) and decoding (right panel) models are depicted. In encoding models, one attempt to predict the neural activity from the auditory representation by finding a set of weights (one for each feature / time lag) that minimizes the difference between true (black) and predicted (red) values. In decoding with a classifier (right), brain activity in multiple electrodes is used to make a discrete prediction of the category of a stimulus. Note that decoding models can also use regression to predict continuous auditory feature values from brain activity, though only classification is shown above.*

The least-squares solution

As described above, generalized linear models offer a balance between model complexity and model interpretability. While any kind of non-linear transformation can be made to raw input or output features *prior* to fitting, the model itself will then find *linear* relationships between the input and output features. At its core, this means finding one weight per feature such that, when each feature is weighted and summed, it either minimizes or maximizes the value of some function (often called a "cost" function). A common formulation for the cost function is to include "loss" penalties such as model squared error (Hastie et al., 2009) on both the training and the validation set of data. The following paragraphs describe a common way to define the loss (or error) in linear regression models, and how this can be used to find values for model coefficients.

In the case of least-squares regression, we define the predictions of a model as the dot product between the weight vector and the input matrix:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}$$

In this case, the cost function is simply the squared difference between the predicted values and the actual values for the output variable. It takes the following form:

$$CF_{LS} = error = \frac{1}{n}(\hat{\boldsymbol{y}} - \boldsymbol{y})^T(\hat{\boldsymbol{y}} - \boldsymbol{y})$$

In this case, $\boldsymbol{X}$ is the input training data and w are the model weights, and the term $\hat{\boldsymbol{y}}$ represents model predictions given a set of data. $\boldsymbol{y}$ is the "true" output values, and n is the total number of data points. Both y and $\hat{y}$ are column vectors where each row is a point in time. CF_{LS} stands for the "least squares" cost function. In this case it contains a single loss function that measures the average squared difference between model predictions and "true" outputs.

If there are many more data points than features (a rule of thumb is to have at least ten times more data points than features, though this is context-dependent), then finding a set of weights that minimizes this loss function (the squared error) has a relatively simple solution, known as the *Least Squares Solution* or the *Normal equation*. It is the solution obtained by maximum likelihood with the assumption of Gaussian error. The least square solution is:

$$weights_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Where X is the (*n* time points or observations by *m* features) input matrix, and y is an output vector of length *n observations*. When $\boldsymbol{X}$ and $\boldsymbol{y}$ have a mean of zero, the expression ($\frac{\boldsymbol{X}^T\boldsymbol{y}}{\boldsymbol{n}}$) is the cross-covariance between each input feature and the output. This is then normalized by the auto-covariance matrix of the input features ($\frac{\boldsymbol{X}^T\boldsymbol{X}}{\boldsymbol{n}}$). The output will be a vector of length *m* feature weights that defines how to mix input features together to make one predicted output. It should be noted that while this model weight solution is straightforward to interpret and quick to find, it has several drawbacks such as a tendency to "overfit" to data, as well as the inability to impose relationships between features (such as a smoothness constraint). Some of these will be discussed further in section 4.3, Using regularization to avoid overfitting.

From regression to classification

While classification and regression seem to perform very different tasks, the underlying math between them is surprisingly similar. In fact, a small modification to the regression equations results in a model that makes predictions between two classes instead of outputting a continuous variable. This occurs by taking the output of the linear model and passing it through a function that maps this output onto a number representing the probability that a sample comes from a given class. The function that does this is called the *link function*.

$$p_{class} = f^{-1}(\boldsymbol{X}\boldsymbol{w} + \boldsymbol{b})$$

Where p is the probability of belonging to one of the two classes and f^{-1} is the inverse of the link function (called the *inverse link function*). For example, in logistic regression, *f* is given by the logistic function:

$$log\left(\frac{p}{1-p}\right) = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{b}$$

$\boldsymbol{X}\boldsymbol{w}$ is the weighted sum of the inputs, and the scalar $\boldsymbol{b}$ is a bias term. Taken together, this term defines the angle ($\boldsymbol{w}$) and distance from origin ($\boldsymbol{b}$) of a line in feature space that separates the two classes, often called the *decision plane*.

Datapoints will be categorized as belonging to one class of another depending on which side of the line they lie. The quantity $\boldsymbol{X}\mathbf{w} + \mathbf{b}$ provides a normalized distance from each sample in $\boldsymbol{X}$ to the classifier's decision plane (which is positioned at a distance, b, from the origin). This distance can be associated with a particular probability that the sample belongs to a class. Note that one can also use a step function for the link function, thus generating binary YES/NO predictions about class identity.

While the math behind various classifiers will differ, they are all essentially performing the same task: define a means of "slicing up" feature space such that datapoints in one or another region of this space are categorized according to that region's respective class. For example, *Support Vector Machines* also find a linear relationship that separates classes in feature spaces, with an extra constraint that controls the distance between the separating line and the nearest member of each class (Hastie et al., 2009).

Using regularization to avoid overfitting

The analytical least-squares solution is simple, but often fails due to *overfitting* when there are a high number of feature dimensions (*m*) relative to observations (*n*). In overfitting, the weights become too sensitive to fluctuations in the data that would average to zero in larger data sets. As the number of parameters in the model grows, this sensitivity to noise increases. Overfitting is most easily detected when the model performs well on the training data, but performs poorly on the testing data (see section 5, *Validating the model*).

Neural recordings are often highly variable either because of signal to noise limitations of the measures or because of the additional difficulty of producing a stationary internal brain state (Sahani & Linden, 2003; Theunissen et al., 2001). At the same time, there is increasing interest in using more complex features to model brain activity. Moreover, the amount of available data is often severely restricted, and in extreme cases there are fewer datapoints than weights to fit. In these cases the problem is said to be *underconstrained*, reflecting the fact that there is not enough data to properly constrain the weights of the model. To handle such situations and to avoid overfitting the data, it is common to employ *regularization* when fitting models. The basic goal of regularization is to add constraints (or equivalently priors) on the weights to effectively reduce the number of parameters (*m*) in the model and prevent overfitting. Regularization is also called shrinking, as it shrinks the number or magnitude of parameters. A common way to do this is to use a penalty on the total magnitude of all weight values. This is called imposing a "norm" on the weights. In the Bayesian framework, different types of penalties correspond to different priors on the weights. They reflect assumptions on the probability distribution of the weights *before* observing the data (Naselaris et al., 2011; Wu et al., 2006).

In machine learning, norms follow the convention ℓN, where N is generally 1 or 2 (though it could be any value in between). The ℓN norm for a vector is defined as follows:

$$\left|\left|x\right|\right|_n = \left(\sum_i |x_i|^n\right)^{\frac{1}{n}}$$

When applied to the model weights, the ℓN norm reflects the magnitude of all weights combined. This can be added to the model's cost function, supplementing the traditional least squares loss function. For example, using the $\ell 2$ norm (in a technique called *Ridge Regression*) adds an extra penalty to the squared sum of all weights, resulting in the following value for the regression cost function:

$$CF_{Ridge} = \frac{1}{n}(\boldsymbol{Xw} - \boldsymbol{y})^2 + \lambda\left|\left|\boldsymbol{w}\right|\right|^2$$

Where $\boldsymbol{w}$ is the model weights, and λ is a hyper-parameter (in this case called the Ridge parameter) that controls the relative influence between the weight magnitude vs. the mean squared error. Ridge regression corresponds to a Gaussian prior on the weight distribution with variance given by $\frac{1}{\lambda}$. For small values of λ, the optimal model fit will be largely driven by the squared error, for large values, the model fit will be driven by minimizing the magnitude of model weights. As a result, all of the weights will trend towards smaller numbers. For Ridge regression, the weights can also be obtained analytically:

$$weights_{Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{I}\lambda)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

There are many other forms of regularization, for example, $\ell 1$ regularization (also known as Lasso Regression) adds a penalty for the sum of the absolute value of all weights and causes many weights to be close to 0, while a few may remain larger (known as fitting a *sparse* set of weights). It is also common to simultaneously balance $\ell 1$ and $\ell 2$ penalties in the same model (called *Elastic* Nets, (Hastie et al., 2009)).

In general, regularization tends to reduce the *variance* of the weights, restricting them to a smaller space of possible values and reducing their sensitivity to noise. In the case of ℓN regression, this is often described as placing a finite amount of magnitude that is spread out between the weights. The N in ℓN regression controls the extent to which this magnitude is given to a small subset of weights vs. shared equally between all weights. For example, in Ridge regression, large weights are penalized more, which encourages **all** weights to be smaller in value. This encourages weights that smoothly vary from one to another, and may discourage excessively high weights on any one weight which may be due to noise. Regularization reduces the likelihood that weights will be overfit to noise in the data and improves the testing data score. $\ell 2$ regularization also has the advantage of having an analytical solution, which can speed up computation time. An exhaustive description of useful regularization methods and their effect on analyses can be found in Hastie et al., (2009).

Parameters that are not directly fit to the training data (such as the Ridge parameter) are called *hyper-parameters* or *free parameters*. They exist at a higher level than the fitted model weights, and influence the behavior of the model fitting process in different ways (e.g., the number of non-zero weights in the model, or the extent to which more complex model features can be created out of combinations of the original features). They are not determined in the standard model fitting process, however they can be chosen in order to minimize the error on a validation dataset (see below). Changing a hyper-parameter in order to maximize statistics such as prediction score is called *tuning* the parameter, which will be covered in the next section.

In addition, there are many choices made in predictive modeling that are not easily quantifiable. For example, the choice of the model form (e.g., $\ell 2$ vs $\ell 1$ regularization) is an additional free model parameter that will affect the result. In addition, there are often multiple ways to "fit" a model. For example, the least-squares solution is not always solved in its analytic form. If the number of features is prohibitively large, it is common to use numerical approximations to the above equation, such as gradient descent, which uses an iterative approach to find the set of weights that minimizes the cost function. With linear models that utilize enough independent data points, there is always one set of weight parameters that has

the lowest error, often described as a “global minimum”. In contrast, non-linear models have a landscape of both local and global minima, in which small changes to parameter values will *increase* model error and so the gradient descent algorithm will (incorrectly) stop early. In this way, iterative methods may get “stuck” in a local minimum without reaching a global minimum. Linear models do not suffer from the problem of local minima. However, since gradient descent often stops before total convergence, it may result in (small) variations in the final solution given different weight initializations.

Note that for linear time-invariant models (i.e., when the weights of the model do not change over time) and when the second order statistical properties of the stimulus are stationary in time (i.e., the variance and covariance of the stimulus do not change with time), then it is more efficient to find the linear coefficients of the model in the Fourier domain. For stimuli with those time-invariant properties, the eigenvectors of the stimulus auto-covariance matrix ($\frac{X^T X}{n}$ in the normal equation) are the discrete Fourier Transform. Thus, by transforming the cross-correlation between the stimulus and the response ($\frac{X^T y}{n}$) into the frequency domain, the normal equation becomes a division of the Fourier representation of $X^T y$ and the power of the stimulus at each frequency. Moreover, by limiting the estimation of the linear filter weights to the frequencies with significant power (i.e., those for which there is sufficient sampling in the data), one effectively regularizes the regression. See (Theunissen et al., 2001) for an in-depth discussion.

Validating the model

After data have been collected, model features have been determined, and model weights have been fit, it is important to determine whether the model is a “good” description of the relationship between stimulus features and brain activity. This is called *validating* the model. This critical step involves making model predictions using new data and determining if the predictions capture variability in the “ground truth” of data that was recorded.

Validating a model should be performed on data that was not used to train the model, including preprocessing, feature selection, and model fitting. It is common to use *cross-validation* to accomplish this. In this approach, the researcher splits the data into two subsets. One subset is used to train the model (a “training set”), and the other is used to validate the model (a “test set”). If the model has captured a “true” underlying relationship between inputs and outputs, then the model should be able to accurately predict data points that it has never seen before (those in the test set). This gives an indication for the stability of the model’s predictive power (e.g., how well is it able to predict different subsets of held-out data), as well as the stability of the model weights (e.g., placing confidence intervals on the weight values.)

There are many ways to perform cross-validation. For example, in *K*-fold cross validation, the dataset is split into *K* subsets (usually between 5 and 10). The model is fit on *K-1* subsets, and then validated on the held-out subset. The cross validation iterates over these sets until each subset was once a test set. In the extreme case, there are as many subsets as there are datapoints, and a single datapoint is left out for the validation set on each iteration. This is

called Leave One Out cross validation, though it may bias the results and should only be used if very little data for training the model is available (Varoquaux et al., 2016). Because electrophysiology data is correlated with itself (i.e., autocorrelated) in time, it is crucial when creating training/test splits to avoid separating datapoints that occur close to one another in time (for example, by keeping "chunks" of contiguous timepoints together, such as a single trial that consists of one spoken sentence). If this is not done, correlations between datapoints that occur close to one another in time will artificially inflate the model performance when they occur in both the training and test sets. This is because the model will be effectively trained and tested on the same set of data, due to patterns in both the signal and the noise being split between training / test sets. See **Figure 4** for a description of the cross-validation process, as well as the Jupyter notebook "Prediction and Validation", section "Aside: what happens if we don't split by trials?"[3].

Determining the correct hyper-parameter for regularization requires an extra step in the cross-validation process. The first step is the same: the full dataset is split into two parts, training data and testing data (called the "outer loop"). Next, the training data is split once more into training and validation datasets (called the "inner loop"). In the inner loop, a range of hyper-parameter values is used to fit models on a subset of the training data, and each model is validated on the held-out validation data, resulting in one model score per hyper-parameter value for each iteration of the inner loop. The "best" hyper-parameter is chosen by aggregating across inner loop iterations, and choosing the hyper-parameter value with the best model performance. The model with this parameter is then re-tested on the outer loop testing data. The process of searching over many possible hyper-parameter values is called a "grid search", and the whole process of splitting training data into subsets of training / validation data is often called nested-loop cross validation. Efficient hyper-parameter search strategies exist for some learning algorithms (Hastie et al., 2009). However, there are caveats to doing this effectively, and the result may still be biased with particularly noisy data (Varoquaux et al., 2016).

[3]http://beta.mybinder.org/v2/gh/choldgraf/paper-encoding_decoding_electrophysiology/master?filepath=notebooks/Prediction%20and%20Validation.ipynb

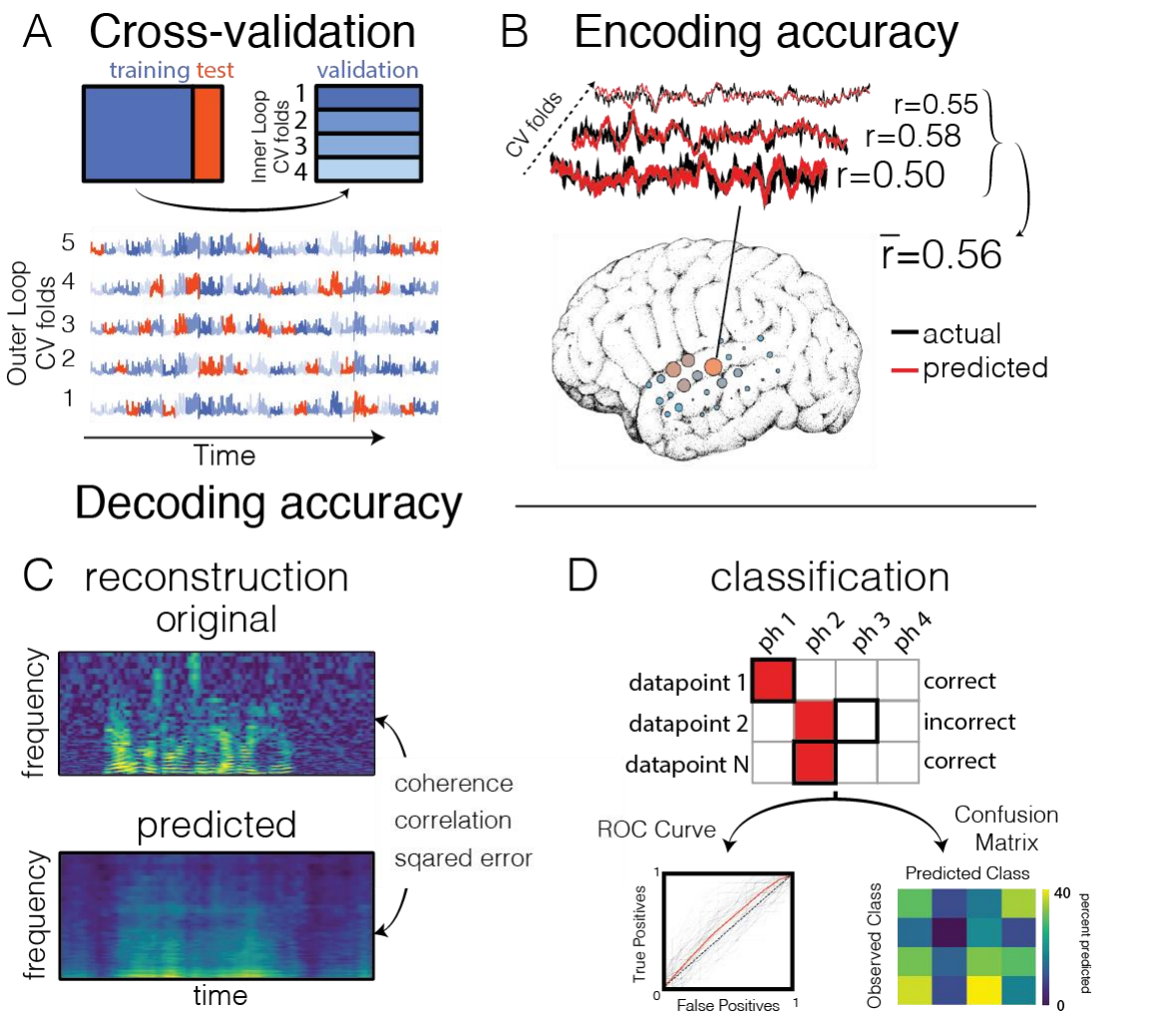


Figure 4 – Validation and prediction. *A. Cross validation is used to tune hyperparameters and validate the model. In one iteration of the outer loop, the data is split into training and test sets. Right: an inner loop is then performed on the training set, with a different subset of training data (blue shades) held out as a validation set for assessing hyperparameter performance. The hyperparameter with the highest mean score across inner-loop iterations is generally chosen for a final evaluation on the test set. Lower: The same neural timeseries across five iterations of the outer-loop. Each iteration results in a different partitioning of the data into training, test, and validation sets. Note that timepoints are grouped together in time to avoid overfitting during hyperparameter tuning.* ***B.*** *Examples of actual and predicted brain activity for various cross-validated testing folds. The overall prediction score is averaged across folds, and displayed on the surface of the subject's reconstructed brain.* ***C.*** *In decoding models when performing stimulus reconstruction (regression), a model is fit for each frequency band. Model predictions may be combined to create a predicted spectrogram. The predicted and original auditory spectrograms are compared using metrics such as mean squared error. D. When using classification for decoding, the model predicts one of several classes for each test datapoint. These predictions are validated with metrics such as the Receiver Operating Characteristic (ROC, left) that shows the performance of a binary classifier system as its discrimination threshold is varied. The ROC curve is shown for each outer CV iteration (black) as well as the mean across CV iterations (red). If the classifier outputs the labels above chance level, the Area Under the ROC curve (AUC) will be larger than 0.5. Alternatively, the model performance can be compared across classes resulting in a confusion matrix*

(right), which shows for what percent of the testing set a class was predicted (columns) given the actual class (rows). The ith row and jth column represents the percent of the time that a datapoint of class i was predicted to belong to class j.

Metrics for regression prediction scores

As described previously, inputs and outputs to a predictive model are generally created using one or more non-linear transformations of the raw stimulus and neural activity. The flexible nature of inputs and outputs in regression means that there are many alternative fitted models. In general, a model's performance is gauged from its ability to make predictions about data it has never seen before (data in a validation or test set) requiring a criterion to perform objective comparisons among all those models. The definition of model performance depends on the type of output for the model (e.g., a time series in regression vs. a label in categorization). It will also depend on the metric of error (or loss function) used, which itself depends on assumptions about the noise inherent in the system (e.g., whether it is normally-distributed). Assumptions about noise will depend on both the neural system being studied (e.g., single units vs. continuous variables such as high-frequency activity in ECoG) as well as the kind of model being used (Paninski, 2004). The metric of squared error (described below) assumes normally-distributed noise, and will be assumed for continuous signals in the remainder of the text.

Coefficient of Determination (R^2)

Encoding models as well as decoding models for stimulus reconstruction use regression, which outputs a continuously varying value. The extent to which regression predictions match the actual recorded data is called model *goodness of fit (GoF)*. A robust measure is the *Coefficient of Determination (R^2),* defined as the squared error between the predicted and actual activity, divided by the squared error that would have occurred with a model that simply predicts the mean of the true output data.

$$SSE_{tot} = \sum_i (y_i - \bar{y})^2$$
$$SSE_{reg} = \sum_i (y_i - \hat{y}_i)^2$$
$$R^2 = 1 - \frac{SSE_{reg}}{SSE_{tot}}$$

where $\hat{y}_i$ is the predicted value of y at timepoint i, and $\bar{y}$ is the mean value of y over all timepoints. The first two terms are both called the *sum of squared error.* One is the error defined by the model (the difference between predicted and actual values), and the other is the error defined by the output's deviation around its own mean (closely related to the output variance). Computing the ratio of errors provides an index for the increase in output variability explained by the regression model. If R^2 is positive it means that the variance of the model's error is less than the variance of the testing data, if it is zero then the model makes predictions no better than a model that simply predicts the mean of the testing data, and if it is negative

then the variance of the model's error is larger than the variance of the testing data (this is only possible when the linear model is being tested on data on which it was not fit).

The Coefficient of Determination, when used with a linear model and without cross-validation, is related to Pearson's correlation coefficient, *r*, by $R^2 = r^2$. However, on held-out data R^2 can be negative whereas the correlation coefficient squared (r^2) must be positive. Finally, R^2 is directly obtained from the sum of square errors which is the value that is minimized in regression with normally-distributed noise. Thus, it is a natural choice for GoF in the selection of the best hyper-parameter in regularized regression.

Coherence and mutual information

Another option for assessing model performance in regression is coherence. This approach uses Fourier methods to assess the extent to which predicted and actual signals share temporal structure. This is a more appropriate metric when the predicted signals are time series, and is given by the following form:

$$\gamma(\omega)^2 = \frac{\langle X(\omega)Y^*(\omega)\rangle\langle X^*(\omega)Y(\omega)\rangle}{\langle X(\omega)X^*(\omega)\rangle\langle Y^*(\omega)Y(\omega)\rangle}$$

where $X(\omega)$ and $Y(\omega)$ are complex numbers representing the stimulus and neural Fourier component at frequency ω, and $X^*(\omega)$ represents the complex conjugate. It is common to calculate the coherence at each frequency, ω, and then convert the output into *Gaussian Mutual Information (MI)*, an information theoretic quantity with units of $bits/sec$ (also known as the channel capacity) that characterizes an upper bound for information transmission for signals with a particular frequency power spectrum, and for noise with normal distributions. The Gaussian MI is given by:

$$MI_{norm}(\omega) = -\int_0^\infty log2(1-\gamma^2(\omega))d\omega$$

While this metric is more complex than using R^2, it is well-suited to the temporal properties of neural timeseries data. In particular, it provides a data-driven approach to determining the relevant time scales (or bandwidth) of the signal and circumvents the need for smoothing the signal or its prediction before estimating GoF values such as R^2 (Theunissen et al., 2001).

Metrics for classification prediction scores

Common statistics and estimating baseline scores

It is common to use *classification* models in decoding, which output a discrete variable in the form of a predicted class identity (such as a brain state or experimental condition). In this case, there is a simple "yes/no" answer for whether the prediction was correct. As such, it is common to report the percent correct of each class type for model scoring. This is then compared to a percent correct one would expect using random guessing (e.g., $100 * \frac{1}{n_{classes}}$). If there are different numbers of datapoints represented in each class, then a better baseline is the percentage of datapoints that belong to the most common class (e.g. $100 * \frac{n_A}{n_A+n_B}$). It should be noted that these are *theoretical* measures of guessing levels, but a better guessing level can often be estimated from the data (Rieger et al., 2008). For example, it is common to use a

permutation approach to randomly distribute labels among examples in the training set, and to repeat the cross validation several hundred times to obtain an estimate of the classification rate that can be obtained with such "random" datasets. This classification rate then serves as the “null” baseline. This approach may also reveal an unexpected transfer of information between training and test data that leads to an unexpectedly high guessing level.

ROC Curves

It is often informative to investigate the behavior of a classifier when the bias parameter, $\boldsymbol{b}$, is varied. Varying $\boldsymbol{b}$ and calculating the ratio of “true-positive” to “false-positives” creates the *Receiver Operating Characteristic* (ROC) curve of the classifier (Green & Swets, 1988). This describes the extent to which a classifier is able to separate the two classes. The integral over the ROC curve reflects the separability of the two classes independent of the decision criterion, providing a less-biased metric than percent correct (Hastie et al., 2009).

A geometric interpretation may help to understand how the ROC curve is calculated. The classifier's decision surface is an oriented plane in the space spanned by the features (e.g., a line in a 2-D space, if there are only two features). In order to determine the class of each sample, the samples are projected onto the normal vector of the decision plane by calculating $\boldsymbol{Xw}$**.** Samples on one side of the place will result in a positive value for $\boldsymbol{Xw}$, while samples on the other side of the plane will be negative. This corresponds to the two classes, and results in two histograms for the values of $\boldsymbol{Xw}$, one for each class. The decision criterion $\boldsymbol{b}$ can then be varied, resulting in different separations of the samples into two classes. By varying $\boldsymbol{b}$ for a range of values, and comparing the *predicted* vs the *true* labels for each value of $\boldsymbol{b}$, one calculates false positives (false alarms) and true positives (hits) for several decision planes with the same orientation but different positions. Calculating these values for many positions of the decision boundary constructs the ROC curve. A demonstration of the ROC curve and how it relates to the model’s hyperplane can be found in the provided jupyter notebooks.

The Area Under the Curve (AUC) is simply the total amount of area under the ROC curve, and is often reported as a summary statistic of the ROC curve. If the classifier is performing at chance, then the AUC will be .5, and if it correctly labels all datapoints for all decision thresholds, then the AUC will be 1. More advanced topics relating to classifier algorithms are covered in Hastie et al. (2009) and Pedregosa, Grisel, Weiss, Passos, & Brucher (2011).

The Confusion Matrix

In the case of multi-class classification (e.g. multinomial logistic regression), it is common to represent the results using a *confusion matrix.* In this visualization, each row is the “known” class, and each column is a predicted class. The *i, j*th value represents the number of times that a datapoint known to belong to class *i* was predicted to belong to class *j.* As such, the diagonal line represents correct predictions (where $class_{true} = class_{predicted}$), and any off-diagonal values represent incorrect predictions (see Figure 4D).

Confusion matrices are useful because they describe a more complex picture of how the model predictions perform. This makes it possible to account for more complex patterns in the model’s predictions. To capture information about systematic errors (for example if stimulus labels fall into subsets of groups between which the model cannot distinguish), one can use

confusion matrices to estimate the mutual information that fully describes the joint probabilities between the predicted class and the actual class (e.g., Chang et al., 2010; Elie & Theunissen, 2016).

What is a "good" model score?

Determining whether a model's predictive score is "good" or not is not trivial. Many regression and classification scoring metrics are a continuously varying number, and deciding a cutoff point above which a score is not only "statistically significant" but also large enough in effect size to warrant reporting is a challenging problem. This is particularly critical for applications such as Brain Computer Interfaces.

Statistical significance

A common practice in model fitting is to determine which models pass some criteria for statistical "significance". This usually means assessing whether the model is able to make predictions above chance (e.g., a coefficient of determination significantly different from zero in the case of regression, or an AUC greater than .5 in the case of classification). To assess importance and model generalizability, the researcher needs to compare the prediction of the new model to those obtained in other models (i.e. with other feature spaces or other, usually simpler, architectures). If improvements in GoF are clearly observed, then the researcher may investigate the model properties (such as the model weights) to determine which features were most influential in predicting outputs.

As mentioned above, there are multiple challenges with using predictive power to assess the performance of an encoding / decoding model. When fitting model parameters, most models assume that output signals have independent and either Gaussian- or Poisson-distributed noise. If this assumption does not hold (either because the signal and the noise are poorly estimated by the model, or because the noise is not actually Gaussian/Poisson), then the model parameters will be biased and the model less reliable, leading to considerations about whether the assumptions made by the model are valid. Note, however, that there have been recent efforts to fit non-linear models of the input/output function without explicitly assuming distributions of error (Fitzgerald, Sincich, & Sharpee, 2011).

Moreover, as with any statistic of brain activity, metrics for predictive power can be artificially inflated. For example, signals that are averaged, smoothed, or otherwise have strong low-frequency power will tend to give larger prediction scores, but may not represent the true relationship between stimuli and brain features. This is one reason to use metrics that are designed with time-series in mind, such as coherence, which does not depend on a particular level of smoothing applied to the data.

Estimating the prediction score ceiling

Another useful technique involves determining the highest possible prediction score one would expect given the variability in the data collected. A given R^2 value may be interpreted as "good" or "bad" based off the maximum expected R^2 possible for the dataset. This is called the "noise ceiling" of the data, and it allows one to calculate the percent of *possible* variance explainable by the model, instead of the percent of *total* variance explained by the model.

There is no guaranteed way to calculate the noise ceiling of a model, as it must be estimated from the data at hand. However, there have been attempts at defining principled approaches to doing so. These follow the principle that the recorded neural data is thought to be a combination of "signal" and "noise".

$$data_{stim\_i} = signal_{stim\_i} + noise$$

Note that in this case, only the signal component of the data is dependent on a given stimulus.

One may estimate the noise ceiling of a model based off of the signal-to-noise ratio (SNR) of the neural response to repetitions of the same stimulus. In this case, one randomly splits these repetitions into two groups and calculates the mean response to each, theoretically removing the noise component of the response in each group. The statistic of interest (e.g., R^2) is then calculated between each group. This process is repeated many times, and the resulting distribution of model scores can be used to calculate the noise ceiling. This process is explained in more detail in Hsu, Borst, & Theunissen, 2004 (section 4), and code for performing this is demonstrated in the Jupyter notebooks associated with this manuscript.

It is possible to perform the same approach using *different* stimuli by assuming that signals and noise have particular statistics. For example, the signal can be assumed to be restricted to low frequencies and the noise to have a normal distribution. If these assumptions hold, then it may be possible to estimate the maximum prediction score, but this risks arriving at a conservative estimate of this value due to some parts of the signal being treated as noise and averaged out. It is also important to note that these approaches assume a linear, invariant neural response to the stimulus, and it is more difficult to assess the theoretical maximum prediction score of the non-linear relationship between inputs and outputs (Sahani & Linden, 2003).

A note on multiple comparisons

The ability to perform multivariate analyses is both a blessing and a curse. On one hand, one can relate the activity of many stimulus features to a neural signal within a single modeling framework. On the other, this introduces new considerations when controlling for multiple comparisons and statistical inference.

The most notable benefit for multiple comparisons in the encoding / decoding model framework is the fact that input variables are considered jointly, meaning that it is not always necessary to run an independent test for each variable of interest. Instead, the researcher may inspect the pattern of activity across all model coefficients. For example, (Holdgraf et al., 2016) fit spectro-temporal receptive fields when electrocorticography patients heard degraded speech sentences. The authors compared the shape of the receptive field rather than performing inference on individual model coefficients. As such, relatively fewer statistical analyses were carried out by focusing on *patterns* in the receptive field rather than each parameter independently.

While predictive modeling can reduce the number of statistical comparisons by considering the joint pattern of coefficients across features, it also introduces new challenges for statistical comparisons. For example, natural stimuli offer an opportunity to investigate the relationship between neural activity and many different sets of features (e.g., spectrotemporal features,

articulatory features, and words; de Heer et al., 2017). As new features are used to fit models, there is an increased likelihood of a type 1 error. In these cases, it is crucial to define well-formulated hypotheses *before* fitting models with many different input features. Alternatively, one may use an encoding / decoding framework as an exploratory analysis step for the purpose of generating new hypotheses about the representation of stimulus features in the brain. These should then be confirmed on held-out data that has not yet been analyzed, or by follow-up experiments that are designed to test the hypotheses generated from the exploratory step. Ultimately it should be emphasized that while predictive models consider input features simultaneously, they are not a silver bullet for multiple comparisons problems, especially when performing statistical inference on individual model parameters (Bennett, Baird, Miller, & Wolfrod, 2009; Curran-Everett, 2000; Maris & Oostenveld, 2007).

Another challenge for multiple comparisons comes with the choice of model and the parameters associated with this model. While this paper focuses on linear models with standard regularization techniques (Ridge regression), there are myriad architectures for linking input and output activity. It is tempting to try several types of encoding / decoding models when exploring data, and researchers should be careful that they are not introducing "experimenter free parameters" that may artificially inflate their Type 1 error rate.

Finally, the model itself often also has so-called *hyperparameters* that control the behavior of the model and the kind of structure that it finds in the input data. These hyperparameters have a strong influence on the outcome of the analysis, and should be tuned so that the model performs well on held-out validation data. Importantly, researchers cannot use the same set of data to both tune hyperparameters and test their model. Instead, it is best practice to use an *inner loop* (see above). This reduces the tendency of the model to over fit to training data (Hastie et al., 2009; Naselaris et al., 2011; Wu et al., 2006). If performing statistical inference on model parameters, this should be done *outside* of the inner-loop, after hyperparameters have already been tuned.

Interpreting the model

If one concludes that the model is capturing an important element of the relationship between brain activity and stimulus properties, one may use it to draw conclusions about the neural process under study. While encoding and decoding models have similar inputs and outputs, they can be interpreted in different, and often complementary ways (Weichwald et al., 2015). The proper method for fitting and interpreting model weights is actively debated, and the reader is urged to consult the current and emerging literature focused on predictive models of brain function (Naselaris et al., 2011; Varoquaux et al., 2016). In the following sections, we describe some challenges and best-practices in using predictive power to make scientific statements about the brain.

Encoding models

The simplest method for interpreting the results of a model fit is to investigate its weights. In a linear model, a positive weight for a given feature means that higher values of that feature correspond to higher values in the neural signal (they are correlated), a negative weight

suggests that increases in the feature values are related to a decrease in the neural signal (they are anti-correlated). If the magnitude of a weight is zero (or very small) it means that fluctuations in the values for that feature will have little effect on the neural signal. As such, investigating the weights amounts to describing the features that a particular neural signal will respond to, presumably because that feature (or one like it) is represented within the neural information at that region of the processing hierarchy. Note that the values of the different features have to be appropriately normalized during model training so that differences in the scale of features does not influence the magnitude of feature weights. This is typically done by z-scoring the values of each feature separately by subtracting its mean and dividing by its standard deviation.

If stimulus features have been chosen such that they have an interpretable meaning, then it is straightforward to assess meaning to the weight of each feature. In addition, if the features have a natural ordering to them (such as increasing frequency bands of a spectrogram, along with multiple time lags for each band), then the pattern of weights represents a receptive field for the neural signal. For example, spectrotemporal receptive fields have been shown to map onto higher-order acoustic features (Woolley, Gill, Fremouw, & Theunissen, 2009) and to increase in complexity as one moves through the auditory pathway (Miller, Escabí, Read, & Schreiner, 2002; Sen et al., 2001; Sharpee et al., 2011). This approach has also been used in humans to investigate the tuning properties as one moves across the superior temporal gyrus (Hullett et al., 2016). It is also possible to use statistical methods to find patterns in model coefficients across large regions of cortex. For example, (Huth et al., 2016, 2012) fit semantic word models (where each coefficient corresponded to one word) to each voxel in the human cortex. The authors then used Principle Components Analysis to investigate model coefficient covariance across widely distributed regions of the brain, finding consistent axes along which these coefficients covaried with one another.

Finally, another approach towards interpreting encoding models entails comparing model performance across multiple feature representations. For example, in de Heer et al. (2017), the authors investigated the representation of three auditory features (spectral, articulatory, and semantic features) across the cortical surface. They accomplished this by partitioning variance explained by each feature set individually, as well as by joint models incorporating combinations of these features. This enabled them to determine the extent to which each feature is represented across the cortex.

Decoding models

In a decoding approach, model weights are attached to each neural signal. Higher values for a signal mean that it is more important in predicting the output value of the stimulus / class used in the model. Interpreting the weights of decoding models can be challenging, as weights with a large amplitude do not necessarily mean that the neural signal encodes information about the stimulus (See “7. Differences between encoding and decoding models” for a more thorough discussion of this idea). It is important to rely on the statistical reliability of the model weight

magnitudes (e.g., low variance across random partitions of data) to extract interpretable features (Reichert et al., 2014).

Finally, it should be noted that in some cases, decoding models are used purely for making optimal predictions about stimulus values. For instance, in neurorehabilitation, decoding models have been used to predict 3D trajectories of a robotic arm for motor substitution (Hochberg et al., 2012). In this case, decoding is approached as an engineering problem, wherein the goal is to obtain the highest decoding predictions and interpreting model weights is of less importance.

General comments on interpretation

It is possible to use the predictive power of either encoding models (e.g., the R^2 of a model) or decoding models (e.g., the AUC calculated from an ROC curve) to make statements about the nature of stimulus feature representations in the brain. For example, if two models are fit on the same neural data, each with a different set of input features, one may compare the variance explained in the testing data by each model. By fitting multiple models, each with a different feature representation, and comparing their relative prediction scores, one may investigate the extent to which each of these feature representations are a "good" description of the neural response (Huth et al., 2016). However, comparing models with different types or numbers of features is not straightforward, as there are often relationships between the features used in each model, as well as difference in the number of parameters used. In this case, a variance partitioning approach can also be used to distinguish the variance exclusively explained by two (or more) models from the one exclusively explained by one and not the other. This is done by comparing the prediction scores of each model separately, as well as a joint model that includes all possible parameters (de Heer et al., 2017; Lescroart et al., 2015).

It is also possible to investigate the weights and predictive power across models trained in different regions of the brain to investigate how the relationship between stimulus features and brain activity varies across cortex. By plotting a model's predictive power as a function of its neural location, one may construct a tuning map that shows which brain regions are well-predicted by a set of features (Huth et al., 2016). Moreover, by summarizing receptive fields by the feature value that elicits the largest response in brain activity, and plotting the "preferred feature" for each region of the brain, one may construct a *tuning map* that describes how the neural response within a particular set of features is distributed in the brain (Hullett et al., 2016; Huth et al., 2016; Moerel et al., 2013).

By choosing the right representations of features to include in the model, it may be possible to reliably predict all of the variability in brain activity that is dependent on the controlled experimental parameters. Note that the activity that arises from non-experimental factors, e.g. from internal states not controlled in the experiment or from neural and measurement noise, cannot be predicted. This goal requires special considerations for choosing stimuli and experimental design, which will be discussed in the final section.

Differences between encoding and decoding models

Differences in terminology and causality

While it is tempting to treat encoding and decoding models as two sides of the same coin, there are important differences between them in an experimental context. Encoding and decoding models have different assumptions about the direction of causality that may influence the possible interpretations of the model depending on the experiment being conducted.

Encoding models are often called *Forward* models, reflecting the direction of time from stimulus to neural activity. Conversely, decoding models are often called *Backward* or *Inverse* models, as they move "backwards" in time in a traditional sensory experiment (Crosse, Di Liberto, Bednar, & Lalor, 2016; Thirion et al., 2006). However, it should be noted that this is not always the case, as sometimes a decoded value (e.g., a movement) is actually driven by neural activity. For this reason we prefer the more specific terminology of *encoding* and *decoding*.

The nature of the experiment may also influence the terminology employed. For example, in an experimental paradigm in which stimuli in the world give rise to recorded brain activity (e.g., an experiment where subjects listen to speech), an encoding model naturally models the direction of causality from stimuli to brain activity. As such, it is called a *causal* model. On the other hand, in this experiment a decoding model operates in the opposite direction, inferring properties of the world from the neural activity. This is often called an *acausal* model.

The importance of specifying the direction of causality, and accounting for this in model choice and interpretation, is discussed in greater detail in Weichwald et al., (2015). The following sections describe some important considerations.

Differences in regression

It is possible for decoding models to be constructed with a regression framework, similarly to how encoding models operate. For example, in Mesgarani & Chang, (2012) and Pasley et al. (2012), the experimenters fit one model for each stimulus feature being decoded. This amounts to simply reversing the terms in the standard regression equations:

$$weights_{encoding} = (X^T X)^{-1} X^T y$$
$$weights_{decoding} = (Y^T Y)^{-1} Y^T x$$

It is tempting in this case to collect the coefficients of each decoding model and interpret this as if they came from an encoding model. However, it's important to note that a primary role of regression is to account for correlations between input features when estimating model coefficients. As explained in detail in Weichwald et al., (2015), if a stimulus feature X_i causally influences a neural feature Y_i, and if the stimulus feature X_i is *correlated* with another stimulus feature X_j (for example, if they share correlated noise, or if the stimulus features are naturally correlated), the decoder will give significant weights for both X_i and X_j, even though it is only X_i that influences the neural signal. This fact has important implications in the interpretation of model weights.

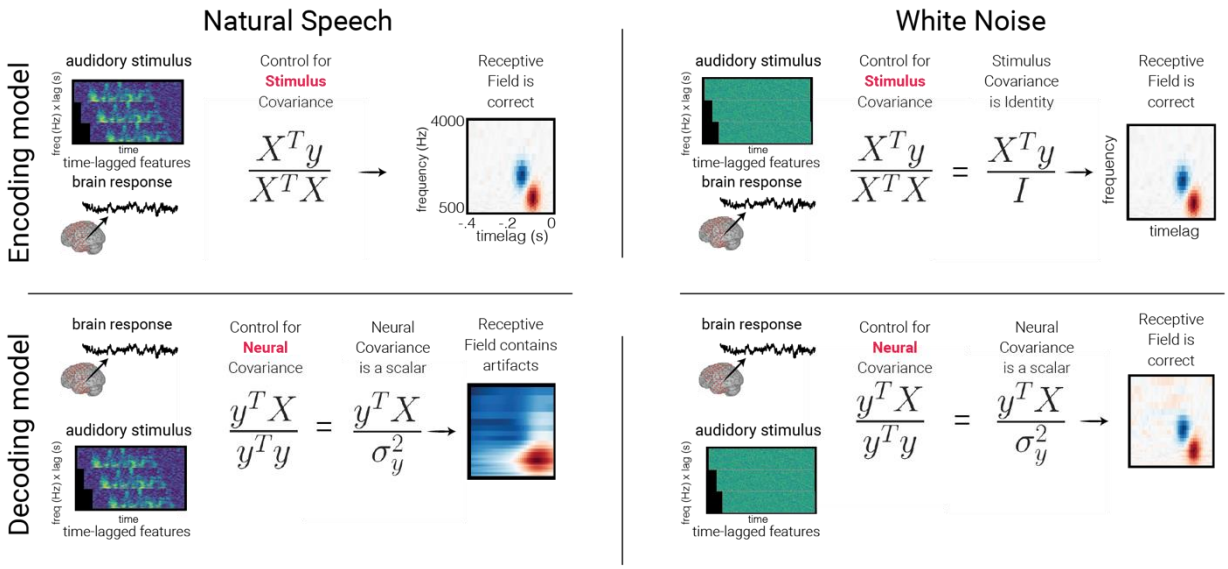


*Figure 5– **Comparing encoding and decoding weights.** An example of how using an encoding or a decoding framework can influence model results. In this case, we attempt to find the relationship between spectral features of sound and the neural activity. Left, upper: Using an encoding model, we naturally control for the covariance of the stimulus. Because the stimulus is natural speech, the correlations between lags and frequencies are accounted for, and the correct receptive field is recovered. Left, lower: In a decoding model, the X and y terms are reversed, and we instead control for the covariance of the neural activity. Because we have only one neural channel, the covariance term becomes a scalar (the variance of the neural signal). The decoded model weights are smeared in time and frequency. This is because of correlations that exist between these stimulus features. Right, upper/lower: The same approach applied to white noise stimuli instead of natural speech. In this case, there are no correlations between stimulus features, and so the covariance matrix becomes the identity matrix, making the receptive field in the encoding and decoding approach roughly the same. While this example is shown for receptive field modeling, the same caveat applies to any modeling framework where there are correlations between either inputs or outputs.*

Consider the case of receptive field modeling, in which auditory stimuli are presented to the individual, and a model is fit to uncover the spectral features to which the neural activity responds. In the encoding model, correlations between stimulus features are explicitly accounted for (X^TX), while in the decoding model, correlations between the *neural* features are accounted for (Y^TY). While it is possible to retrieve a receptive field using a decoding paradigm (e.g., by fitting one decoding model for each frequency / time-lag and collecting coefficients into a STRF), correlations in the stimulus features will skew the distribution of model coefficients. This might result in a STRF that is smoothed over a local region in delay / frequency. An encoding model should (theoretically) take these stimulus correlations into account, and only assign non-zero coefficient values to the proper features (see Figure 5). In this case it is important to consider the regularizer used in fitting the model, as there are differences in how regularization techniques distribute model weights with correlated features (Mesgarani, David, Fritz, & Shamma, 2009).

Differences in classification

The direction of causality also has important implications in the interpretation of classifiers. It is common to fit a classifier that predicts a stimulus type or neural state using neural features as inputs. In this case, it is tempting to interpret the magnitude of each weight as the extent to which that neural signal carries information about the state being decoded. However, this may not be the case. Following the logic above, if a neural signal with *no* true response to a stimulus is correlated with a neural signal that *does* respond to a stimulus, the classifier may (mistakenly) give positive magnitude to each. As such, one must exercise caution when making inferences about the importance of neural signals using model coefficients in an *Acausal* decoding model (Haufe et al., 2014; Mesgarani et al., 2009).

For example, monitoring the activity of brain regions *not* involved in representing stimulus features but instead reflecting some internal state (e.g. attention) may improve the quality of the decoder performance if attention is correlated with stimulus presentation. Such an effect would be due to the multivariate nature of the decoding model and could, in principle, be detected with additional univariate analyses. This is true of many decoding models, and may cause erroneous conclusions about an electrode's role in processing sensory features. However, as explained in Weichwald et al. (2015), the potential difficulty for causal interpretations in decoding approaches does not negate their usefulness: encoding and decoding models can be used in a complementary fashion to describe potential causal relationships between stimulus and corresponding neural activity in different brain regions.

Experimental Design

While much of this paper has covered the technical and data analytic side of predictive modeling, it is also important to design experiments with predictive models in mind. Fitting encoding and decoding models effectively requires particular considerations for the experimental manipulations and stimulus choices. We will discuss some of these topics below.

Task Design

While traditional experiments manipulate a limited number of independent variables between conditions, the strength of predictive modeling lies in using complex stimuli with many potential features of interest being presented continuously and overlapping in time. This has the added benefit that complex stimuli are generally closer to the "real world" of human experience. This adds to the experiment's *external validity*, which can be difficult to achieve with traditional experimental designs(Campbell & Stanley, 2015).

The simplest task for an encoding model framework is to ask the subject to passively perceive a stimulus presented to them. For example, Huth et al asked subjects to listen to series of stories told in the podcast *The Moth* (Huth et al., 2016). There was no explicit behavioral manipulation required of the subjects, other than attending to the stories. Using semantic features extracted from the audio, as well as BOLD activity collected with fMRI, the researchers were able to build encoding models that described how semantic categories drove the activity across wide regions of the cortex.

The use of complex stimuli does not preclude performing experimental manipulation. For example, Holdgraf et al. (2016) presented a natural speech stimulus to ECoG subjects, who were asked to passively listen to the sounds. These sentences came in triplets following a *degraded -> clean -> degraded* structure. By presenting the same degraded speech stimulus *before* and *after* the presentation of a non-degraded version of the sentence, the experimenters manipulated the independent variable of comprehension, and tested its effect on the neural response to multiple speech features.

It is also possible to ask subjects to actively engage in the task to influence how their sensory cortex interacts with the stimuli. Mesgarani et al used a decoding paradigm to predict the spectrogram of speech that elicited a pattern of neural activity (Mesgarani & Chang, 2012). They asked the subject to attend to one of two natural speech streams, the classic cocktail party effect. Thus, they experimentally manipulated the subject's attention, while the natural speech stimuli were kept the same. They compared the decoded spectrogram as a function of which speaker the subject was attending to, suggesting that attention modulates the cortical response to spectro-temporal features.

Stimulus construction

Choosing the proper stimuli is a crucial step in order to properly construct predictive models. A model's ability to relate stimulus features to brain activity is only as good as the data on which it is trained. For a model to be interpretable, it must be fit with a rich set of possible feature combinations that cover the stimulus statistics that are typical for the individual under study, and for the feature representations of interest. For example, it is difficult to make statements about how the brain responds to semantic information if the stimuli presented do not broadly cover semantic space.

There are many stimulus sets that are commonly used in predictive modeling of the auditory system. For example, the TIMIT corpus is a collection of spoken English sentences that are designed to cover a broad range of acoustic and linguistic features (Zue, Seneff, & Glass, 1990). This may be appropriate for studying lower-order auditory processes, though it is unclear whether stimuli such as these are useful for more abstract semantic processes, as the sentences do not follow any high-level narrative. Efforts have been made to construct more semantically rich stimuli (e.g., Huth et al, 2016), though it is difficult to properly tag a stimulus with the proper timing of linguistic features (e.g. phoneme and word onsets). A database with many types of linguistic / auditory stimuli can be found at *catalog.ldc.upenn.edu.*

How much data to collect?

The short answer to this question is always "as much as you possibly can". However, in practice many studies are time-limited in their ability to collect large quantities of data. One should take care to include enough stimuli such that the model has the right amount of data to make predictions on test set data. It is not possible to know exactly how much data is needed as this depends on both the number of parameters in the model as well as the noise in the signal being predicted. However, it is possible to estimate the amount of training samples required to achieve a reasonable predictive score given further assumptions about the complexity of the model and the expected noise variance (similar to traditional statistical power estimation).

Ideally, one should conduct pilot studies in order to determine the minimum number of trials, time-points, and other experimental manipulations required to model the relationship between inputs / outputs to some degree of desired accuracy. It is useful to plot a model's predictive score on testing data as a function of the number of data points included in fitting the model, this is called a *Learning Curve*. At some point, increasing the amount of data in the model fit will no longer result in an improvement in prediction scores. One should collect *at least* enough data such that predictive scores remain stable as more data is added. For insight into what is meant by "stable", see the simulation performed by Willmore and Smyth on a spiking neuron. These authors showed the shape of the reconstruction error curve for a number of fitting procedures and as a function of the number of stimulus presentations, finding that error decreases as the number of presentations goes up, and eventually bottoms-out (Willmore and Smyth, 2003, Fig. 5).

Finally, it is also advised to include multiple repetitions of stimuli that will be used purely for validating the model. This has two substantial benefits. First, having multiple instances of the brain's response to the same stimulus makes it easier to estimate the ceiling on model performance (see section Metrics for regression prediction scores). Second, if these repetitions happen at different points throughout the experiment, it is possible to use them to assess the degree of *stationarity* in the neural response. Most models assume that the relationship between the stimulus features and the brain activity will be stable over time. This is often not the case as brains are inherently plastic (e.g., Holdgraf et al., 2016; Meyer et al., 2014), and may change their responsiveness to stimuli based on experimental manipulations or broader changes such as levels of internal or external attention. Recording the neural response to the same stimulus throughout the experiment provides a metric of whether the assumption of stationarity holds.

Conclusions

Predictive modeling allows researchers to relate neural activity to complex and naturalistic stimuli in the world. Encoding models provide an objective methodology to determine the ability of different feature representations to account for variability in the neural response. Decoding models play a complementary role to encoding models, and allow for the reconstruction of stimuli from ensembles of neural activity, opening the door for future advancements in neuroprosthetics. Both approaches have been successfully used to model the neural response of single units (e.g. Theunissen et al., 2001), high-frequency electrode activity (e.g., Stephanie Martin et al., 2016; Stéphanie Martin et al., 2014; Mesgarani & Chang, 2012), and BOLD responses to low-level stimulus features (Nishimoto et al., 2011). They have also been used to investigate the neural response to higher-level stimulus features (e.g. Çukur, Nishimoto, Huth, & Gallant, 2013; Huth et al., 2016), as well as to investigate how this response changes across time or condition (e.g. J. Fritz, Shamma, Elhilali, & Klein, 2003; Meyer et al., 2014; Slee & David, 2015).

There are many caveats that come with a predictive modeling framework, including considerations for feature extraction, model selection, model validation, model interpretation, and experimental design. We have discussed many of these issues in this review and have

provided python tutorials to guide the reader in implementing these methods. We urge the reader to examine the citations provided for further details and to follow advances in this field closely as our understanding of its drawbacks and its potential continues to evolve.

Chapter 3 – Decoding models for speech reconstruction

Introduction

A promising use of regression models for modeling neural activity is their ability to predict patterns of stimulus features that elicited the recorded neural activity. This is called *decoding*, and allows one to reconstruct a new stimulus, given a pattern of brain activity. This chapter describes attempts at performing speech decoding using electrocorticography in order to study the neural representation of spoken (overt) and imagined (covert) speech.

Citation

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N. E., Rieger, J., Schalk, G., Knight, R.T., & Pasley, B.N. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. Front. Neuroeng. 7, 14. doi:10.3389/fneng.2014.00014.

Abstract

Auditory perception and auditory imagery have been shown to activate overlapping brain regions. We hypothesized that these phenomena also share a common underlying neural representation. To assess this, we used electrocorticography intracranial recordings from epileptic patients performing an out loud or a silent reading task. In these tasks, short stories scrolled across a video screen in two conditions: subjects read the same stories both aloud (overt) and silently (covert). In a control condition the subject remained in a resting state. We first built a high gamma (70-150 Hz) neural decoding model to reconstruct spectrotemporal auditory features of self-generated overt speech. We then evaluated whether this same model could reconstruct auditory speech features in the covert speech condition. Two speech models were tested: a spectrogram and a modulation-based feature space. For the overt condition, reconstruction accuracy was evaluated as the correlation between original and predicted speech features, and was significant in each subject ($p<10^{-5}$; paired two-sample t-test). For the covert speech condition, dynamic time warping was first used to realign the covert speech reconstruction with the corresponding original speech from the overt condition. Reconstruction accuracy was then evaluated as the correlation between original and reconstructed speech features. Covert reconstruction accuracy was compared to the accuracy obtained from reconstructions in the baseline control condition. Reconstruction accuracy for the covert condition was significantly better than for the control condition ($p< 0.005$; paired two-sample t-test). The superior temporal gyrus, pre- and post-central gyrus provided the highest reconstruction information. The relationship between overt and covert speech reconstruction depended on anatomy. These results provide evidence that auditory representations of covert speech can be reconstructed from models that are built from an overt speech data set, supporting a partially shared neural substrate.

Introduction

Mental imagery produces experiences and neural activation patterns similar to actual perception. For instance, thinking of moving a limb activates the motor cortex, internal object visualization activates the visual cortex, with similar effects observed for each sensory modality (Kosslyn, Ganis, & Thompson, 2001; Kosslyn & Thompson, n.d.; Roth et al., 1996; Stevenson & Case, 2005). Auditory imagery is defined as the mental representation of sound perception in the absence of external auditory stimulation. Behavioral and neural studies have suggested that structural and temporal properties of auditory features, such as pitch (Halpern, 1989), timbre (Halpern, Zatorre, Bouffard, & Johnson, 2004; Pitt & Crowder, 1992), loudness (Intons-Peterson, 1980) and rhythm (Halpern, 1988) are preserved during music imagery (Hubbard, 2013). However, less is known about the neural substrate of speech imagery. Speech imagery (inner speech, silent speech, imagined speech, covert speech or auditory verbal imagery) refers to our ability to "hear" speech internally without the intentional movement of any extremities, such as the lips, tongue, hands or auditory stimulation (Brigham & Kumar, 2010).

The neural basis of speech processing has been a topic of intense investigation for over a century (Hickok & Poeppel, 2007). The functional cortical organization of speech comprehension includes Heschl's gyrus (primary auditory cortex), the superior temporal gyrus (STG) and sulcus (STS; e.g., Wernicke's area). Speech production depends on premotor, motor and posterior inferior frontal regions (e.g., Broca's area; Heim, Opitz, & Friederici, 2002; Duffau et al., 2003; Billingsley-Marshall et al., 2007; Towle et al., 2008; Price, 2012). How these brain areas interact to encode higher-level components of speech such as phonological, semantic, or lexical features, as well as their role in covert speech, remains unclear. Increasing evidence suggests that speech imagery and perception activate the same cortical areas. Functional imaging studies have reported overlapping cortical regions during overt and covert speech generation in inferior frontal lobes, sensorimotor cortex regions, supplementary motor areas, and anterior cingulate gyri. Some regions commonly associated with motor aspects of speech production were not active during the silent task (Palmer et al., 2001; Rosen, Ojemann, Ollinger, & Petersen, 2000; Yetkin et al., 1995). Transcranial magnetic stimulation over motor sites and inferior frontal gyrus induced speech arrest in both overt and covert speech production (Aziz-Zadeh, Cattaneo, Rochat, & Rizzolatti, 2005). Finally, brain lesion studies have shown high correlation between overt and covert speech abilities, such as rhyme and homophones judgment (Geva, Bennett, Warburton, & Patterson, 2011) for patients with aphasia.

Imagery-related brain activation could result from top-down induction mechanisms including memory retrieval (Kosslyn, 2005; Kosslyn et al., 2001) and motor simulation (Guenther, Ghosh, & Tourville, 2006; Price, 2011; Tian & Poeppel, 2012). In memory retrieval, perceptual experience may arise from stored information (objects, spatial properties and dynamics) acquired during actual speech perception and production experiences (Kosslyn, 2005). In motor simulation, a copy of the motor cortex activity (efference copy) is forwarded to lower sensory cortices, enabling a comparison of actual with desired movement, and permitting online behavioral adjustments (Jeannerod, 2003). Despite findings of overlapping brain activation during overt and covert speech (Aleman, 2004; Aziz-Zadeh et al., 2005; Geva, Correia, &

Warburton, 2011; Hinke et al., 1993; McGuire et al., 1996; Palmer et al., 2001; Rosen et al., 2000; Yetkin et al., 1995), it is likely that covert speech is not simply overt speech without moving the articulatory apparatus. Behavioral judgment studies showed that aphasic patients indicated inner speech impairment, while maintaining relatively intact overt speech abilities, while others manifested the reverse pattern (Geva, Bennett, et al., 2011). Similarly, imaging techniques showed different patterns of cortical activation during covert compared to overt speech, namely in the premotor cortex, left primary motor cortex, left insula, and left superior temporal gyrus (Huang, Carr, & Cao, 2002; Pei et al., 2011; Shuster & Lemieux, 2005). This suggests that brain activation maps associated with both tasks are dissociated at least in some cases (Aleman, 2004; Feinberg, Gonzalez Rothi, & Heilman, 1986; Geva, Jones, Crinion, Baron, & Warburton, 2011; Shuster & Lemieux, 2005). The extent to which auditory perception and imagery engage similar underlying neural representations remains poorly understood.

To investigate similarities between the neural representations of overt and covert speech, we employed neural decoding models to predict auditory features experienced during speech imagery. Decoding models predict information about stimuli or mental states from recorded neural activity (Bialek, Rieke, de Ruyter van Steveninck, & Warland, 1991). This technique has attracted increasing interest in neuroscience as a quantitative method to test hypotheses about neural representation (Kay, Naselaris, Prenger, & Gallant, 2008; Naselaris, Kay, Nishimoto, & Gallant, 2011; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009; Pasley et al., 2012; Warland, Reinagel, & Meister, 1997). For instance, decoding models have allowed predicting continuous limb trajectories (Carmena et al., 2003; Hochberg et al., 2006, 2012) from the motor cortex. In the visual domain, visual scenes can be decoded from neural activity in the visual cortex (Warland et al., 1997; Kay et al., 2008). Similarly, this approach has been used to predict continuous spectrotemporal features of speech. We used this approach to compare decoding accuracy during overt and covert conditions in order to evaluate the similarity of speech representations during speech perception and imagery.

We hypothesized that speech perception and imagery share a partially overlapping neural representation in auditory cortical areas. We reasoned that if speech imagery and perception share neural substrates, the two conditions should engage similar neural representations. Thus, a neural decoding model trained from overt speech should be able to predict speech features in the covert condition. (Pasley et al., 2012) showed that auditory spectrotemporal features of speech could be accurately reconstructed, and used to identify individual words during various listening tasks. In this study, we used a similar neural decoding model trained on sounds from self-generated overt speech. This model was then used to decode spectrotemporal auditory features from brain activity measured during a covert speech condition. Our results provide evidence for a shared neural representation underlying speech perception and imagery.

To test these hypotheses we used electrocorticography (ECoG), which provides high spatiotemporal resolution recordings of non-primary auditory cortex (Leuthardt, Schalk, Wolpaw, Ojemann, & Moran, 2004). In particular, the high gamma band (HG, ~70-150 Hz) reliably tracks neuronal activity in all sensory modalities (Lachaux, Axmacher, Mormann, Halgren, & Crone, 2012) and correlates with the spike rate of the underlying neural population (Lachaux et al., 2012; Miller et al., 2007; Boonstra, Houweling, & Muskulus, 2009). HG activity in

auditory and motor cortex has been linked to speech processing (Towle et al., 2008; Pasley et al., 2012), and served as the input signal for all tested neural decoding models.

Materials and methods

Subjects and data acquisition

Electrocorticographic (ECoG) recordings were obtained using subdural electrode arrays implanted in 7 patients undergoing neurosurgical procedures for epilepsy (Table 1). All patients volunteered and gave their informed consent (approved by the Albany Medical College Institutional Review Board) before testing. The implanted electrode grids (Ad-Tech Medical Corp., Racine, WI; PMT Corporation, Chanhassen, MN) consisted of platinum–iridium electrodes (4 mm in diameter, 2.3 mm exposed) that were embedded in silicon and spaced at an inter-electrode distance of 0.6-1cm. Grid placement and duration of ECoG monitoring were based solely on the requirements of the clinical evaluation (see **Figure 6**).

Subject	Age	Sex	Handed-ness	FSIQ	VIQ	PIQ	LL	Seizure Focus	Grid/Strip Locations and Contact Numbers
S1	30	M	Right	74	64	90	Bi-lateral	Left temporal	Left temporal (35) Left temporal pole (4) Left fronto-parietal (48) Left occipital pole (4)
S2	29	F	Right	90	91	90	Left	Left temporal	Left temporal (35) Left fronto-parietal (56) Left temporal (4) Left occipital pole (4)
S3	26	F	Right	112	106	117	Left	Left temporal	Left temporal (35) Left fronto-parietal (64) Left temporal (4) Left occipital pole (4)
S4	56	M	Right	84	82	87	Left	Left temporal	Left temporal (35) Left fronto-parietal (56) Left occipital pole (4)
S5	26	M	Right	102	103	100	Left	Right temporal	Right temporal (35) Right fronto-parietal (64) Right frontal pole (6) Right occipital pole (6)
S6	45	M	Right	98	93	105	Left	Left frontal	Left front-temporal (54) Left temporal (4)
S7	29	F	Right	84	111	95	Bi-lateral	Left temporal	Left temporal (68) Left fronto-parietal (40) Left frontal pole (4) Left parietal (4) Left temporal (4)

Table 1. Clinical profiles of subjects. All of the subjects had normal cognitive capacity and were functionally independent. Full scale (FSIQ), verbal (VIQ) and performance (PIQ) intelligence has was based on the Wechsler Adult Intelligence Scale (WAIS-III) test. Language lateralization (LL) was based on the Wada test.

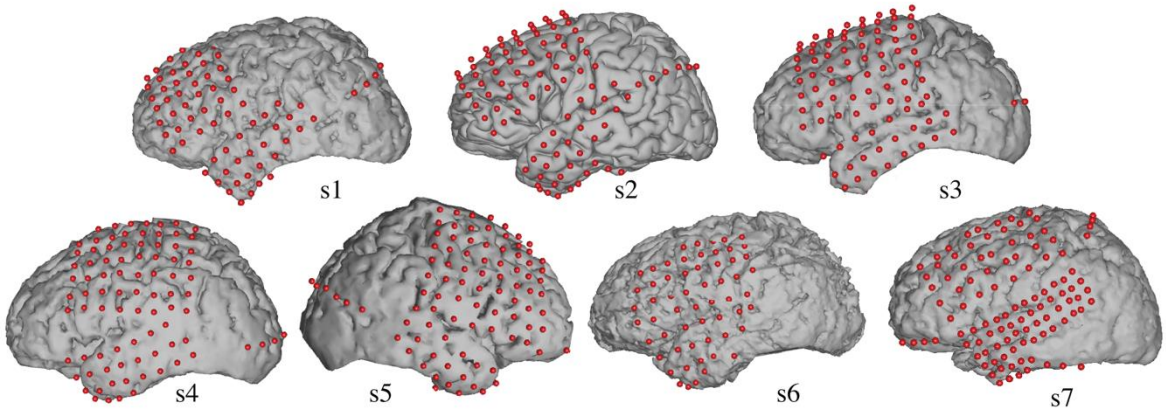


Figure 6 - Electrode locations. *Grid locations for each subject are overlaid on cortical surface reconstructions of each subject's MRI scan.*

ECoG signals were recorded at the bedside using seven 16-channel g.USBamp biosignal acquisition devices (g.tec, Graz, Austria) at a sampling rate of 9600 Hz. Electrode contacts distant from epileptic foci and areas of interest were used for reference and ground. Data acquisition and synchronization with the task presentation were accomplished using BCI2000 software (Schalk, 2010; Schalk, McFarland, Hinterberger, Birbaumer, & Wolpaw, 2004). All channels were subsequently downsampled to 1,000 Hz, corrected for DC shifts, and band pass filtered from 0.5 to 200 Hz. Notch filters at 60 Hz, 120 Hz and 180 Hz were used to remove electromagnetic noise. The time series were then visually inspected to remove the intervals containing ictal activity as well as channels that had excessive noise (including broadband electromagnetic noise from hospital equipment or poor contact with the cortical surface). Finally, electrodes were re-referenced to a common average. The high gamma frequency band (70-150 Hz) was extracted using the Hilbert transform.

In addition to the ECoG signals, we acquired the subject's voice through a dynamic microphone (Samson R21s) that was rated for voice recordings (bandwidth 80-12000 Hz, sensitivity 2.24 mV/Pa) and placed within 10 cm of the patient's face. We used a dedicated 16-channel g.USBamp to amplify and digitize the microphone signal in sync with the ECoG data. Finally, we verified the patient's compliance in the covert task using an eye-tracker (Tobii T60, Tobii Sweden).

Experimental paradigms

The recording session included three conditions. In the first condition, text excerpts from historical political speeches or a children's story (i.e., Gettysburg Address (Roy & Basler, 1955), JFK's Inaugural Address (Kennedy, 1961), or Humpty Dumpy ("Mother Goose's Nursery Rhymes," 1867) were visually displayed on the screen moving from right to left at the vertical center of the screen. The rate of scrolling text ranged between 42-76 words/min, and was adjusted based on the subject's attentiveness, cognitive/verbal ability, and comfort prior to experimental recordings. In the first condition, the subject was instructed to read the text aloud

(overt condition). In the second condition, the same text was displayed at the same scrolling rate, but the subject was instructed to read it silently (covert condition). The third condition served as the control and was obtained while the subject was in a resting state condition (baseline control). For each condition, a run lasted between 6 and 8 min, and was repeated 2-3 times depending on the mental and physical condition of the subjects.

Auditory speech representations

We evaluated the predictive power of a neural decoding model based on high gamma signals (see section 2.4 for details) to reconstruct two auditory feature representations: a spectrogram-based and a modulation-based representation. The spectrogram is a time-varying representation of the amplitude envelope at each acoustic frequency. This representation was generated by an affine wavelet transform of the sound pressure waveform using a 128 channel-auditory filter bank mimicking the frequency analysis of the auditory periphery (Taishih Chi, Ru, & Shamma, 2005). The 128 acoustic frequencies of the initial spectrograms were subsequently downsampled to 32 acoustic frequency bins – with logarithmically spaced center frequencies ranging from 180-7,000 Hz.

The modulation representation is based on a non-linear transformation of the spectrogram. Spectral and temporal fluctuations reflect important properties of speech intelligibility. For instance, comprehension is impaired when temporal modulations (<12 Hz) or spectral modulations (4 cycles/kHz) are removed (Elliott & Theunissen, 2009). In addition, low and intermediate temporal modulation rates (< 4 Hz) are linked with syllable rate, whereas fast modulations (> 16 Hz) are related to syllable onsets and offsets. Similarly, broad spectral modulations are associated with vowel formants, whereas narrow spectral modulations are associated with harmonics (Shamma, 2003). The modulation representation was generated by a 2-D affine wavelet transform of the 128 channel auditory spectrogram. The bank of modulation-selective filters spanned a range of spectral scales (0.5–8 cycle/octave) and temporal rates (1–32 Hz), and was estimated from studies of the primary auditory cortex (T Chi, Gao, Guyton, Ru, & Shamma, 1999). The modulation representation was obtained by taking the magnitude of the complex-valued output of the filter bank, and subsequently reduced to 60 modulation features (5 scales x 12 rates) by averaging along the frequency dimension. These operations were computed using the NSL Matlab toolbox (http://www.isr.umd.edu/Labs/NSL/Software.htm). In summary, the neural decoding model predicted 32 spectral frequency features and 60 rate and scale features in the spectrogram-based and modulation-based speech representation, respectively.

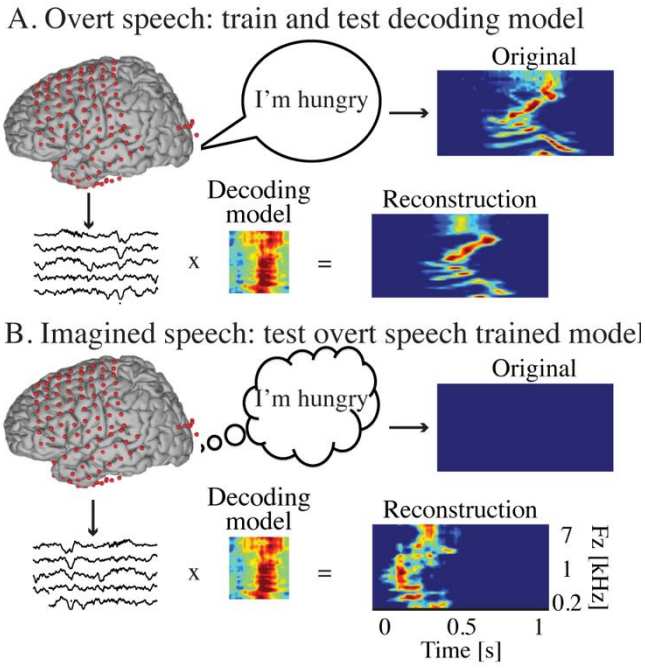


Figure 7 - Decoding approach. ***(A)*** *The overt speech condition was used to train and test the accuracy of a neural-based decoding model to reconstruct spectrotemporal features of speech. The reconstructed patterns were compared to the true original (spoken out loud) speech representation (spectrogram or modulation-based).* ***(B)*** *During covert speech, there is no behavioral output, which prevents building a decoding model directly from covert speech data. Instead, the decoding model trained from the overt speech condition is used to decode covert speech neural activity. The covert speech reconstructed patterns were compared to identical speech segments spoken aloud during the overt speech condition (using dynamic time warping realignment).*

Decoding model and reconstruction procedure

Overt speech decoding

The decoding model was a linear mapping between neural activity and the speech representation (**Figure 7**A). It modeled the speech representation (spectrogram or modulation) as a linear weighted sum of activity at each electrode as follows:

$$\hat{S}(t,p) = \sum_{\tau}\sum_{n} g(\tau,p,n)R(t-\tau,n), \qquad (1)$$

where $R(t-\tau,n)$ is the high gamma activity of electrode n at time $(t-\tau)$, where τ is the time lag ranging between -500ms and 500ms. $\hat{S}(t,p)$ is the estimated speech representation at time t and speech feature p , where p is one of 32 acoustic frequency features in the spectrogram-based representation (**Figure 10**B) or one of 60 modulation features (5 scales x 12 rates) in the modulation-based representation (**Figure 12**B; see section 2.3 for details). Finally, $g(\tau,p,n)$ is the linear transformation matrix, which depends on the time lag, speech feature, and electrode channel. Both speech representations and the neural high gamma response data were synchronized, downsampled to 100 Hz, and standardized to zero mean and unit standard deviation prior to model fitting.

Model parameters, the matrix g described above, were fit using gradient descent with early stopping regularization – an iterative linear regression algorithm. We used a jackknife resampling technique to fit separately between 4 and 7 models (Efron, 1982), and then averaged the parameter estimates to yield the final model. To deal with auto-correlated neural activity and speech features, the data were first divided into 7-second blocks. Then, 90% of the data were randomly partitioned into training set and 10% into testing set. Within the training set, 10% of the data were used to monitor out-of-sample prediction accuracy to determine the early stopping criterion and minimize overfitting. The algorithm was terminated after a series of 30 iterations failing to improve performance. Finally, model prediction accuracy (see section 2.5 for details) was evaluated on the independent testing set. Model fitting was performed using the STRFLab MATLAB toolbox (http://strflab.berkeley.edu/).

Covert speech decoding

Decoding covert speech is complicated by the lack of any measurable behavioral or acoustic output that is synchronized to brain activity. In other words, there is no simple ground truth by which to evaluate the accuracy of the model when a well-defined output is unavailable. To address this, we used the following approach. First, the decoding model was trained using data from the overt speaking condition. Second, the same model (equation 1) was applied to data from the covert condition to predict speech features imagined by the subject (**Figure 7**B), as follows:

$$\hat{S}_{covert}(t,p) = \sum_{\tau}\sum_{n} g(\tau,p,n)R_{covert}(t-\tau,n), \qquad (2)$$

where $\hat{S}_{covert}(\mathrm{t},\mathrm{p})$ is the predicted covert speech representation at time t and speech feature p, and $\mathrm{R}_{covert}(\mathrm{t}-\tau,\mathrm{n})$ is the high gamma neuronal response of electrode n at time $(\mathrm{t}-\tau)$, where τ is the time lag ranging between -500ms and 500ms. Finally, $\mathrm{g}(\tau,\mathrm{p},\mathrm{n})$ is the linear model trained from the overt speech condition. To evaluate prediction accuracy during covert speech, we made the assumption that the covert speech representation should match the spectrotemporal content of overt speech. In this sense, overt speech is used as the "ground truth". Because subjects read the same text segments in both overt and covert conditions, we computed the similarity between the covert reconstructions and the corresponding original speech sounds recorded during the overt condition. To account for timing differences between conditions, we used dynamic time warping to realign the covert reconstruction to the original overt speech sound, as described in the next section.

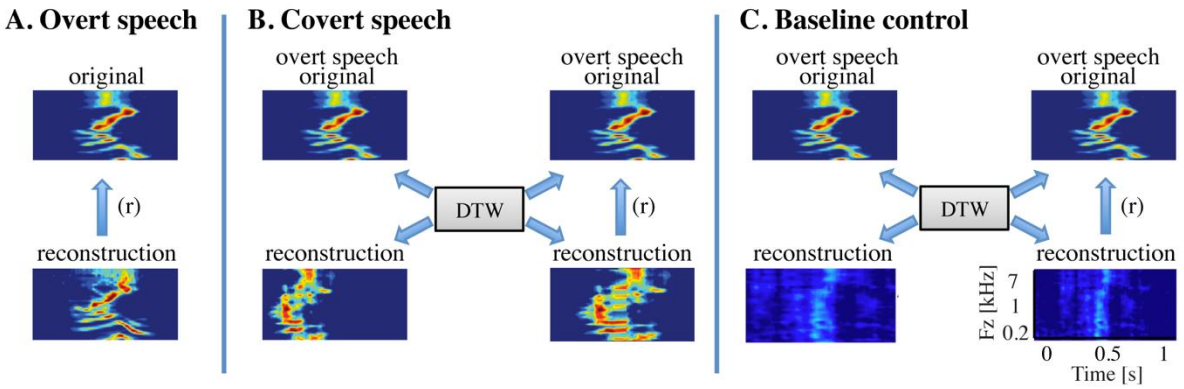


Figure 8 - Speech realignment. ***(A)*** *Overt speech analysis – the overall reconstruction accuracy for the overt speech condition was quantified by computing directly the correlation coefficient (Pearson's r) between the reconstructed and original speech representations* ***(B)*** *Covert speech analysis – the covert speech reconstruction is not necessarily aligned to the corresponding overt speech representation due to speaking rate differences and repetition irregularities. The reconstruction was thus realigned to the overt speech stimuli using dynamic time warping. The overall reconstruction accuracy was then quantified by computing the correlation coefficient (Pearson's r) between the covert speech reconstruction and the original speech representation.* ***(C)*** *Baseline control analysis – a resting state (baseline control) condition was used to assess statistical significance of covert speech reconstruction accuracy. Resting state activity was used to generate a noise reconstruction and dynamic time warping was applied to align the noise reconstruction to overt speech as in (B). Because dynamic time warping has substantial degrees of freedom, due to its ability to stretch and compress speech segments, the overall reconstruction accuracy for the baseline control condition is significantly higher than zero. However, direct statistical comparisons between the covert and baseline conditions are valid as equivalent analysis procedures are applied to both covert and resting state neural data.*

Dynamic time warping

We used a dynamic time warping (DTW) algorithm to realign the covert speech reconstruction with the corresponding spoken audio signal from the overt condition, allowing a direct estimate of the covert reconstruction accuracy (**Figure 8**B). For the overt speech reconstructions, dynamic time warping was not employed (**Figure 8**A), unless otherwise stated. DTW is a standard algorithm used to align two sequences that may vary in time or speed (Sakoe & Chiba, 1978); Giorgino 2009). The idea behind DTW is to find the optimal path through a local similarity matrix d, computed between every pair of elements in the query and template time series, $X \in \mathbb{R}^{P \times N}$ and $Y \in \mathbb{R}^{P \times M}$ as follows:

$$d(n,m) = f(x_n, y_m), \qquad d \in \mathbb{R}^{NxM}, \tag{3}$$

where d is the dissimilarity matrix at time n and m, f can be any distance metric between sequence x and y at time n and m, respectively. In this study, we used the Euclidean distance, defined as $d(n,m) = \sqrt{\sum_p^P (x_{np} - y_{mp})^2}$. Given φ, the average accumulated distortion between both warped signals is defined by:

$$d_\varphi(x,y) = \sum_{k=1}^{K} \frac{d(\varphi_x(k), \varphi_y(k))}{C_\varphi}, \tag{4}$$

where φ_x and φ_y are the warping functions of length K (that remap the time indices of X and Y, respectively), and C_φ is the corresponding normalization constant (in this case N+M), ensuring that the accumulated distortions are comparable along different paths. The optimal warping path φ, chooses the indices of X and Y in order to minimize the overall accumulated distance.

$$D(X,Y) = \min_{\varphi} d_{\varphi}(X,Y), \quad (5)$$

where D is the accumulated distance or global dissimilarity. The alignment was computed using Rabiner-Juan step patterns (type 3; . This step pattern constrained the sets of allowed transitions between matched pairs to:

$$[\varphi_x(k+1) - \varphi_x(k),\ \varphi_y(k+1) - \varphi_y(k)] \in \{(1,2),(2,1),(1,1)\} \quad (6)$$

In addition, we assumed that the temporal offsets between covert speech and original overt speech would be less than 2 sec, and thus introduced a global constraint – the Sakoe-Chiba band window (Sakoe & Chiba, 1978), defined as follows:

$$\left|\varphi_x(k) - \varphi_y(k)\right| \leq T \quad (7)$$

where T = 2 sec was the chosen value that defines the maximum-allowable width of the window. Finally, to reduce computational load, the entire time series was broken into 30 sec segments, and warping was applied on each individual pair of segments (overt, covert, or baseline control reconstruction warped to original speech representation). The warped segments were concatenated and the reconstruction accuracy was defined on the full time series of warped data. The DTW package in R (Giorgino, 2009) was used for all analyses.

Baseline control condition (resting state)

To assess statistical significance of the covert reconstruction accuracy, we applied the same decoding steps (sections 2.4.2 – 2.4.3) to a baseline control condition taken from data recorded during a separate resting state recording session. The overt speech decoding model was applied to neural data from the baseline control, as follows:

$$\hat{S}_{baseline}(t,p) = \sum_{\tau}\sum_{n} g(\tau,p,n) R_{baseline}(t-\tau,n), \quad (8)$$

where $\hat{S}_{baseline}(t,p)$ is the predicted baseline reconstruction at time t and speech feature p, and $R_{baseline}(t-\tau,n)$ is the high gamma neural response during resting state. Finally, $g(\tau,p,n)$ is the linear model trained from the overt speech condition. We also used DTW to realign the baseline control reconstruction with the spoken audio signal from the overt condition, allowing a direct estimate of the control condition decoding predictions (Figure 3.C).

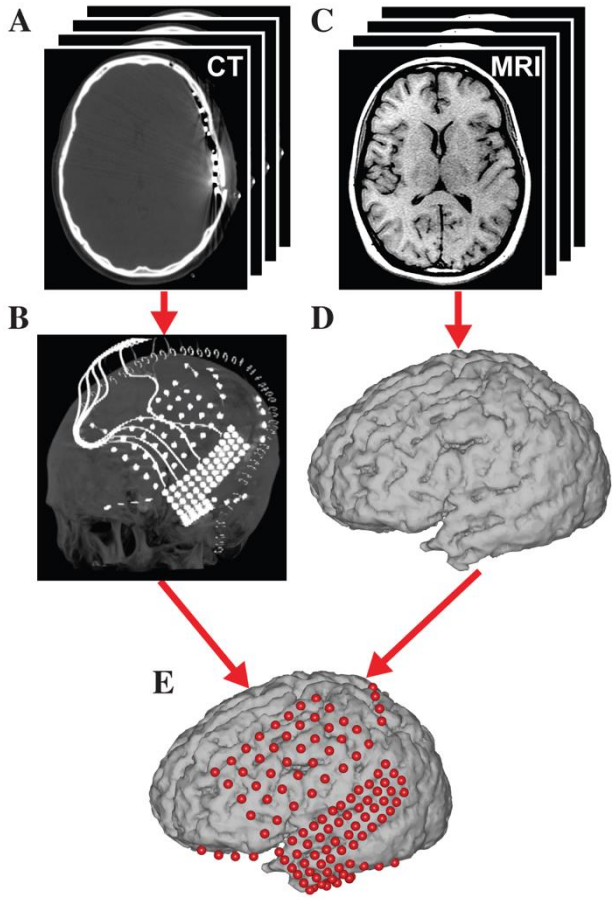


Figure 9 - Brain mapping and electrode localization. ***(A)*** *Post-operative CT scans (1 mm slices) and* ***(C)*** *pre-operative structural MRI scans (1.5 mm slices, T1-weighted) were acquired for each subject. From these scans, grid position* ***(B)*** *and the cortical surface* ***(D)*** *were reconstructed providing a subject-specific anatomical model* ***(E)*** *(see section 2.7 for details).*

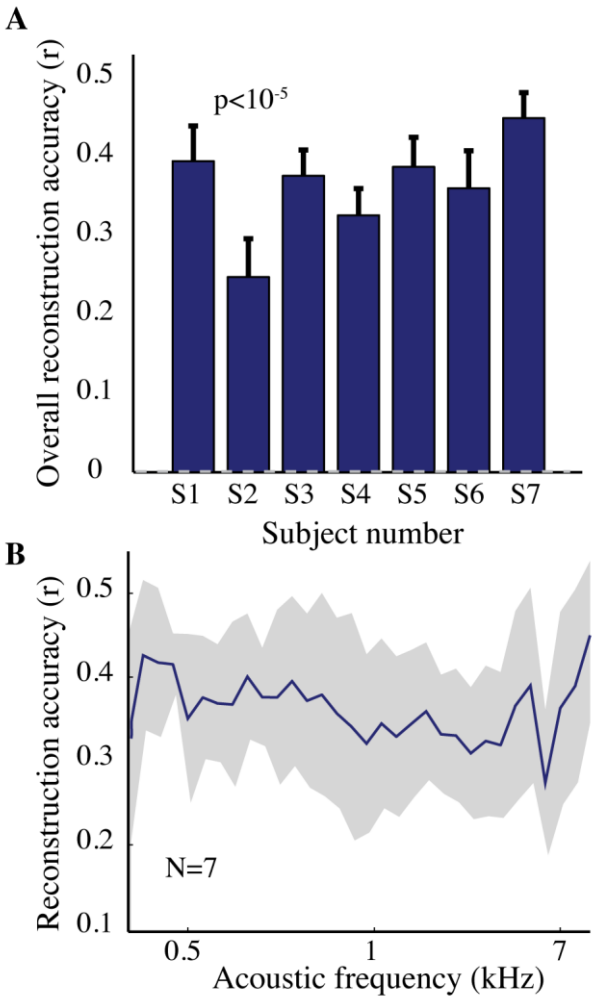


Figure 10 - Overt speech reconstruction accuracy for the spectrogram-based speech representation. (A) *Overall reconstruction accuracy for each subject using the spectrogram-based speech representation. Error bars denote standard error of the mean (SEM). Overall accuracy is reported as the mean over all features (32 acoustic frequencies ranging from 0.2-7 kHz). The overall spectrogram reconstruction accuracy for the overt speech was greater than baseline control reconstruction accuracy in all individuals ($p<10^{-5}$; Hotelling's t-test). Baseline control reconstruction accuracy was not significantly different from zero ($p>0.1$; one-sample t-test; grey dashed line)* ***(B)*** *Reconstruction accuracy as a function of acoustic frequency averaged over all subjects (N=7) using the spectrogram model. Shaded region denotes SEM over subjects.*

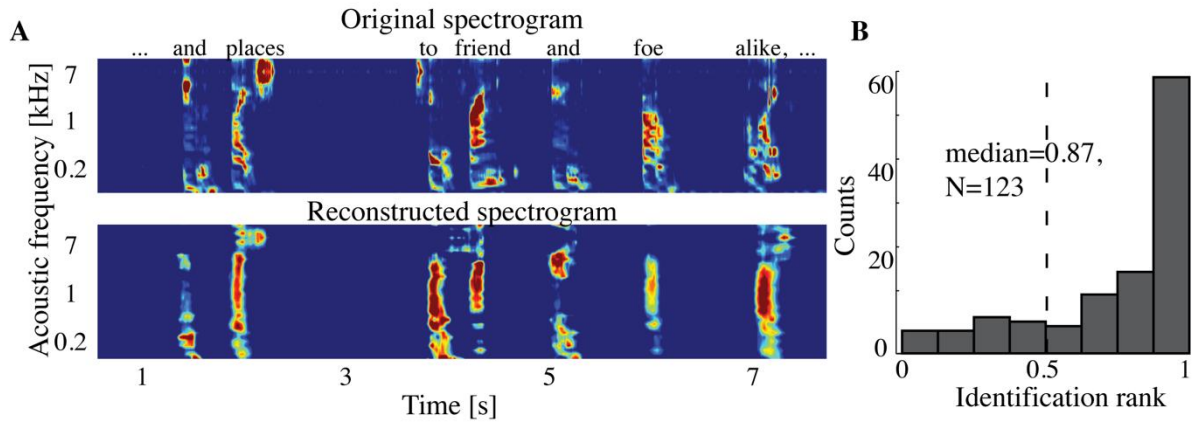


Figure 11 - Overt speech reconstruction and identification. ***(A)*** *Top panel: segment of the original sound spectrogram (subject's own voice), as well as the corresponding text above it. Bottom panel: same segment reconstructed with the decoding model.* ***(B)*** *Identification rank. Speech segments (5 sec) were extracted from the continuous spectrogram. For each extracted segment (N=123) a similarity score (correlation coefficient) was computed between the target reconstruction and each original spectrogram of the candidate set. The similarity scores were sorted and identification rank was quantified as the percentile rank of the correct segment. 1.0 indicates the target reconstruction matched the correct segment out of all candidate segments; 0.0 indicates the target was least similar to the correct segment among all other candidates; (dashed line indicates chance level = 0.5; median identification rank = 0.87; $p<10^{-5}$; randomization test).*

Evaluation

In the overt speech condition, reconstruction accuracy was quantified by computing the correlation coefficient (Pearson's r) between the reconstructed and original speech representation using data from the independent test set. For each cross-validation resample, we calculated one correlation coefficient for each speech feature over time – leading to 32 correlation coefficients (one for each acoustic frequency features) for the spectrogram-based model and 60 correlation coefficients (5 scale x 12 rate features) for the modulation-based model. Overall reconstruction accuracy was reported as the mean correlation over resamples and speech components (32 and 60 for the spectrogram and modulation representation, respectively). Standard error of the mean (SEM) was calculated by taking the standard deviation of the overall reconstruction accuracy across resamples. To assess statistical significance (see section 2.6 for details), overt speech reconstruction accuracy was compared to the accuracy obtained from the baseline control condition (resting state).

In the covert speech condition, we first realigned the reconstructions and original overt speech representations using dynamic time warping (Figure 3.B). Then, we computed the overall reconstruction accuracy using the same procedure as in the overt speech condition. To evaluate statistical significance (see section 2.6 for details), DTW was also applied to the baseline control condition prior to assessing the overall reconstruction accuracy (Figure 3.C).

To further assess the predictive power of the reconstruction process, we evaluated the ability to identify specific blocks of speech utterances within the continuous recording (**Figure 16**).

First, 24-140 segments of speech utterances (5 sec duration) were extracted from the original and reconstructed spectrogram representations. Second, a confusion matrix was constructed where each element contained the similarity score between the target reconstructed segment and the original reference segments from the overt speech spectrogram. To compute the similarity score between each target and reference segment, DTW was applied to temporally align each pair and the mean correlation coefficient was used as the similarity score. The confusion matrix reflects how well a given reconstructed segment matches its corresponding original segment versus other candidates. The similarity scores were sorted, and identification accuracy was quantified as the percentile smaller than the rank of the correct segment (Pasley et al., 2012). At chance level, the expected percentile rank is 0.5, while perfect identification is 1.0.

To define the most informative areas for overt speech decoding accuracy, we isolated for each electrode its corresponding decoding weights, and used the electrode-specific weights to generate a separate reconstruction for each electrode. This allowed calculating a reconstruction accuracy correlation coefficient for each individual electrode. We applied the same procedure to the baseline condition. Baseline reconstruction accuracy was subtracted from the overt values to generate subject-specific informative area maps (**Figure 13**). The same technique was used in the covert speech condition, except that DTW was applied to realign separately each electrode-specific reconstruction to the original overt speech. Similarly, baseline reconstruction accuracy (with DTW realignment) was subtracted from the covert values to define the informative areas (**Figure 17**).

Statistics

To assess statistical significance for the difference between overt speech and baseline control reconstruction accuracy, we used Hotelling's t statistic with a significance level of $p<10^{-5}$. This test accounts for the dependence of the two correlations on the same group (i.e. both correlations are relative to the same original overt speech representation; ; (Birk, 2013). It evaluates whether the correlations between overt speech reconstruction accuracy and baseline reconstruction accuracy differed in magnitude taking into account their intercorrelation, as follows:

$$t = \frac{(r_{jk} - r_{jh})\sqrt{(n-3)(1+r_{kh})}}{\sqrt{2|R|}} \qquad (9)$$

where r_{jk} is the correlation between original overt speech and reconstruction, r_{jh} is the correlation between original overt speech and baseline reconstruction and r_{kh} is the correlation between overt speech reconstruction and baseline reconstruction; df = n – 3 is the effective sample size (Kaneoke et al., 2012) and where

$$|R| = 1 + 2r_{jk}r_{jh}r_{kh} - r_{jk}^2 - r_{jh}^2 - r_{kh}^2 \qquad (10)$$

At the population level (**Figure 10**A), statistical significance was performed using Student's t-tests ($p<10^{-5}$) after first applying Fisher's Z transform to convert the correlation coefficients to a normal distribution (Fisher, 1915).

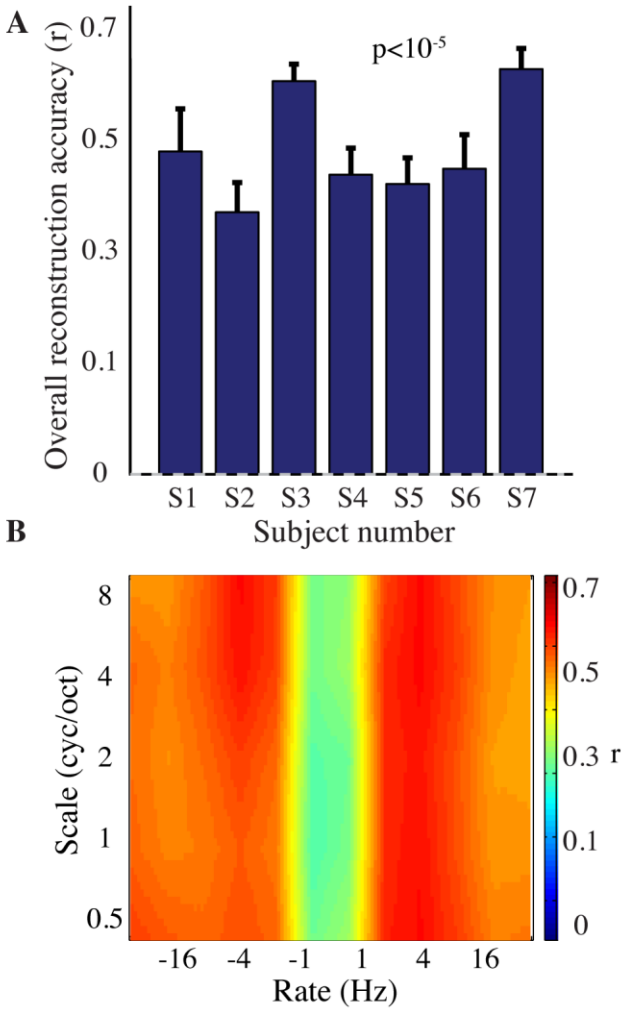


Figure 12 - Overt speech reconstruction accuracy for the modulation-based speech representation. (A) *Overall reconstruction accuracy for each subject using the modulation-based speech representation. Error bars denote SEM. Overall accuracy is reported as the mean over all features (5 spectral and 12 temporal modulations ranging between 0.5-8 cyc/oct and -32-32 Hz, respectively). The overall modulation reconstruction accuracy for the overt speech was greater than baseline control reconstruction accuracy in all individuals ($p<10^{-5}$; Hotelling's t-test). Baseline control reconstruction accuracy was not significantly different from zero ($p>0.1$; one-sample t-test; grey dashed line).* ***(B)*** *Reconstruction accuracy as a function of rate and scale averaged over all subjects (N=7).*

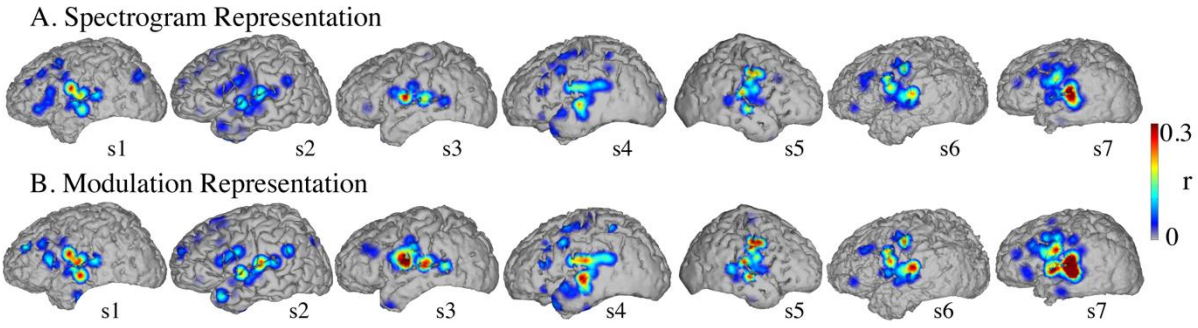


Figure 13 - Overt speech informative areas. *Reconstruction accuracy correlation coefficients were computed separately for each individual electrode and for both overt and baseline control conditions (see section 3.1.3 for details). The plotted correlation values are calculated by subtracting the correlation during baseline control from the overt condition. The informative area map was thresholded to $p<0.05$ (Bonferroni correction)* ***(A)*** *Spectrogram-based reconstruction accuracy* ***(B)*** *modulation-based reconstruction accuracy.*

Test of significance in the covert speech condition was equivalent to the overt condition (equation 9; $p<0.05$; Hotelling's t test), except that the reconstructions and original overt speech representations were first realigned using dynamic time warping. Since DTW induces an artificial increase in correlation by finding an optimal warping path between any two signals (including potential noise signals), this procedure causes the accuracy for baseline reconstruction to exceed zero correlation. However, because the equivalent data processing sequence was applied to both conditions, any statistical differences between the two conditions were due to differences in the neural input signals.

At the population level (**Figure 14**), we directly compared the reconstruction accuracy in all three conditions (overt, covert and baseline control). DTW realignement to the original overt speech was first applied separately for each condition. Reconstruction accuracy was computed as the correlation between the respective realigned pairs. Statistical significance was performed using Fisher's Z transform and one-way ANOVA ($p<10^{-6}$), followed by post hoc t-test ($p<10^{-5}$ for overt speech; $p<0.005$ for covert speech).

For individual subjects, significance of identification rank was computed using a randomization test ($p<10^{-5}$ for overt speech ; $p<0.005$ for covert speech ; $p>0.5$ for baseline control). We shuffled the segment label in the candidate set 10,000 times to generate a null distribution of identification ranks under the hypothesis that there is no relationship between target and reference speech segments. Time-varying speech representations are auto-correlated. To maintain temporal correlations in the data, and preserve the exchangeability of the trial labels, the length of the extracted segments was chosen sufficiently long (5 seconds). The proportion of shuffled ranks greater than the observed rank yields the p-value that the observed accuracy is due to chance. Identification accuracy was assessed for each of the three experimental conditions (overt reconstruction, covert reconstruction, baseline control reconstruction). At the population level, significant identification performance was tested using a one-sided, one-sample t-test ($p<10^{-5}$ for overt speech ; $p<0.05$ for covert speech ; $p>0.5$ for baseline control).

For the informative electrode analysis, statistical significance of overt speech reconstruction was determined relative to the baseline condition using Hotelling's t statistic (equation 9; Hotelling's t test). Electrodes were defined as "informative" if the overt speech reconstruction accuracy was significantly greater than baseline ($p<0.05$; Hotelling's t test with Bonferroni correction). The same procedure was used for covert speech informative areas (equation 9; $p<0.05$; Hotelling's t test with Bonferroni correction), except that DTW was used in both covert speech and baseline control condition.

To investigate possible anatomical differences between overt and covert informative areas, all significant electrodes (either overt, covert or both conditions; $p<0.05$; Bonferroni correction) were selected for an unbalanced two-way ANOVA, with experimental condition (overt and covert) and anatomical region (superior temporal gyrus, pre- and post-central gyrus) as factors. Figure 18 shows significant electrodes in these regions across subjects, co-registered with the Talairach brain template (Lancaster et al., 2000).

Coregistration

Each subject had postoperative anterior–posterior and lateral radiographs (**Figure 9**), as well as computer tomography (CT) scans to verify ECoG grid locations. Three-dimensional cortical models of individual subjects were generated using pre-operative structural magnetic resonance (MR) imaging. These MR images were co-registered with the post-operative CT images using Curry software (Compumedics, Charlotte, NC) to identify electrode locations. Electrode locations were assigned to Brodmann areas using the Talairach Daemon (http://www.talairach.org, (Lancaster et al., 2000). Activation maps computed across subjects were projected on this 3D brain model, and were generated using a custom Matlab program (Gunduz et al., 2012).

Results

Overt Speech

Spectrogram-based reconstruction

The overall spectrogram reconstruction accuracy for overt speech was significantly greater than baseline control reconstruction accuracy in all individual subjects ($p<10^{-5}$; Hotelling's t-test, **Figure 10**A). At the population level, mean overall reconstruction accuracy averaged across all subjects $(N = 7)$ was also significantly higher than baseline control condition ($r=0.41$, $p<10^{-5}$; Fisher's Z transform followed by paired two-sample t-test). The baseline control reconstruction accuracy was not significantly different from zero ($r=0.0$, $p>0.1$; one-sample t-test; dashed line; **Figure 10**A). Group averaged reconstruction accuracy for individual acoustic frequencies ranged between r=~0.25 – 0.5 (**Figure 10**B). An example of a continuous segment of the original and reconstructed spectrogram is depicted for a subject with left hemispheric coverage in **Figure 11**A. In this subject, the reconstruction quality permitted accurate identification of individual decoded speech segments (**Figure 11**B). The median identification rank (0.87, N=123 segments) was significantly greater than chance level (0.5, $p<10^{-5}$; randomization test). Identification performance was significant in each individual subject ($p<10^{-5}$; randomization test). Across all

subjects, identification performance was significant for overt speech reconstruction (**Figure 16**; $rank_{overt}$=0.91 > 0.5, $p<10^{-6}$; one-sided one-sample t-test), whereas the baseline control condition was not significantly greater than chance level ($rank_{baseline}$ = 0.48 > 0.5, p>0.5 one-sided one-sample t-test).

Modulation-based reconstruction

We next evaluated reconstruction accuracy of the modulation representation. The overall reconstruction accuracy was significant in all individual subjects ($p<10^{-5}$; Hotelling's t-test **Figure 12**A). At a population level, mean overall reconstruction accuracy averaged over all patients $(N = 7)$ was also significantly higher than the baseline reconstruction (r=0.55, $p<10^{-5}$; Fisher's Z transform followed by paired two-sample t-test). The baseline control reconstruction accuracy was not significantly different from zero (r=0.02, p>0.1; one-sample t-test; dashed line; **Figure 12**A). Group averaged reconstruction accuracy for individual rate and scale was highest for temporal modulations above 2 Hz (**Figure 12**B).

Informative Areas

Figure 13 shows the significant informative areas (map thresholded at p<0.05; Bonferroni correction), quantified by the electrode-specific reconstruction accuracy (see section 2.5 for details). In both spectrogram and modulation-based representations the most accurate sites for overt speech decoding were localized to the superior temporal gyrus, pre and post central gyrus, consistent with previous spectrogram decoding studies (Pasley et al., 2012).

Covert Speech

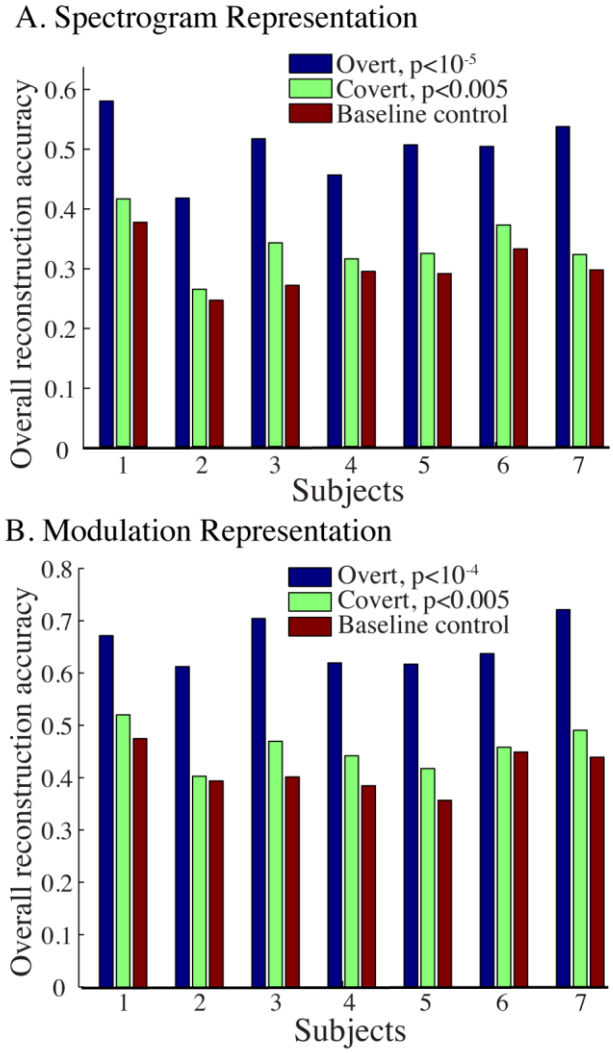


Figure 14 - Overall reconstruction accuracy using dynamic time warping realignment. *Overall reconstruction accuracy for each subject during overt speech, covert speech and baseline control conditions after dynamic time warping realignment.* ***(A)*** *Spectrogram-based representation* ***(B)*** *Modulation-based representation.*

Spectrogram-based reconstruction

Figure 14A shows the overall reconstruction accuracy for overt speech, covert speech, and baseline control after DTW realignment to the original overt speech was applied separately for each condition. The overall reconstruction accuracy for covert speech was significantly higher than the control condition in 5 out of 7 individual subjects ($p<0.05$; Hotelling's t-test; $p>0.05$ for the non-significant subjects). At the population level, there was a significant difference in the overall reconstruction accuracy across the three conditions (overt, covert and baseline control; $F_{(2, 18)} = 35.3$, $p<10^{-6}$; Fisher's Z transform followed by one-way ANOVA). Post-hoc t-tests confirmed that covert speech reconstruction accuracy was significantly lower than overt speech

reconstruction accuracy (r_{covert} = 0.34 < r_{overt} = 0.50, $p<10^{-5}$; Fisher's Z transform followed by paired two-sample t-test), but higher than the baseline control condition (r_{covert} = 0.34 > $r_{baseline}$ = 0.30, p<0.005; Fisher's Z transform followed by a paired two-sample t-test). **Figure 15**A illustrates a segment of the reconstructed covert speech spectrogram and its corresponding overt segment (realigned with DTW). We next evaluated identification performance (N=123 segments) for covert speech and baseline control conditions in this subject (**Figure 15**B). In the covert speech condition, the median identification rank equaled 0.62, and was significantly higher than chance level of 0.5 (p<0.005; randomization test), whereas the baseline control condition was not significant (median identification rank = 0.47, p>0.5; randomization test). Several of the remaining subjects exhibited a trend toward higher identification performance, but were not significant at the p<0.05 level (**Figure 16**; randomization test). At the population level, mean identification performance across all subjects was significantly greater than chance for the covert condition ($rank_{covert}$ = 0.55 > 0.5, p<0.05; one-sided one-sample t-test), and not significant for the baseline control ($rank_{baseline}$ = 0.48 > 0.5, p>0.5; one-sided one-sample t-test). These results provide preliminary evidence that neural activity during auditory speech imagery can be used to decode spectrotemporal features of covert speech.

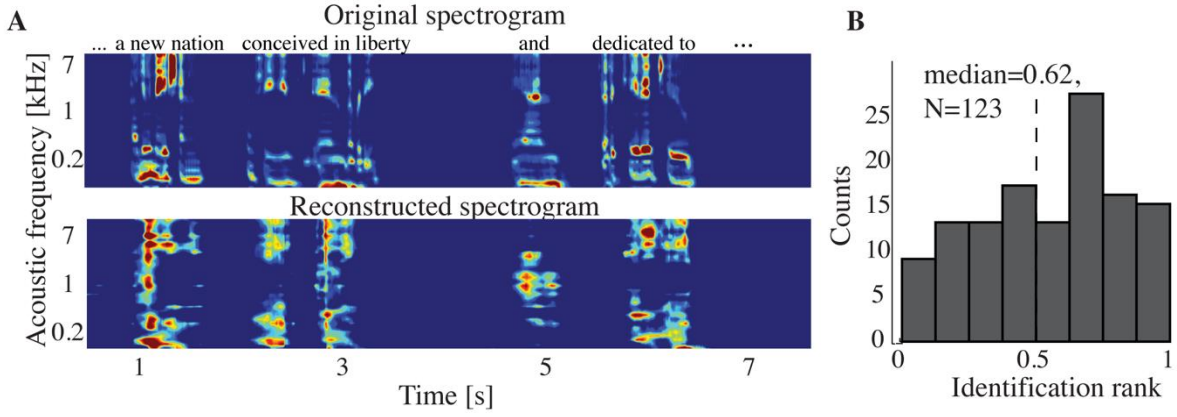


Figure 15 - Covert speech reconstruction. ***(A)*** *Top panel: a segment of the overt (spoken out loud) spectrogram representation. Bottom panel: the same segment reconstructed from neural activity during the covert condition using the decoding model.* ***(B)*** *Identification rank. Speech segments (5 sec) were extracted from the continuous spectrogram. For each target segment (N=123) a similarity score (correlation coefficient) was computed between the target reconstruction and each original spectrogram in the candidate set. The similarity scores were sorted and identification rank was quantified as the percentile rank of the correct segment. 1.0 indicates the target reconstruction matched the correct segment out of all candidate segments; 0.0 indicates the target was least similar to the correct segment among all other candidates. (dashed line indicates chance level = 0.5; median identification rank = 0.62; p<0.005; randomization test).*

Modulation-based reconstruction

Reconstruction accuracy for the modulation-based covert speech condition was significant in 4 out of 7 individuals (p<0.05; Hotelling's t-test; p>0.1 for non-significant subjects; **Figure 14**B). At the population level, the overall reconstruction accuracy across the three conditions (overt, covert and baseline control) was significantly different ($F_{(2, 18)}$ = 62.1, $p<10^{-6}$; one-way ANOVA). Post-hoc t-tests confirmed that covert speech reconstruction accuracy was significantly lower

than overt speech reconstruction accuracy (r_{covert} = 0.46 < r_{overt} = 0.66, $p<10^{-5}$; Fisher's Z transform followed by a paired two-sample t-test), but higher than the baseline control condition (r_{covert} = 0.46 > $r_{baseline}$ = 0.42, p<0.005; Fisher's Z transform followed by a paired two-sample t-test).

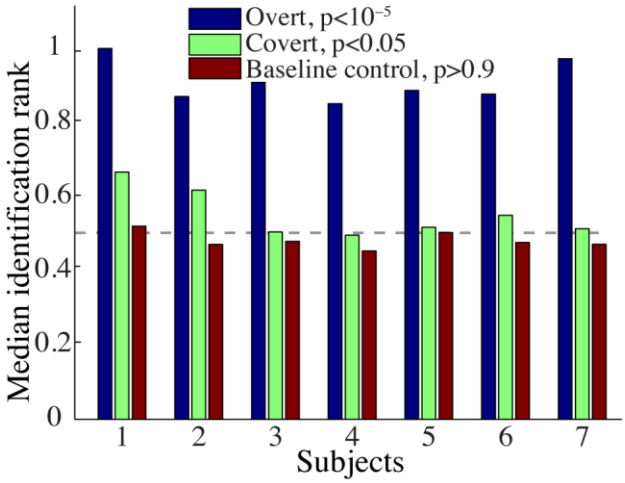


Figure 16 - Overt and covert speech identification. *Median identification rank for each subject during overt speech, covert speech and baseline control conditions (see section 2.5 for more details). At the group level, $rank_{overt}$ = 0.91 and $rank_{covert}$ = 0.55 are significantly higher than chance level (0.5; randomization; grey dashed line), whereas $rank_{baseline}$=0.48 is not significantly different.*

Informative Areas

Significant informative areas (map thresholded at p<0.05; Bonferroni correction), quantified by the electrode-specific reconstruction accuracy (see section 2.5 for details) are shown in **Figure 17**. As observed in the overt condition, brain areas involved in covert spectrotemporal decoding were also concentrated around STG, pre and post central gyri.

Anatomical differences between overt and covert informative areas were assessed for significant electrodes (either overt, covert or both conditions; p<0.05; Bonferroni correction), using an unbalanced two-way ANOVA, with experimental condition (overt and covert speech) and anatomical region (superior temporal gyrus, pre- and post-central gyrus) as factors. Figure 18 shows significant electrodes across subject, co-registered with the Talairach brain template (Lancaster et al., 2000). The main effect of experimental condition was significant for the spectrogram-based ($F_{(1, 116)}$ =19.6, $p<10^{-6}$) and modulation-based reconstructions ($F_{(1, 156)}$ =16.9, $p<10^{-4}$), indicating that the magnitude of reconstruction accuracy for overt speech (spectrogram: mean difference with baseline (r) = 0.06; modulation: mean difference = 0.1) was higher than for covert speech (spectrogram: mean difference = 0.006; modulation: mean difference = 0.01) at the level of single electrodes. The main effect of anatomical region was also significant (spectrogram: $F_{(2, 116)}$ =3.22, p<0.05, and modulation: $F_{(2, 156)}$ =3.4, p<0.05). However, post hoc t-tests with Bonferroni correction indicated no differences in accuracy at the level of p=0.05: STG (spectrogram: mean difference = 0.05; modulation: mean difference = 0.07), pre (spectrogram: mean difference = 0.02; modulation: mean difference = 0.05), and

post-central gyrus (spectrogram: mean difference = 0.02; modulation: mean difference = 0.01). The interaction between gyrus and experimental condition was significant for the modulation-based reconstruction ($F_{(2, 156)}$ =3.6, p<0.05) and marginally significant for the spectrogram ($F_{(2, 116)}$ =2.92, p=0.058). In the modulation representation, the overt condition resulted in significantly higher accuracy than the covert condition for the STG (mean difference = 0.12; $p<10^{-5}$), but not for the pre-central (mean difference = 0.06; p>0.05) or the post-central gyrus (mean difference = 0.02; p>0.05). This suggests that STG is the cortical area where the spectrotemporal representations of overt and covert speech have the largest absolute difference in reconstruction accuracy. Understanding the differences in the neural representations of overt and covert speech within STG is therefore a key question toward improving the spectrotemporal decoding accuracy of covert speech.

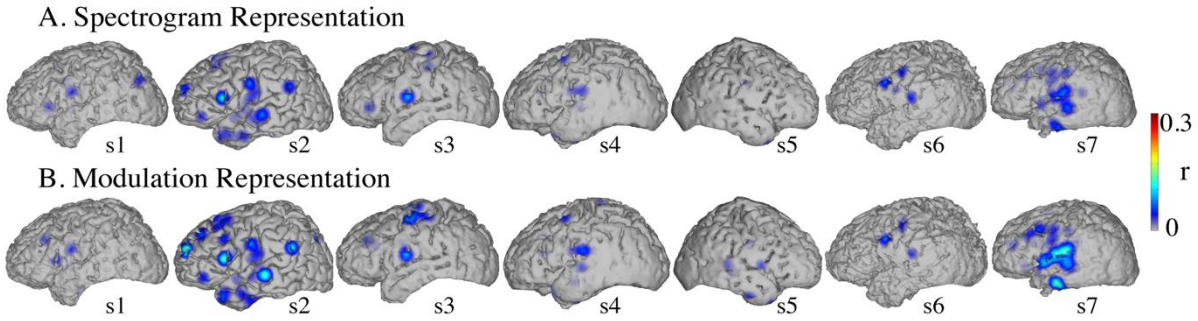


Figure 17 - Covert speech informative areas. Reconstruction accuracy correlation coefficients were computed separately for each individual electrode and for both covert and baseline control conditions (see section 3.1.3 and 3.2.3 for details). The plotted correlation values are calculated by subtracting the correlation during baseline control from the covert condition. The informative area map was thresholded to p<0.05 (Bonferroni correction)**(A)** Spectrogram-based reconstruction accuracy **(B)** modulation-based reconstruction accuracy.

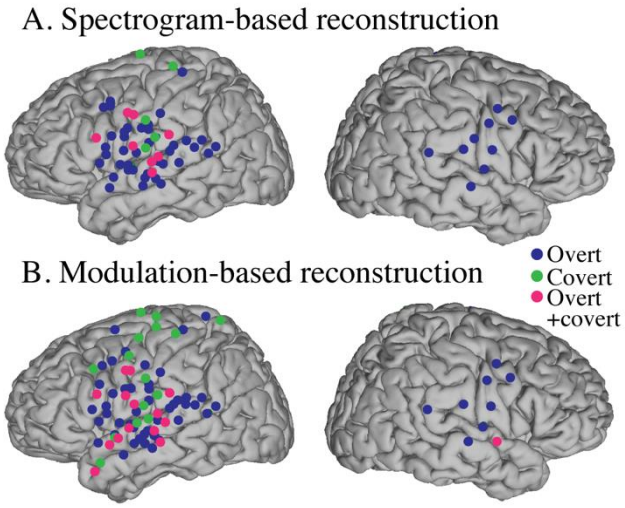


Figure 18 - ***Region of interest analysis of significant electrodes.*** *Significant electrodes (either overt, covert or both; p<0.05; Bonferroni correction) in STG, Pre- and Post-central gyrus across subjects, co-registered with the Talairach brain template (Lancaster et al., 2000).*

Discussion

We evaluated a method to reconstruct overt and covert speech from direct intracranial brain recordings. Our approach was first to build a neural decoding model from self-generated overt speech, and then to evaluate whether this same model could reconstruct speech features in the covert speech condition at a level of accuracy higher than expected by chance. This technique provided a quantitative comparison of the similarity between auditory perception and imagery in terms of neural representations based on acoustic frequency and modulation content. Our results indicated that auditory features of covert speech could be decoded from models trained from an overt speech condition, providing evidence of a shared neural substrate for overt and covert speech. However, comparison of reconstruction accuracy in the two conditions also revealed important differences between overt and covert speech spectrotemporal representation. The predictive power during overt speech was higher compared to covert speech and this difference was largest in STG sites consistent with previous findings of a partial overlap of the two neural representations (Geva, Jones, et al., 2011; Huang et al., 2002; Pei et al., 2011; Shuster & Lemieux, 2005). In addition, we compared the quality of the reconstructions by assessing how well they could be identified. The quality of overt speech reconstruction allowed a highly significant identification, while in the covert speech condition, the identification was only marginally significant. These results provide evidence that continuous features of covert speech can be extracted and decoded from ECoG signals, providing a basis for development of a brain-based communication method for patients with disabling neurological conditions.

Previous research demonstrated that continuous spectrotemporal features of auditory stimuli could be reconstructed using a high gamma neural-based decoder (Pasley et al., 2012). In this study, we analyzed auditory stimuli from self-generated speech as opposed to external auditory stimulation. During self-produced speech, neural activity in human auditory cortex is reported

to be suppressed (Creutzfeldt, Ojemann, & Lettich, 1989) (Flinker et al., 2010) which has been attributed to the effect of efference copy or corollary discharge sent from the motor cortex onto sensory areas (Jeannerod, 2003). Despite this effect, we observed that high gamma activity in the superior temporal gyrus, pre- and post-central gyrus during vocalization was sufficient to reliably reconstruct continuous spectrotemporal auditory features of speech.

There is accumulating evidence that imagery and perception share similar neural representations in overlapping cortical regions (Palmer *et al.* 2001; Yetkin *et al.* 1995; Rosen *et al.* 2000; Aziz-Zadeh *et al.* 2005; Cichy *et al.* 2012; Geva et al., 2011c) . It has been proposed that an efference copy is generated from the motor cortex through motor simulation and sent to sensory cortices enabling a comparison of actual with desired movement and permitting online behavioral adjustments (Jeannerod 2003). Similar accounts have been proposed in speech processing (Guenther et al., 2009; Hickok, 2001; Price, 2011; Tian & Poeppel, 2012). Higher order brain areas internally induce lower level sensory cortices activation, even in the absence of actual motor output (covert). The anatomical results reported here are in agreement with these models. The relationship between overt and covert speech reconstruction depended on anatomy. High gamma activity in the superior temporal gyrus, pre- and post-central gyrus provided the highest information to decode both spectrogram and modulation features of overt and covert speech. However, the predictive power for covert speech was weaker than for overt speech. This is in accordance with previous research showing that the magnitude of activation was greater in overt than in covert speech in some perisylvian regions (Palmer et al., 2001; Partovi et al., 2012; Pei et al., 2011) possibly reflecting a lower signal-to-noise ratio (SNR) for HG activity during covert speech. Future work is needed to determine the relative contributions of SNR vs. differences in the underlying neural representations to account for discrepancies between overt and covert speech reconstruction accuracy.

A key test of reconstruction accuracy is the ability to use the reconstruction to identify specific speech utterances. At the group level, using covert reconstructions, identification performance was significant, but at a weaker level ($p=0.032$) than overt speech identification ($p<10^{-4}$). At the individual level, covert speech reconstruction in one subject (out of seven) was accurate enough to identify speech utterances better than chance level. This highlights the difficulty in applying a model derived from overt speech data to decode covert speech. This also indicates that the spectrotemporal neural mechanisms of overt and covert speech are partly different, in agreement with previous literature (Aleman, 2004; Basho, Palmer, Rubio, Wulfeck, & Müller, 2007; Pei et al., 2011; Shuster & Lemieux, 2005). Despite these difficulties, it is possible that decoding accuracy may be improved by several factors. First, a major difficulty in this approach is the alignment of covert speech reconstructions to a reference speech segment. Variability in speaking rate, pronunciation, and speech errors can result in suboptimal alignments that may be improved by better alignment algorithms or by more advanced automatic speech recognition techniques (e.g., Hidden Markov Models). Second, a better scientific understanding of the differences between overt and covert speech representations may provide insight into how the decoding model can be improved to better model covert speech neural data. For example, the current study uses a simple model that assumes the auditory representation of covert speech imagery is equivalent to that of overt speech. If systematic differences in

spectrotemporal encoding can be identified during covert speech, then the spectrotemporal tuning of the decoding model can be biased to reflect these differences in order to optimize the model for covert speech data. Further investigation of the differences in overt and covert spectrotemporal neural representation offers a promising avenue for improving covert speech decoding.

Chapter 4 – Encoding models reveal tuning plasticity in human auditory cortex

Introduction

In this chapter we move from the world of *decoding* towards *encoding*. This is a process by which we attempt to uncover the subset of stimulus features that elicit patterns of recorded neural activity. It is possible to use this approach to describe the tuning properties of neural activity. This chapter describes an attempt to use this encoding modeling framework in electrocorticography electrodes for the purposes of studying tuning plasticity in response to degraded speech.

Citation

Holdgraf, C. R., de Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J. J., Knight, R.T., Theunissen, F.E. (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. Nat. Commun. 7, 13654. doi:10.1038/ncomms13654.

Abstract

Experience shapes our perception of the world on a moment-to-moment basis. This robust perceptual effect parallels a change in the neural representation of stimulus features, though the nature of this representation and its plasticity are not well-understood. Spectrotemporal receptive field (STRF) mapping describes the neural response to acoustic features, and has been used to study contextual effects on auditory receptive fields in animal models. We performed a STRF plasticity analysis on electrophysiological data from recordings obtained directly from the human auditory cortex. Here we report rapid, automatic plasticity of the spectrotemporal response of recorded neural ensembles, driven by previous experience with acoustic and linguistic information and with a neurophysiological effect in the sub-second range. This plasticity reflects increased sensitivity to spectrotemporal features, enhancing the extraction of more speech-like features from a degraded stimulus and providing the physiological basis for the observed “perceptual enhancement” in understanding speech.

Introduction

Auditory perception encompasses a sequence of feature extraction steps, with increasingly complex acoustic features extracted at each stage of neural processing (Eggermont, 2001; Theunissen & Elie, 2014). Auditory neuroscientists have used synthetic and natural sounds as stimuli while recording the neural activity of single auditory neurons to investigate the nature of these computations. This research has led to an understanding of cortical auditory processing as a modulation filter bank (Chi, Ru, & Shamma, 2005). At the level of auditory cortex, sounds are decomposed not only in frequency channels (as in the auditory periphery) but also in terms of join spectral and temporal modulations. The filters in this modulation filter bank are the neurons' spectro-temporal receptive fields or STRFs (Depireux, Simon, Klein, & Shamma, 2001; L. M. Miller, Escabí, Read, & Schreiner, 2002; Theunissen, Sen, & Doupe, 2000). The decomposition of sounds into a modulation filter bank facilitates many tasks, including the discrimination of speech from non-speech (Mesgarani, Slaney, & Shamma, 2006) and the extraction of communication signals from noise (Moore, Lee, & Theunissen, 2013).

Several studies have examined a STRF based feature representation at different levels of the auditory hierarchy (Atencio, Sharpee, & Schreiner, 2012; L. M. Miller et al., 2002; Woolley, Fremouw, Hsu, & Theunissen, 2005), but it is less understood if and how these representations interact with each other. For example, the presence of a higher-level response (such as the recognition of task-relevant stimuli) may alter the way that stimulus features are represented at lower levels in the auditory processing stream (Gilbert & Sigman, 2007). It has been shown that the tuning of auditory neurons change during behavioral tasks (Fritz, Shamma, Elhilali, & Klein, 2003; Rabinowitz, Willmore, King, & Schnupp, 2013; Rabinowitz, Willmore, Schnupp, & King, 2011; Shamma & Fritz, 2014), revealing that the STRFs describing this tuning are plastic. Further, neuroanatomical (Atiani et al., 2014; Coull, Frith, Büchel, & Nobre, 2000; Davis & Johnsrude, 2007) and neurophysiological (David, Fritz, & Shamma, 2012; Yin, Fritz, & Shamma, 2014) research have highlighted the importance of top-down mechanisms for inducing such task-dependent STRF plasticity. These results were all obtained from single-unit recordings in animal models, and top-down manipulation was generally modulated with active attentional manipulations or task-relevant demands.

Human speech perception is another area in which top-down and bottom-up mechanisms are in constant interplay (Block & Siegel, 2013; Cusack, Deeks, Aikman, & Carlyon, 2004; Schroeder, Wilson, Radman, Scharfman, & Lakatos, 2010). The act of understanding speech requires that auditory information entering the auditory periphery is interpreted through the lens of previous experience with natural sounds and language. It is assumed that this experience plays a role in shaping the response to speech in the cortex. Recent research using human electrophysiology has shown that experience with sound or contextual information about its content corresponds to differing patterns of low-frequency activity in both auditory and premotor cortex. For example, activity in the theta band of neural signals is reported to track the temporal structure in the speech envelope (Fontolan, Morillon, Liegeois-Chauvel, & Giraud, 2014; Giraud & Poeppel, 2012; Gross et al., 2013) and this tracking increases as noise levels are decreased in the speech stimulus (Peelle, Gross, & Davis, 2013). In addition, power in theta and beta frequency bands have been implicated in top-down processing during speech perception

(Fontolan et al., 2014). It has been suggested that these signals reflect the brain's attempt to find relevant information in the speech signal, and to filter out noise or competing auditory streams (Lakatos et al., 2013). While these approaches delineate differing patterns of neural activity that reflect top-down processes, they do not quantify changes in the spectro-temporal tuning of cortical activity, a feature representation that is believed to be encoded in auditory cortical neurons.

To investigate how contextual effects modulate auditory cortical activity, it is necessary to investigate the feature representations that are encoded in auditory brain areas. STRF models have been used as a standard for characterizing the tuning of neurons in primary auditory cortex (Theunissen & Elie, 2014). Recent research has shown that STRF modeling may be applied to human electrocorticography (ECoG) in order to characterize the spectrotemporal tuning of electrodes in response to speech (Hullett, Hamilton, Mesgarani, Schreiner, & Chang, 2016; Martin et al., 2014; Pasley et al., 2012) and to investigate plasticity in the auditory cortical response (Mesgarani & Chang, 2012). In particular, the high-frequency broadband (HFB; 70-150 Hz) component recorded with ECoG has both the spatial resolution to localize activity to discrete regions of the brain, and the temporal resolution to resolve the fine-grained pattern of acoustic features. HFB is believed to reflect local cortical activity typically obtained with 4-10 mm electrode spacing (Wodlinger, Degenhart, Collinger, Tyler-Kabara, & Wang, 2011). HFB activity is felt to represent a broadband increase in power, most readily detected in frequencies centered from 70-150Hz (K. J. Miller, Zanos, Fetz, den Nijs, & Ojemann, 2009). This permits using HFB activity in ECoG to study the representation of spectro-temporal speech features in human auditory cortex and investigate how this representation changes during language processing.

Here, we perform a passive listening speech task in electrocorticography subjects. In this task, subjects hear degraded speech before and after experience with an unfiltered speech context. We first document that perceptual enhancement to the degraded sound is boosted after experience with the unfiltered speech, enabling speech comprehension. We then use STRF modeling techniques to investigate if this perceptual enhancement coincides with a shift in auditory cortical tuning to spectrotemporal speech features. We use regularized regression techniques to estimate a STRF for each recoding electrode. It is estimated that the HFB activity of a single electrode reflects the activity of hundreds of thousands of neurons (Crone, Korzeniewska, & Franaszczuk, 2011; Ray, Crone, Niebur, Franaszczuk, & Hsiao, 2008). Thus, we effectively calculate an ensemble spectrotemporal receptive field, which we refer to as an eSTRF. We subsequently use this acronym to explicitly distinguish our results from those obtained with single auditory units. We show that providing an unfiltered speech context prior to a degraded speech stimulus causes an automatic, rapid shift in auditory cortical eSTRFs, enhancing their sensitivity to speech features. These findings provide evidence of an automatic mechanism in which experience with a contextually appropriate speech sentence causes behavioral perceptual enhancement for subsequent degraded speech signals, along with a tuning shift towards speech-specific spectrotemporal auditory features in auditory cortical areas.

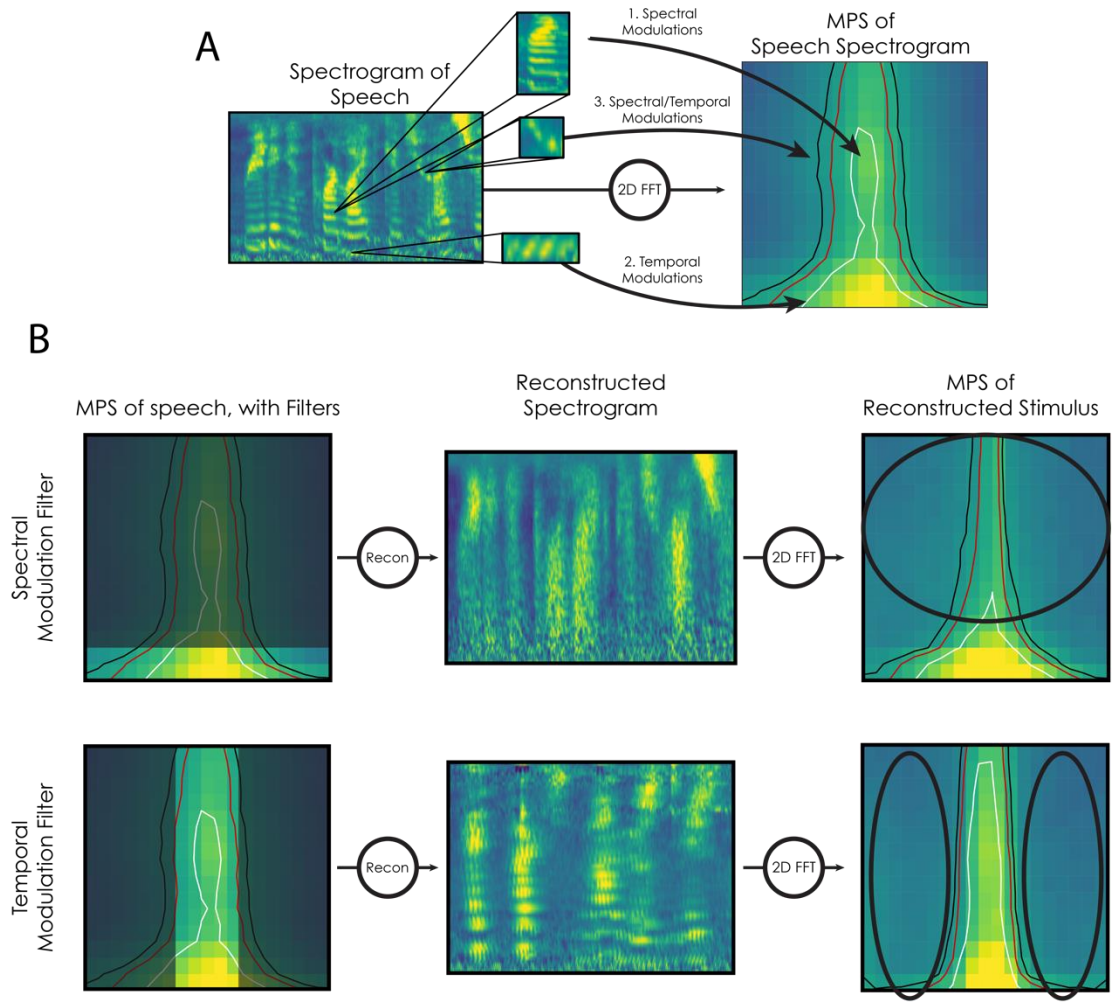


Figure 19 - Stimulus Creation By Filtering Modulation Power Spectrum
(A) The modulation power spectrum describes the oscillatory patterns present in a time-frequency representation of sound. Left, the spectrogram of unfiltered speech is shown. Right, the modulation power spectrum (MPS, calculated from a 2-D FFT) is shown. Patterns in the spectrogram are reflected as power in temporal or spectral axes of the MPS. Rapid spectral fluctuations (e.g., harmonic stacks from pitch, 1) are represented near the middle/top of the MPS. Rapid temporal fluctuations (e.g., plosives, 2) are represented near the bottom/sides of the MPS. Joint spectral/temporal fluctuations (e.g., rising pitch and phoneme changes, 3) are represented in the upper corners of the MPS. (B) Left column: Filtered speech was created by filtering either spectral (top) or temporal (bottom) regions of the MPS space. MIDDLE column: spectrograms of the resulting filtered speech is shown. Right column: re-calculating the MPS on the filtered speech spectrogram shows that the MPS is now lacking power in the filtered regions.

Results

ECoG Behavioral Task

A passive listening filtered-speech task was used to study the neural response to degraded speech before and after hearing an unfiltered speech context. Filtered speech stimuli were created by filtering out portions of the Modulation Power Spectrum (MPS) of each sentence (see Figure 1, Methods and Supplementary Audio File 1) with low-pass filters. The corner frequency of each filter was chosen to render speech unintelligible by removing key spectral or temporal modulations (Elliott & Theunissen, 2009). Electrocorticography subjects (n=7) heard a filtered version of a speech utterance (hereafter described as the BEFORE condition), followed by an unfiltered version of the sound (MIDDLE condition), and finally by a repetition of the filtered version (AFTER condition). The first filtered speech presentation is incomprehensible, while the second filtered speech presentation is understandable due to experience with the unfiltered speech context. See **Figure 20** for a description of task design.

Behavioral Control Study

Due to limitations of the ECoG recording environment, it was not possible to obtain behavioral response data from ECoG patients, and a separate task was performed on control subjects to validate and quantify the perceptual effects generated in our stimuli sequences. In one task, subjects heard a single filtered version of each stimulus (with no unfiltered speech context), and were asked to type any words that they understood. The percent correct was calculated for each sentence. Without any unfiltered speech context, subjects recognized 3.5 +/- 0.4% of filtered speech words, replicating previous studies with the same filtering technique (Elliott & Theunissen, 2009). In a second task, subjects listened to filtered speech sentences along with a number of different context sentences, mimicking the "filtered-unfiltered-filtered" structure of the behavioral task that the patients performed. Subjects typed out any words that they understood after the second presentation of the filtered speech sentence. Without any context, subjects understood 4.53 +/- .82% words. When they were given a contextual sentence that was different from the filtered speech sentence, subjects understood 10.5 +/- 1.3% words, representing the perceptual enhancement due to stimulus repetition or general activation of auditory streams involved in the processing of intact speech. When the contextual sentence was the same sentence as the filtered speech, subjects understood 77.7 +/- 1.5% of words. As such, there was a roughly 67.2% increase in comprehension relative to hearing a different contextual sentence, representing the perceptual enhancement we focus on in this paper (two-sample t-test, p=1e-5, df=16). This perceptual enhancement reflects multiple speech processes, including the recent activation of the auditory stream in response to clean speech (as in the different sentence case) as well as the activation of cognitive areas involved in language processing resulting from speech comprehension (see **Figure 20** and Supplementary Figure 1).

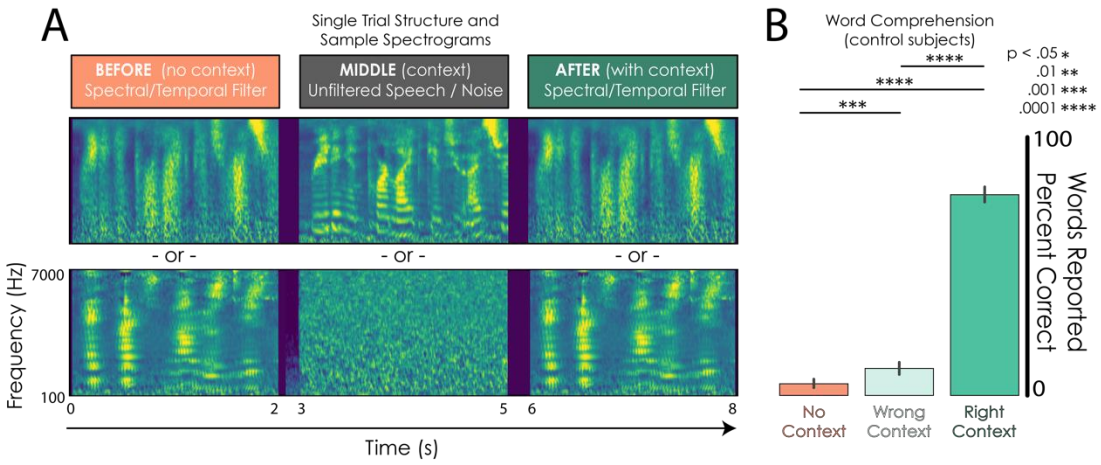


Figure 20 - Behavioral task and speech intelligibility results *(A) Trials consisted of three steps: BEFORE, MIDDLE, AFTER. In the first step called BEFORE (left column), subjects heard a filtered speech stimulus that lacked the key modulations for speech intelligibility. Stimuli were filtered either with a spectral modulation filter (top), removing spectral envelope modulations above 0.5 cycles/kHz or a temporal modulation filter (bottom), removing temporal envelope modulations above 3 Hz (see methods). In the second step called MIDDLE (center column), subjects heard the unfiltered version of the spoken sentence. A subset of 3 subjects had a 50% chance of hearing either the unfiltered version or pink noise with a matched frequency power spectrum. In the third step called AFTER (right column), the same filtered speech stimulus was repeated. Subjects attended to a fixation cross presented during each stimulus and passively listened to the presented sounds. (B) In a separate behavioral task, non-clinical subjects were asked to type any words they heard after the first filtered speech presentation (BEFORE and here labelled No Context), after a filtered speech sentence that followed a different unfiltered sentence (AFTER with Wrong Context), or after a filtered speech sentence that followed the matching unfiltered sentence (AFTER with Right Context). Mean +/- standard error % words correct is shown. More details and results obtained using other contextual stimuli to further explore the stimulus information required for the perceptual enhancement can be found in Supplementary Figure 1 and methods.*

High-Frequency Broadband Activity

All analyses in this study were based on the on high-frequency broadband activity (HFB; 70-150Hz) of Electrocorticographic (ECoG) recordings. This HFB signal is characterized by increase in power across a large range of frequencies, and reflects local neuronal firing within ~2mm of each electrode, representing the combined activity of ~500,000 cortical neurons (Crone et al., 2011; Wodlinger et al., 2011). HFB can provide low-noise single-trial evoked responses (see **Figure 24**, as well as Supplementary Movie 1 and 2) that has been used for speech decoding and encoding models in humans (Mesgarani & Chang, 2012; Pasley et al., 2012), making it a good candidate for STRF modeling (see Supplementary Figure 4).

To define speech-selective electrodes, the mean post-stimulus HFB activity was first calculated for every speech trial. For each electrode, we used standard bootstrapping methods to calculate a bootstrap distribution of its mean evoked HFB activity across trials. The .5th percentile of this distribution was then calculated as a lower bound on post-stimulus activity

(corresponding to a 99% confidence interval). This process was repeated for each electrode, and electrodes with a lower bound greater than 0 were defined as speech selective. A subset of electrodes in each subject had significant responses to speech stimuli over baseline (confidence interval test across trials, see **Figure 21**), generally centered around perisylvian regions. These electrodes are subsequently called speech-responsive (Speech-R) and made up 92 of 468 total electrodes (19.6%, see Supplementary Figure 2).

For all Speech-R electrodes on temporal and perisylvian cortex, the mean difference in HFB activity between the BEFORE and AFTER conditions was estimated. There was a significant increase in HFB activity in the AFTER condition (cluster-based permutation test, p=.003, see Figure 2B). This increase in activity could reflect sentence independent changes in arousal (e.g., increased HFB activity to any auditory stimuli), or changes due to the activation of speech and language network resulting in a shift in gain or tuning of speech features in the degraded signal. However, only changes in tuning would lead to sentence specific (and in our experimental paradigm, trial by trial) effects. In subjects that also had pink noise control trials, there was no difference in evoked HFB activity between the AFTER and BEFORE conditions (see Supplementary Figure 3A).

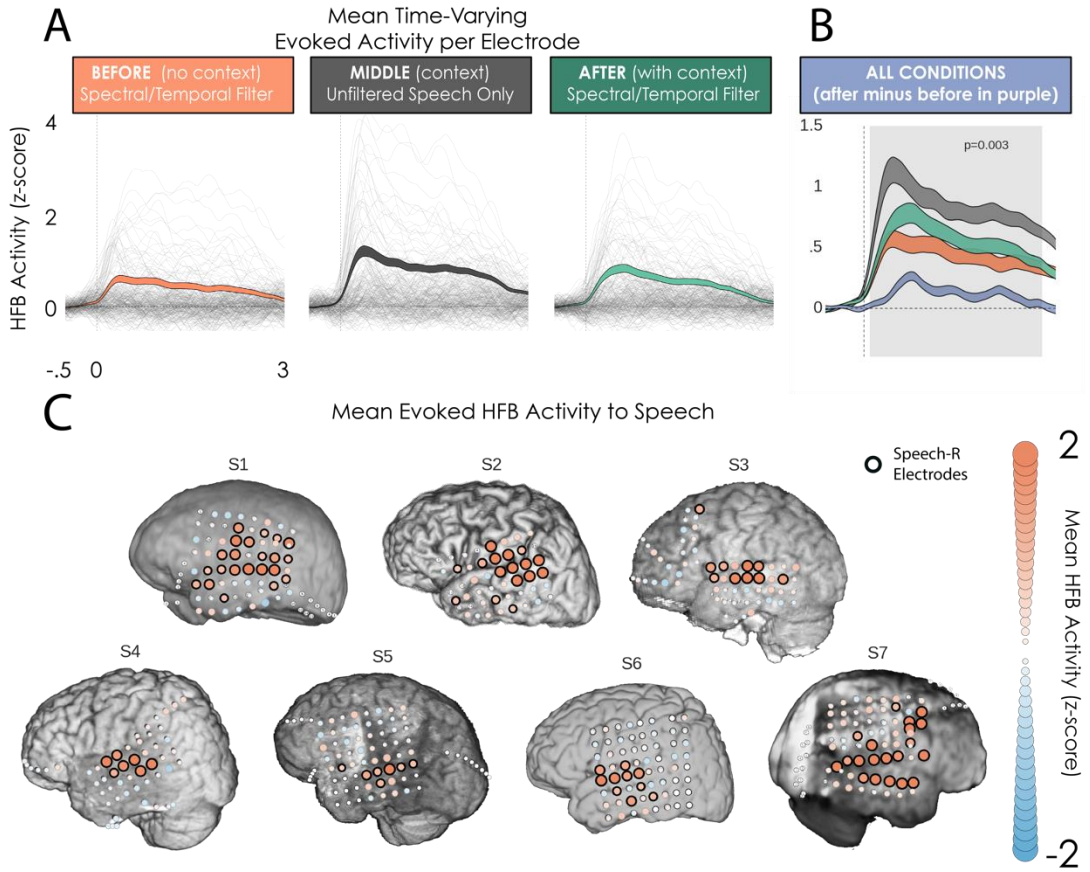


Figure 21 - Evoked High Frequency Broadband (HFB) activity - *(A) Mean HFB activity for each condition (BEFORE filtered - orange; MIDDLE unfiltered - grey; AFTER filtered - green). Individual traces are mean for each temporal lobe electrode. Shaded color traces are grand mean +/- standard error across all electrodes. Units are z-scores over baseline. (B) Grand mean +/- standard error for all*

active temporal lobe electrodes in each condition. The difference between AFTER and BEFORE conditions is shown in purple. Shaded regions represent significant differences between BEFORE and AFTER conditions (cluster-based permutation test, p=.003, n=92, see methods). (C) Electrode coverage and average HFB activity for each subject. Electrode colors/sizes represent the mean evoked HFB activity. Dark outlines show electrodes with HFB activity significantly different from zero (two-tail permutation test , p<0.01), called Speech-R electrodes. These electrodes had reliable increases in activity in response to speech stimuli. Speech-R electrodes located on the temporal lobe and perisylvian regions were included in eSTRF analyses.

Between-condition HFB Coherence

We next investigated whether the difference between the BEFORE and AFTER condition was only reflected in an overall increase in HFB amplitude or if there was also a difference in the time-varying details of each response. We hypothesized that HFB activity in the AFTER condition would be more similar to the activity in the MIDDLE condition (AFTER/MIDDLE) compared with the BEFORE condition (BEFORE/MIDDLE) on a trial by trial basis (i.e. for individual sentences). This would provide evidence that speech-responsive electrodes responded to features in the filtered speech stimulus that were also present in the unfiltered speech context stimulus.

The time-varying coherence between the BEFORE/MIDDLE and AFTER/MIDDLE HFB activity in each trial was estimated in order to quantify the similarity in the responses. The between-condition coherence for successive windows of 400ms was calculated to evaluate the time course of evoked HFB similarity for each trial, then averaged across trials to calculate the coherence for each time bin between the BEFORE/MIDDLE and AFTER/MIDDLE conditions for each electrode over peri-sylvian cortex. The coherence between AFTER/MIDDLE was higher than the coherence between BEFORE/MIDDLE, indicating it was not only the mean amplitude, but also the time-varying activity that was changing from BEFORE to AFTER (permutation test, p=.006, n=78; see Figure 3, bottom row for all comparisons). These differences in coherences could still be due to an overall change: the time-varying response averaged across all trials/sentences could be more similar to the MIDDLE (clean speech) condition in the AFTER than the BEFORE condition. This could happen if speech intelligibility resulted in simple changes in gain or if the response as measured in the HFB to intelligible clean speech was invariant across sentences. This increase in similarity would then be reflected in individual trial responses and result in increases in our coherence estimates. However, in electrodes for which the HFB response is sensitive to spectro-temporal features of sounds, one could expect to find an additional time-varying response that is sentence specific. To distinguish global changes from changes in sentence specific responses, the same between-condition coherence analysis was performed after subtracting the time-varying averaged HFB response across all trials/sentences for each electrode. After subtracting this global response in each electrode, a significant increase in coherence between the responses in the AFTER/MIDDLE conditions remained (see **Figure 22**). This finding shows that the time-varying and sentence specific response in each trial in the AFTER condition becomes more similar to the corresponding response to the unfiltered speech found in the MIDDLE condition and the effects are not simply due to global enhancement in neural activity.

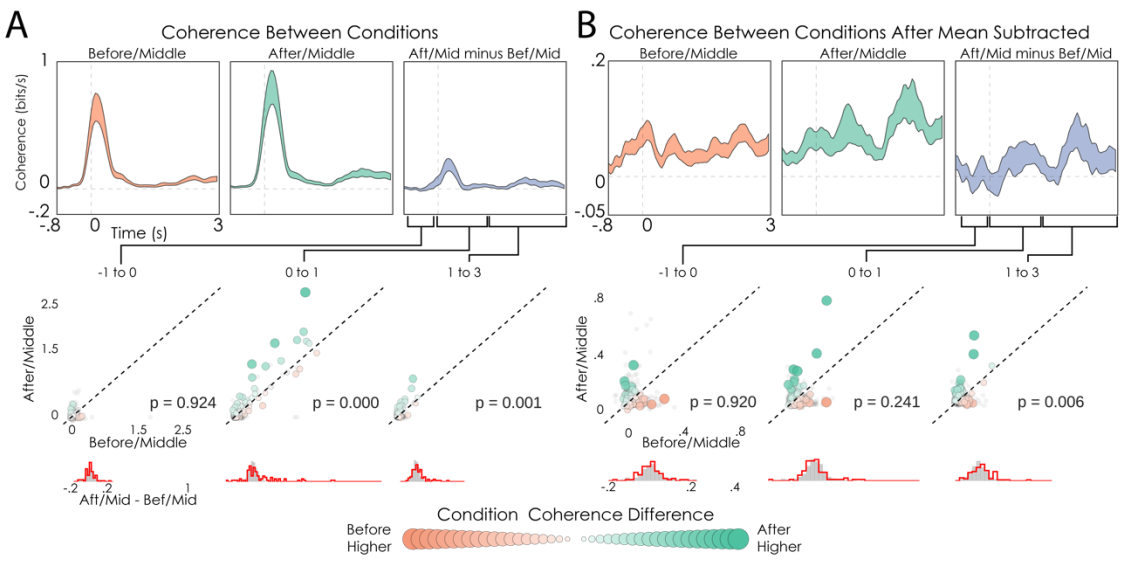


Figure 22 - HFB Similarity Between Conditions *(A) Upper plot, time-varying integrated coherence (in bits/s, see methods) in the evoked HFB response was calculated between pairs of conditions. Coherence was calculated for 400ms windows in 200ms steps. BEFORE/MIDDLE condition is shown in the left subpanels (mean +/- standard error across active electrodes), AFTER/MIDDLE condition in the middle subpanels, and the difference (AFTER/MIDDLE - BEFORE/MIDDLE) is shown on the right subpanels. Lower plot, average coherence values are shown for three time periods of interest: -1 to 0 seconds (left), 0 seconds to 1 second (middle), and 1 second to 2.5 seconds (right). Color and size of points represent the difference AFTER - BEFORE. Inactive electrode values are shown in gray, p-value reflects the difference of condition (AFTER/MIDDLE – BEFORE/MIDDLE). Histograms below show the distribution of AFTER - BEFORE for active electrodes (red) and inactive electrodes (grey). Trial to trial coherence is higher between the AFTER/MIDDLE condition for both post-stimulus windows. (B). Same as in (A), but after subtracting the average evoked response for each electrode. This compares coherence after accounting for global effects in evoked activity that are observed for all stimuli (see main text). Post-stimulus coherence is larger between the AFTER/MIDDLE conditions (permutation test for all comparisons, see bottom histograms for p-values).*

eSTRF Modeling

If the changes in the pattern of high-frequency activity in the AFTER condition relative to the BEFORE condition are related to increased speech comprehension, the AFTER activity should be more similar to the one found in response to clean speech, just as was observed. To further investigate this hypothesis, we next calculated the eSTRFs of all active temporal cortex electrodes in order to detect if they exhibited tuning plasticity related to an increase in speech comprehension. We estimated eSTRFs from stimulus-response (HFB) signals in each condition (BEFORE, MIDDLE, and AFTER). We hypothesized that, relative to the BEFORE condition, eSTRFs in the AFTER condition would shift to be more responsive to unfiltered speech features, providing a potential mechanism for extracting speech-like features from sound and the perceptual enhancement.

eSTRF models were fit for each electrode using a jackknife approach. On each iteration one trial was left out and the model was fit on the remaining trials. The held-out trial was then used to

estimate a goodness of fit (here the coefficient of determination, R^2) and its 99% confidence intervals. Electrodes with a confidence interval that did not overlap with 0 were considered to be electrodes "well-modeled" by the STRF and called STRF-Responsive (STRF-R) electrodes. This yielded 53 of 468 total electrodes (11.3%, see Supplementary Figure 2). The STRF-R electrodes were also generally localized on perisylvian temporal lobe regions (see **Figure 23** for anatomy and **Figure 24** for model score distribution).

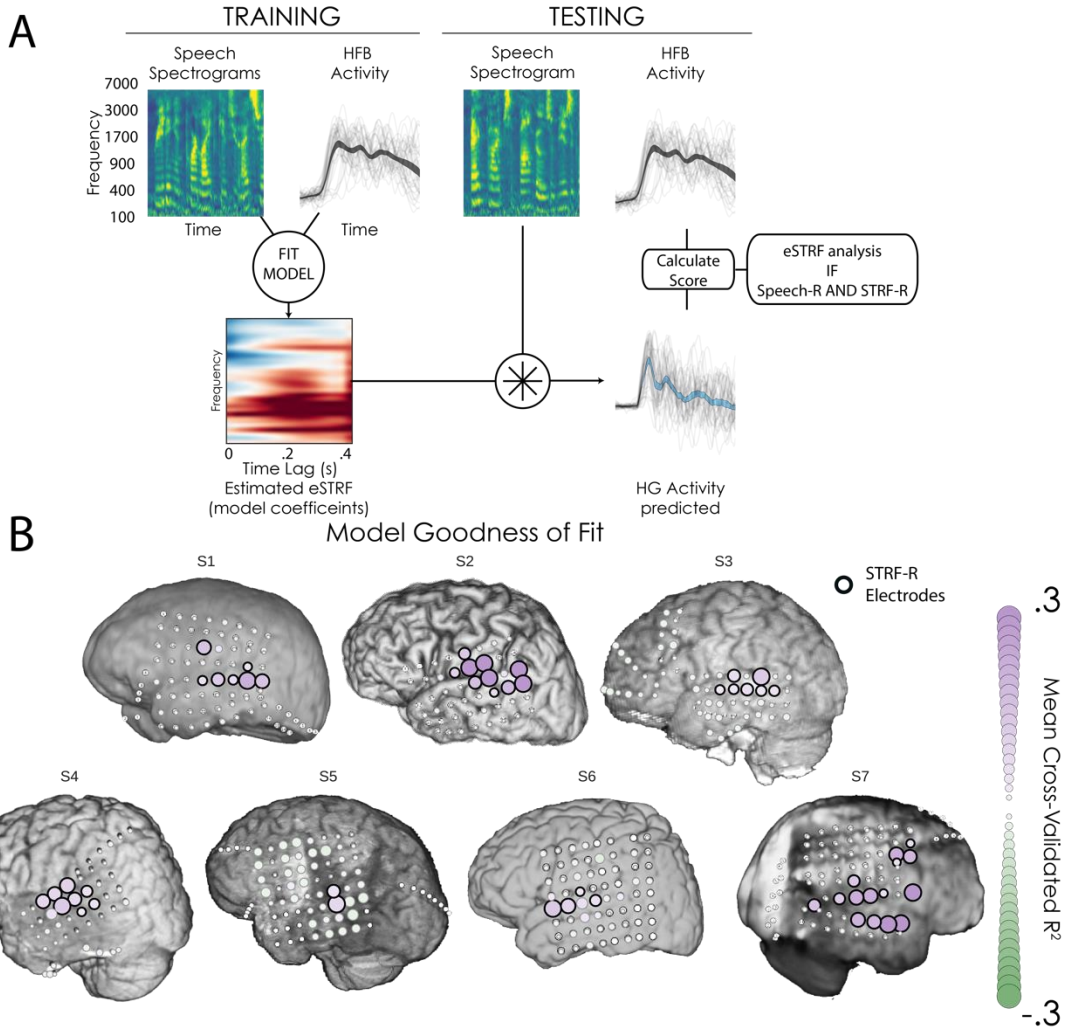


Figure 23 - eSTRF model fitting and goodness of fit across electrodes *(A) Example of model fitting procedure. Auditory spectrograms of sound and evoked HFB activity (top, first/second columns) is used to fit a linear regression model, resulting in a set of model coefficients (eSTRF, lower left). This eSTRF is convolved with a held out auditory spectrogram (top, third column) to generate a predicted HFB activity trace (lower right). The goodness of fit (cross-validated R^2) is calculated between the predicted response and the actual HFB activity in the held out trial (top, fourth column). This process is repeated, leaving out a different trial, until all trials have been included in the test set. (B) Average goodness of fit of the eSTRF model across subjects and electrodes. Size and color of each electrode represents the average model score. Electrodes with black outlines had model scores significantly above 0 (confidence interval test across trials) and are designated STRF-responsive (STRF-R) and were included in further eSTRF*

analyses if they also showed increased HFB activity (Speech-R, see Figure 3). See Supplementary Figure 2 for a comparison of Speech-R and STRF-R electrodes). Negative values of cross-validated R^2 can occur if parts of the neural signal that are not correlated with the stimulus spectrogram are overfit. Note that negative values of R^2 are small and not significantly different from zero as expected, see ***Figure 24*** *for distribution of all R^2 values.*

The coefficients of each eSTRF model (i.e. the specific spectrotemporal gains) were analyzed in order to investigate the nature of the specific spectro-temporal tuning of each electrode. To be included in subsequent analyses, an electrode had to: 1) show evoked HFB activity in response to speech (Speech-R, described above and shown in **Figure 21**); 2) be well-modeled by spectrotemporal features (STRF-R, described above and shown in **Figure 23**), and 3) be located on the temporal lobe or perisylvian cortex, regions traditionally associated with spectrotemporal auditory processing (Mesgarani & Chang, 2012; Pasley et al., 2012; Zatorre, Belin, & Penhune, 2002). This yielded 41 of 468 total electrodes (8.76%, see Supplementary Figure 2).

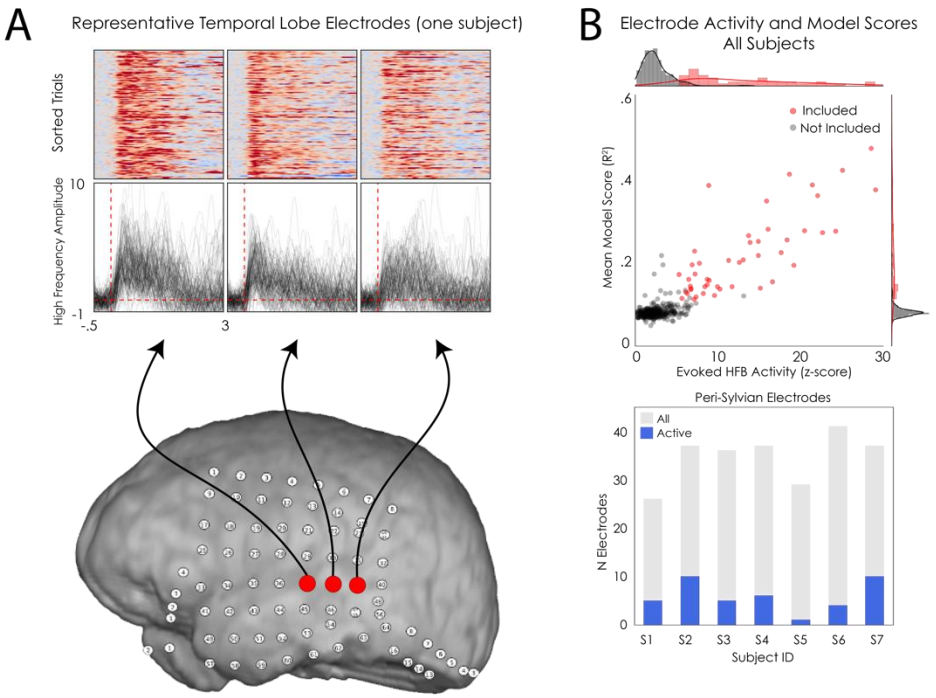


Figure 24 - Single-trial high-frequency broadband activity and relation to model scores *(A) Sample HFB activity for three active STG electrodes. Plots show stacked epoch plots of HFB activity, sorted by activity onset time. An increase in HFB activity is seen at the single trial level. Below, HFB activity from each trial (z-score over baseline) is shown. Temporal electrodes with high HFB activity and high model scores are plotted in red, and were included in eSTRF analyses (see methods). Electrodes that did not meet these criteria are plotted in grey. (B) Top: Mean HFB activity (z-score over baseline, x-axis) are plotted against Model scores (Cross-validated Coefficient of Determination, R^2, y-axis) for each electrode. Bottom: Bar graph showing the number of electrodes that had significant eSTRF predictions per subject. Color represents anatomical location, with most electrodes lying on the temporal lobe.*

Peaks in the eSTRFs were distributed across a wide range of frequencies, and eSTRFs were not well-characterized by simple shapes (e.g. Gabor functions) as seen in typical single unit STRFs (L. M. Miller, Escabí, Read, & Schreiner, 2001; Woolley, Gill, Fremouw, & Theunissen, 2009). This is likely due to the fact that the HFB activity represents the combined firing of ensembles of many thousands of neurons in cortical columns (Crone et al., 2011) (see **Figure 25** for examples). Sparser eSTRFs have also been obtained from ECoG and single unit data using different regularization techniques (David, Mesgarani, & Shamma, 2007; Hullett et al., 2016).

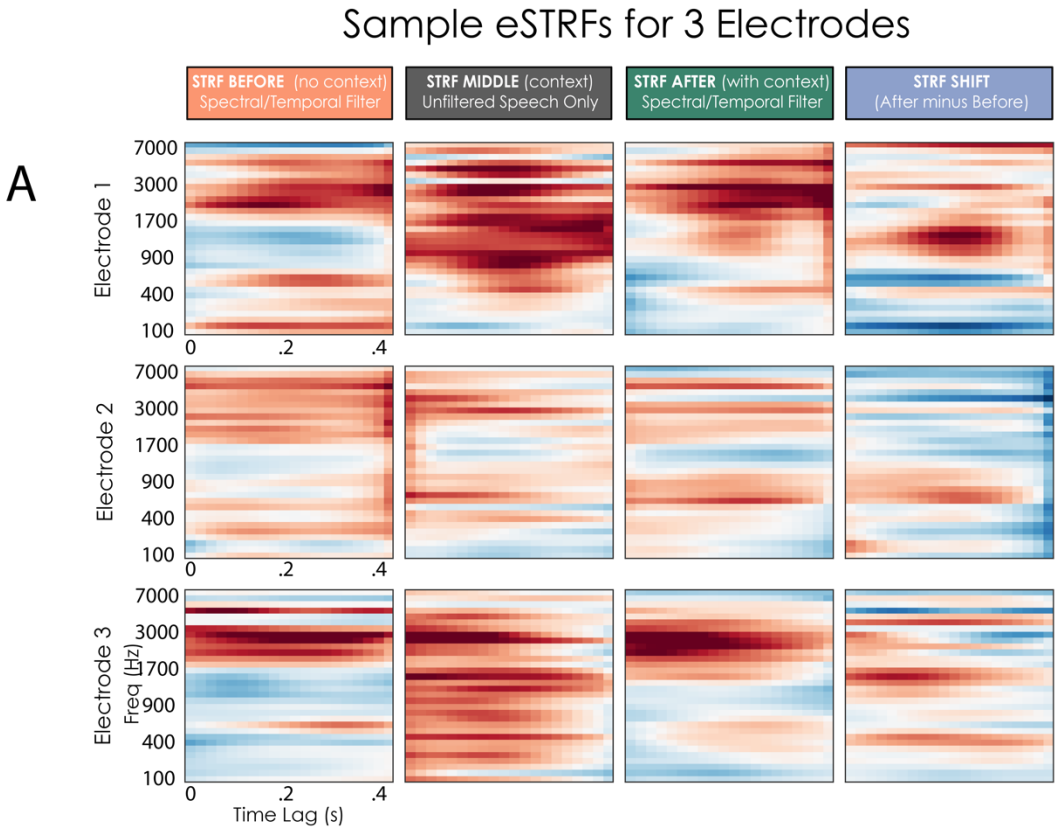


Figure 25 - Sample ensemble spectrotemporal receptive fields (eSTRFs) *(A) eSTRFs for multiple conditions (columns) and electrodes are shown (rows). The right column shows the change in eSTRF (AFTER - BEFORE). The gain of all eSTRFs shown has a color scale in z-score units, where the standard deviation is obtained across cross-validation folds. Subsequent analyses compare the similarity between the BEFORE/MIDDLE eSTRFs (orange) and the AFTER/MIDDLE eSTRFs (green).*

Shifts in eSTRF Modulation Related to Speech Intelligibility

To compare the spectral-temporal features present in speech with those extracted by the eSTRF, we next estimated the gain of the eSTRF in the spectral and temporal modulation domain: the eSTRF Modulation Transfer Function (MTF). The MTF shows which temporal amplitude modulations, which spectral envelope modulations, and which joint spectro-temporal modulations are emphasized (and equivalently attenuated) in the neural response.

The MTF can be compared to the MPS of speech to evaluate the match in tuning between the stimulus (here speech) and the neural filters (see Methods as well as (Singh & Theunissen, 2003; Woolley et al., 2009)). The average MTF functions obtained over all of our electrodes for the BEFORE and AFTER condition are shown in **Figure 26**. Qualitatively, one can observe that these MTFs are matched to the speech MPS shown in the figure. Moreover, the MTF of the shift in eSTRF (AFTER – BEFORE*)* averaged across all electrodes emphasizes the region of the MTF that was both preserved in the filtered speech and shown to be essential for speech intelligibility (Elliott & Theunissen, 2009), suggesting that the observed eSTRF plasticity could facilitate speech perception (see **Figure 26**).

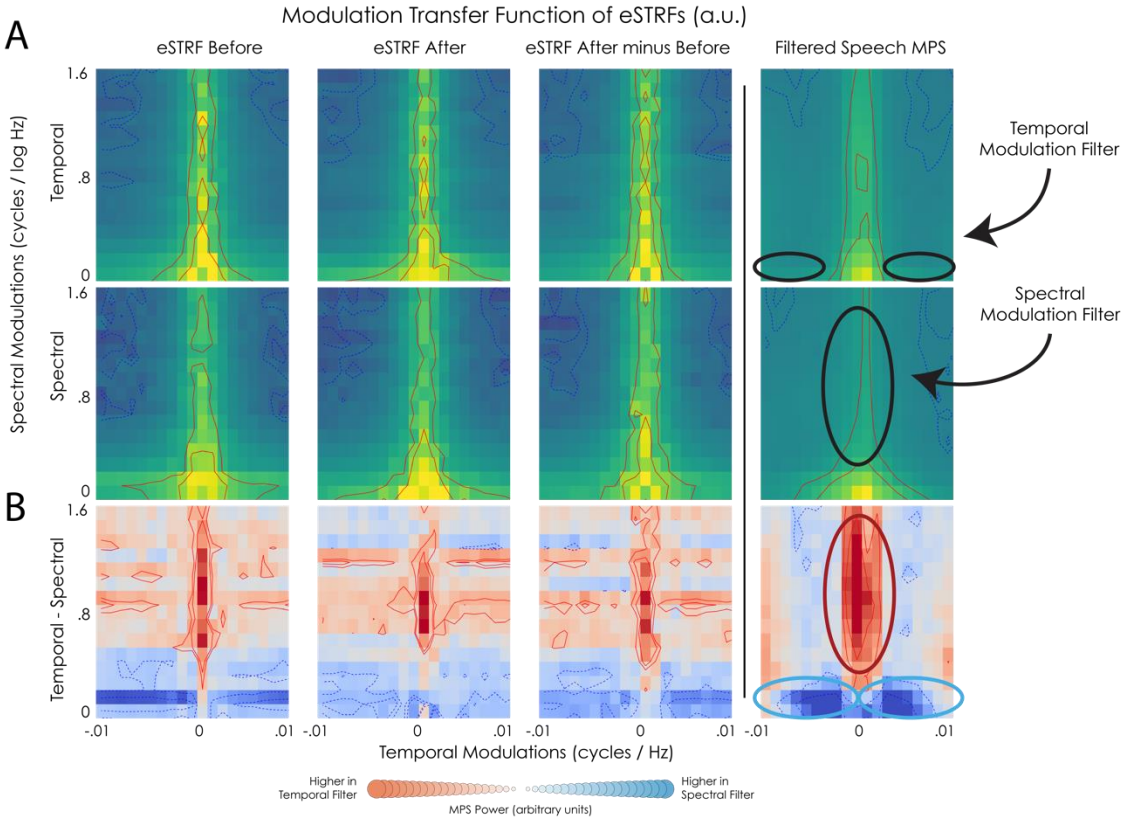


Figure 26 - Modulation Transfer Function (MTF) of eSTRFs *The modulation transfer function (or modulation gain) of each eSTRF was calculated, then averaged across conditions. eSTRFs were estimated separately for temporal-filtered stimuli and for spectral-filtered stimuli. (A) The mean MTF for eSTRFs is shown for the BEFORE (1st column) and AFTER (2nd column), condition. Lines represent 5th, 15th, 85th, and 95th percentiles. The difference in eSTRF was calculated, and the MTF of this difference is shown in the 3rd column. The MPS of the actual filtered speech stimuli is shown in the 4th column for comparison. Top row are eSTRFs fit on temporal-filtered stimuli, bottom row are eSTRFs fit on spectral-filtered stimuli. (B) The difference in the MTF for the two filter types (Temporal - Spectral) was calculated for each electrode. Z-scores for this difference are shown for each condition. The MTF of each eSTRF matches its respective filter type (cluster-based permutation t-test across electrodes, n=41, BEFORE p=.001, AFTER p=.003, AFTER-BEFORE p=.001), suggesting that tuning changes emphasize spectro-temporal modulation that are present in the degraded speech sound and are crucial for speech*

intelligibility. Note that these features were also present in the BEFORE stimulus (since BEFORE and AFTER stimuli were identical) and in the BEFORE eSTRF (as shown in the MTF for the eSTRF BEFORE, left column). This suggests that neural tuning in the BEFORE condition does match filtered speech features, and that this is accentuated after hearing the unfiltered speech.

Filtered Speech eSTRFs Increase Response to Speech Features

We conducted two additional analyses to quantitatively determine whether the AFTER eSTRFs became more sensitive to unfiltered speech features. First, we assessed whether the eSTRF in the AFTER condition was more sensitive to unfiltered speech features than the eSTRF in the BEFORE condition. Each filtered speech eSTRF (BEFORE and AFTER) was used to calculate a predicted response to unfiltered speech. The magnitude of this predicted response reflects the extent to which the eSTRF extracts spectrotemporal features that are present in the input stimulus, in this case unfiltered speech. The root-mean-squared power of the output in the BEFORE and AFTER condition was calculated and compared for each electrode: the power in the AFTER condition was higher than power in the BEFORE condition (mean RMS increase .12+/- .03, p=.0001, n=41; see **Figure 27**).

Next, we used the eSTRFs fit on the unfiltered speech MIDDLE condition to predict HFB activity in the BEFORE and AFTER conditions. The predicted HFB activity was compared to the true HFB activity to assess how well the unfiltered eSTRF characterized the mapping from acoustic features to neural activity. Larger goodness of fit (R^2) values indicate that the mapping of sound features onto HFB activity is more similar to that of the unfiltered speech condition. Goodness of fit scores were higher for the AFTER condition compared with the BEFORE condition (mean R^2 improvement .05 +/- .01, p=.0001, n=41; see **Figure 27**).

Taken together these results together show that the tuning of electrodes in the AFTER condition becomes more similar to tuning acquired in response to unfiltered speech. Moreover, this shift in tuning causes the eSTRF to be more responsive to speech-like features of the stimulus.

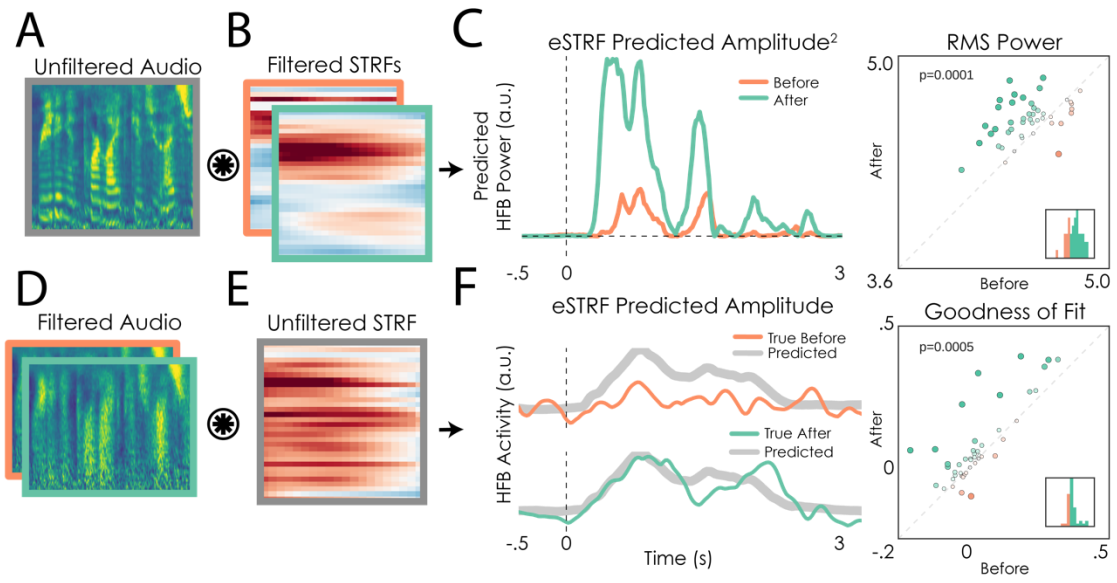


Figure 27 - eSTRF changes overlap with speech features *Top: Spectrograms of unfiltered speech (A) was convolved with the eSTRF fit on each filtered speech condition (B). The size of the predicted response (C, left) depends on the overlap between the eSTRF and the unfiltered speech features. Scatterplot (C, right) shows the predicted response power between BEFORE and AFTER conditions. Size and color represent the difference (AFTER - BEFORE), and inset histogram shows the distribution of differences. Bottom: Spectrograms of filtered speech (D) were convolved with the eSTRF fit on unfiltered speech (E), resulting in a predicted HFB amplitude for that spectrogram (F, left side, grey trace). This was compared with the true HFB activity in each condition (green and orange traces). Correlation between the predicted and actual trace reflects the extent to which an unfiltered eSTRF is predictive of the neural response. The scatterplot shows the comparison between Predicted and True AFTER vs. Predicted and True BEFORE (F, right). Size and color of points represent the difference (AFTER - BEFORE), and inset histogram shows the distribution of this difference.*

Filtered Speech eSTRF Shifts Overlap with MIDDLE eSTRFs

Finally, to directly compare the spectrotemporal tuning between conditions, we calculated the similarity between eSTRFs obtained in each condition using partial correlation (see Methods). Partial correlation measures the correlation between two variables after removing the linear relationship with a third variable. The partial correlation between the eSTRFs in the BEFORE/MIDDLE conditions was estimated after taking into account the eSTRF from the AFTER condition, and vice-versa. Partial correlations between AFTER/MIDDLE were higher than between BEFORE/MIDDLE (**Figure 28**A; mean partial correlation improvement .18 +/- .001; p=.001, n=41, permutation test) indicating that eSTRFs obtained with degraded speech shifted to become more like the eSTRFs obtained with intelligible speech. The majority of the increase in partial correlation is located around the STG (see **Figure 28**C), a region shown in previous research to respond to spectrotemporal features in many sounds (Martin et al., 2014; Mesgarani & Chang, 2012; Pasley et al., 2012; Zatorre et al., 2002). For subjects that also had

pink noise control trials, there was no difference in partial correlation between the BEFORE and AFTER condition (see Supplementary Figure 3B).

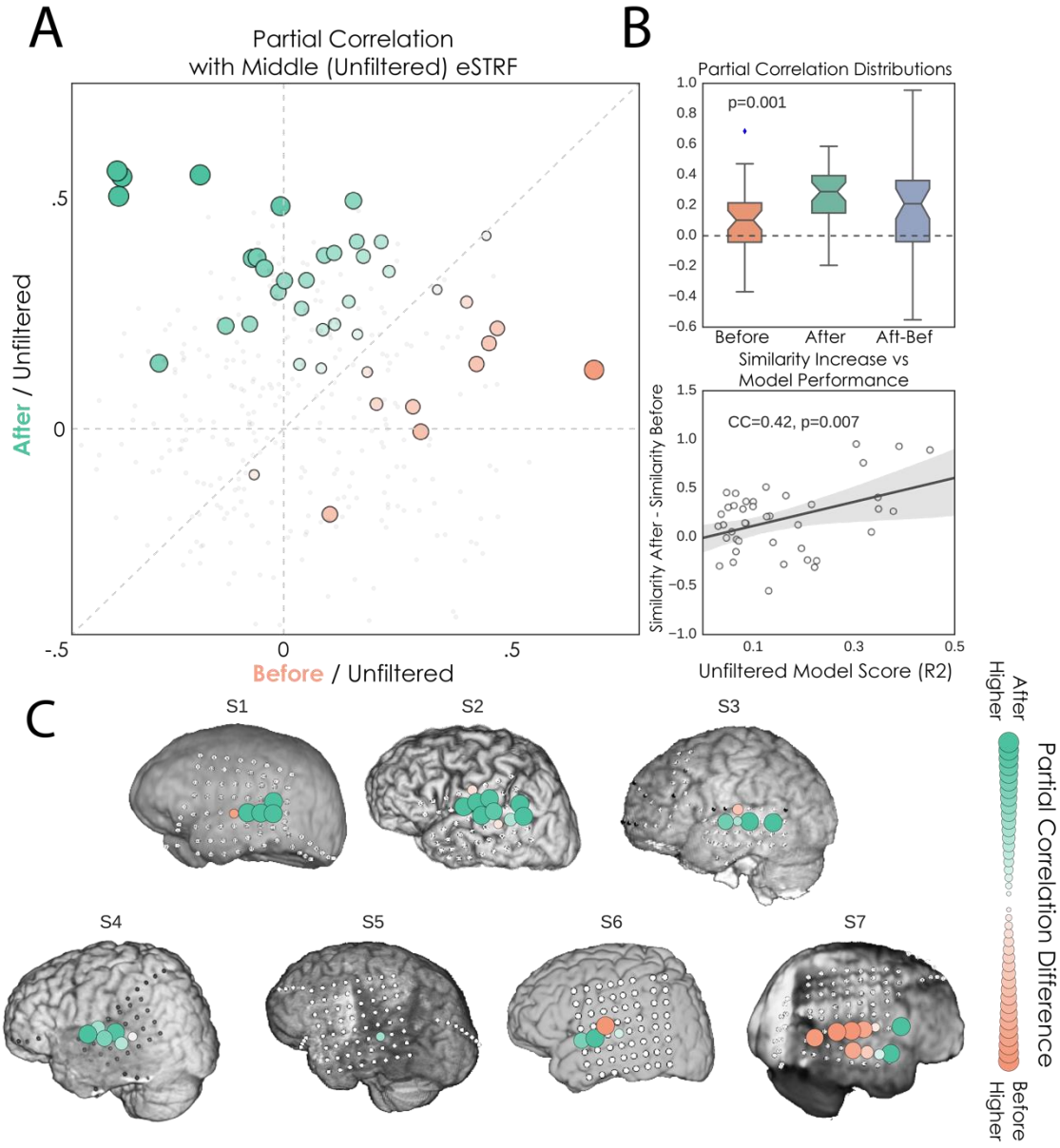


Figure 28 - eSTRF Similarity Analysis *(A) Linear similarity (partial correlation) is shown between eSTRFs for the BEFORE/MIDDLE conditions after regressing out the AFTER Condition (x-axis) and the AFTER/MIDDLE condition after regressing out the BEFORE condition (y-axis). Electrode color and size represent the difference in partial correlations (AFTER/MIDDLE in green, BEFORE/MIDDLE in orange). (B) Top: Median (+/- 75th and 25th percentiles) partial correlations between MIDDLE eSTRF and the BEFORE (left bar) and AFTER (middle bar) condition, as well as difference in partial correlation between AFTER/MIDDLE and BEFORE/MIDDLE conditions (right bar). Bottom: The change in partial correlation (AFTER/MIDDLE - BEFORE/MIDDLE) increases as a function of the electrode's goodness of fit with the eSTRF model (Pearson's r, see figure for stats). (C) Partial correlation differences for each subject and each electrode. Model results are restricted to electrodes*

located on the temporal lobe. Greener colors reflect a higher partial correlation between eSTRFs in the AFTER/MIDDLE conditions. Size of the electrode represents the magnitude of the difference.

Connectivity Analysis

We also examined whether the observed eSTRF plasticity was correlated with changes in functional connectivity measured across electrodes, providing initial clues for whether the tuning shifts could be driven by top-down effects. For this purpose, we calculated the coherence of the HFB amplitude between the electrodes included in the eSTRF analysis and groups of electrodes either in temporal or frontal/premotor cortex. There was a small significant increase in coherence in the AFTER condition relative to the BEFORE condition. Moreover, the coherence in the AFTER condition was closer to that obtained with the unfiltered speech (see Supplementary Figure 5A and Supplementary Methods). These results suggest that there may be changes in functional connectivity that are consistent with a state in the AFTER condition that is closer to the one found during the perception of intelligible speech, potentially explaining the observed changes in the tuning eSTRF for speech like features. This amplitude coherence analysis does not, however, reveal the direction of the effect and it is also possible that the changes in auditory tuning cause the observed changes in functional connectivity.

To examine potential directional effects, we also calculated the Phase Amplitude Coupling (PAC) between the phase of the ECoG signal in the 3-8Hz from electrodes in frontal / premotor regions and the HFB amplitude for electrodes included in the eSTRF analysis in auditory cortex (see Supplementary Fig 5B). We used frequencies from 3 to 8 Hz for the phase calculation as it has been suggested that phase in this frequency range may track the envelope of a perceived speech stimulus and that the low frequency signal could drive responses in higher frequencies (Ding & Simon, 2014; Giraud & Poeppel, 2012; Peelle et al., 2013). In this analysis, however, we did not find significant inter-region cross-frequency activity that was modulated by task condition.

Thus, although changes in functional connectivity were measured both within the temporal lobe and between the frontal cortex and the temporal lobe, we are unable at this point to distinguish top-down from local or bottom up effects. Additional experiments with greater coverage of frontal neural activity and additional analyses are required to determine the direction of the information flow that drives the observed plasticity in the temporal lobe and perisylvian region.

Discussion

After hearing an intact sentence, subjects understand a subsequent noisy version of the same sentence that was previously unintelligible. This robust perceptual enhancement is characterized by an increase in HFB amplitude, onsetting within 300 milliseconds and sustained throughout the speech utterance. Moreover, the time-varying HFB activity becomes more similar to activity during passive listening to unfiltered, intact speech, providing evidence that auditory electrodes shift how they track the time-varying properties of the filtered speech. Finally, a spectrotemporal analysis of human auditory cortical speech responses (eSTRFs) shows

that the perceptual enhancement due to exposure with intact speech is paralleled by a shift in spectrotemporal tuning in auditory cortical areas. This shift in tuning overlaps with speech features, making the cortical population more responsive to unfiltered speech.

These results provide novel evidence that experience with language rapidly and automatically alters auditory representations of spectrotemporal features in the human temporal lobe. Rather than a simple increase or decrease in activity, it is the nature of that activity that changes via a shift in receptive fields. This has implications for encoding models of sound features in human/animal models, as well as in theories of top-down auditory processing. There have been attempts to characterize the neural response to speech under attention-based manipulations. For example, Mesgarani and Chang used a decoding approach to estimate the spectral representation of sound in the auditory cortex during a task in which subjects attended to one of two utterances being played simultaneously (Mesgarani & Chang, 2012). The authors reported that the decoded spectrogram became more similar to the speech stream that was being attended to, suggesting plasticity in the information encoded in cortical electrodes. This effect may be due to enhancing the gain of specific filter channels in the auditory cortex, as we have observed here. The tuning shift can be interpreted as a "spectrotemporal prior" over incoming sounds, priming auditory cortical neurons to respond to particular speech-like qualities. This interpretation is compatible with higher-level theories of categorical (or probabilistic) speech representation, such as such as perceptual warping (Feldman, Griffiths, & Morgan, 2009)

Relatively rapid changes in auditory STRFs have also been demonstrated in animal models. Research that showed task-dependent plasticity in auditory STRFs was initially performed in ferrets who were trained to detect target pitches in a go/no-go task (Fritz et al., 2003). More recently, it was shown that these auditory STRFs were dynamic and shifted due to concurrent top-down and bottom-up demands that depended on particular behavioral tasks (David et al., 2012). This idea is supported in the current study, which revealed rapid cortical plasticity due to the knowledge of high-level auditory features. There have also been studies in animal models that report an invariance to signals embedded in different levels of background noise. For example, Rabinowitz et al showed that neurons higher in the cortical hierarchy were more invariant to noise levels (Rabinowitz et al., 2013). They proposed two separate adaptive gain mechanisms by which neurons separate signal from noise in order to be more sensitive to relevant stimulus features. Similarly, in our study, the perceptual enhancement coming from experience with unfiltered speech could be thought of as a kind of "signal enhancement" in which high-level information causes neurons to vary their gain in order to experience a signal with less noise.

How single-unit STRFs combine to form an ECoG electrode eSTRF is an important next step to bridge the gap between the animal and human literature and advance our understanding of the neural mechanisms that can drive this cortical STRF plasticity. Given the dependence of the behavioral effect on linguistic attributes, we predict that this rapid, automatic shift in the eSTRF originates at least in part from top-down signals in higher-level regions that are part of the language network such as auditory association areas (Bornkessel-schlesewsky & Schlesewsky, 2013; DeWitt & Rauschecker, 2012; Leaver & Rauschecker, 2010; Wassenhove & Schroeder,

2012), or in “non-auditory” regions such as the inferior frontal gyrus or premotor cortices (Horwitz & Braun, 2004).

A prior ECoG study suggested that delta-theta power entraining may provide a mechanism for using temporal structure of the sound to “chunk” relevant auditory streams and facilitate speech processing (Arnal & Giraud, 2012). This theta power entraining might originate in prefrontal cortex and affect lower auditory areas. Although we found functional connectivity effects that were modulated by the task condition, we did not detect inter-regional PAC changes that could provide more substantial evidence for direction of the information flow. Future research with joint frontal/temporal coverage will be needed to explicate the origin of putative top-down processes (e.g. from frontal regions) that might contribute to eSTRF plasticity.

In summary, in this study we demonstrate rapid spectro-temporal plasticity while subjects listened to both normal and degraded speech. We show that the human auditory cortical map is highly dynamic and context dependent, and highlight the importance of studying sensory cortical responses with behaviorally relevant, naturalistic stimuli (Theunissen & Elie, 2014). The dynamical changes observed in these sensory maps are dependent on a spectrotemporal prior related to high-level speech features that increases speech signal identification, enabling perception of a stimulus that was previously incomprehensible.

Methods

Participants and data acquisition

Electrocorticographic (ECoG) recordings were obtained using subdural electrode arrays implanted in 7 patients undergoing neurosurgical procedures for epilepsy (age 22-51; 4F/3M). Recordings took place at the University of California at Irvine (UCI), Columbia University (CU), and John Hopkins University (JH). All patients volunteered and gave their informed consent before testing, and this research was approved by the Committees for the Protection of Human Subjects at UC Berkeley, UC Irvine and the Johns Hopkins Medical School. Grid placement was determined entirely by clinical criteria (see **Figure 21** for reconstructions of subjects). Electrode grids had spacing from 5-10mm (Adtech grids), with the following numbers of channels: JH: (48, 64), IR: (68, 68, 62), CM: (110, 104).

Multi-channel ECoG data were amplified, analog-filtered above .01Hz, and digitally recorded with a sampling rate of 1KHz (JH, CU) or 5KHz (UCI). All channels were subsequently down-sampled to 1KHz, corrected for DC shifts, and band pass filtered from 0.5 to 200 Hz. Notch filters at 60 Hz, 120 Hz and 180 Hz were used to remove electromagnetic line noise. All filters were zero-phase IIR filters implemented with the MNE-python toolbox (Gramfort et al., 2013). The time series were then visually inspected to remove time intervals containing periodic spiking discharges and generalized spiking due to ictal activity. All epileptic channels, as well as channels that had excessive noise including broadband electromagnetic noise from hospital equipment and poor contact with the cortical surface, were removed from analysis. Finally, electrodes were re-referenced to a common average.

Brain Mapping of Electrodes

Each subject had postoperative anterior–posterior and lateral radiographs, as well as computer tomography (CT) scans to verify grid locations. Three-dimensional cortical models of individual subjects were generated using pre-operative structural magnetic resonance (MR) imaging. These MR images were co-registered with the post-operative CT images using Curry software (Compumedics, Charlotte, NC) to identify electrode locations. Cortical activation maps were generated using custom Python software.

ECoG filtered speech passive listening task

ECoG subjects performed a passive listening task that consisted of 50-60 trials. In a single trial, the subject fixated on a cross in the middle of a laptop screen. Three sounds were played successively through laptop speakers. These followed the pattern "filtered speech (BEFORE*)*-> unfiltered speech (MIDDLE*)*-> filtered speech (AFTER*)*", and the speaker/content of the sentence was always the same within a single trial. However, no sentence was repeated within the same subject. Each stimulus was 2-5 seconds long. The inter-stimulus interval was randomly chosen between .5 and 1.5 seconds on each presentation, resulting in a trial length of 12-16 seconds.

In three ECoG subjects, a pink-noise control trial was added to test for the effect of filtered speech repetition on electrode tuning. In these trials, the unfiltered speech context (middle sound presentation) sentence was replaced with energy-matched pink noise. This trial type made up 50% of trials in these subjects. Trials were conducted using the PsychoPy open-source toolbox (Peirce, 2008).

Behavioral Controls for Filtered Speech

Time constraints in the epilepsy ICU environment precluded detailed behavioral assessment of ECoG patients, though post-test, patients typically reported a perceptual enhancement after hearing the unfiltered speech stimuli. An additional behavioral experiment was conducted to assess the degree of perceptual enhancement after hearing the unfiltered speech using different kinds of MIDDLE context sentences. Subjects were divided into three groups. Each subject was asked to listen to speech sentences (explained below) and to type out any words they understood. The mean percentage of words for each sentence was calculated for each subject, and then compared across groups with an unpaired t-test. The first group (n=5) replicated previous intelligibility experiments using the same stimulus set (Elliott & Theunissen, 2009). Filtered sentences were presented to subjects. AFTER each presentation, subjects were asked to type any words that they could understand, and the percent correct of sentence words detected was calculated.

The second group (n=9) was used to control for the effect of stimulus repetition, as well as general changes in arousal due to hearing unfiltered speech. Subjects were presented with the same trial structure used in the ECoG recordings. The BEFORE/AFTER stimuli were always the same filtered speech sentence, and the MIDDLE stimuli was either an unfiltered version of the same sentence or of a different sentence. Using a different sentence in the MIDDLE tests for an effect of filtered speech repetition, as no linguistic or acoustic context matches the filtered

speech sentence. Using the same sentence in the MIDDLE elicits the same perceptual enhancement effect reported in ECoG subjects. Subjects were again asked to type as many words as they understood and the mean percent correct is reported. The third group of subjects (n=15) was used to test linguistic vs acoustic stimulus features on the perceptual enhancement effect. Subjects performed the same task as group 2. However, now the filtered speech context sentence was either the same sentence spoken by a different-gendered speaker, or a different sentence spoken by the same speaker. This provides a coarse split between acoustic context (same speaker, different sentence) and linguistic context (same sentence, different speaker). Subjects were again asked to type any words they understood.

Filtered Speech Sound Creation

Filtered speech was created using a Modulation Transfer Function applied to the joint Spectral-Temporal Modulation Spectrum of the individual speech sentences as described in (Elliott & Theunissen, 2009). This filtering allows one to remove particular frequencies in the joint spectrotemporal envelope of the sound. Briefly, the raw sound waveform is first converted into a time-frequency representation (a spectrogram). Then, a two-dimensional Fourier transform of the sound spectrogram converts this representation into a domain that describes the spectral and temporal modulations that are present in the spectrogram.

The temporal modulations correspond to fluctuations of the amplitude envelope of the sound such as those produced by words and syllables, while the spectral modulations correspond to both coarse (such as speech formants) and fine (such as the harmonics from glottal pulse) repeated structures found along the frequency axis (see **Figure 19** for a visual explanation).

Once the sound has been transformed to this space, the gain of a large portion of frequency modulations (spectral filter) or temporal modulations (temporal filter) is set to 0 (the phase modulation spectrum is left untouched, see **Figure 19** for filtering procedure examples and **Figure 20** for spectrogram examples). This filtered MPS is converted to a spectrogram using an inverse 2-D Fourier transform, and finally back into a time-varying sound wave using a recursive algorithm that selects the appropriate phase shift for each frequency band and recovers the unique sound that corresponds to that spectrogram (Elliott & Theunissen, 2009). The result is a stimulus that sounds speech-like, but is incomprehensible to the naïve listener. Note that the overall frequency power spectrum of the modulation-filtered sounds is unchanged: it is the same as the unfiltered sound.

In this study, two filters were used: a low-pass filter of spectral modulations (.5 cycles / kHz), and a low-pass filter of temporal modulations (3 cycles / Hz). The parameters of these filters were chosen to remove respectively the spectral structure or the temporal structure that is key for speech comprehension (Elliott & Theunissen, 2009).

Neural and Auditory Feature Extraction

Our primary analysis consisted of fitting a linear model that predicted patterns of ECoG High Frequency Broadband (HFB) amplitude as a function of spectral features. Auditory features (inputs to each model) consisted of time and frequency varying amplitudes based on psychoacoustic and physiological studies of language processing (Chi et al., 2005). This

"auditory spectrogram" was obtained by estimating the amplitude envelope for 128 narrow-bands generated by a bank of erb-spaced gammatone filters ranging from 180-7,000 Hz. To obtain the envelope of each narrow-band signal, the output of the filter is half-wave rectified, followed by a non-linear compression, and spectral sharpening. Finally, the output of each frequency band was passed through a leaky integrator with a time constant of 8ms (details on the feature extraction can be found in Chi et al., (2005)). The 128 acoustic frequencies of the initial spectrograms were subsequently down sampled to 32 frequency bands to reduce dimensionality and computational load.

Neural activity (outputs of the model) consisted of the envelope of the HFB activity of each electrode. A window around 21 center frequencies were defined from 70Hz to 140Hz, with the width of each window increasing semi-logarithmically with frequency, following previous studies in ECoG encoding models (Bouchard & Chang, 2014). The raw ECoG signal was first bandpass filtered for each window using a zero-phase IIR filter. Then, the amplitude of the band-passed signal was calculated as the modulus of the Hilbert transform of the signal. Finally, the amplitude for each center frequency was averaged together to attain a single time-varying estimate of HFB activity (Bouchard & Chang, 2014). Before estimating the linear filter, the audio spectral representation and the neural HFB response were downsampled to 50 Hz.

Evoked HFB and Speech Responsive Electrodes

For electrode selection, we baselined each trial using times -800 to -100ms relative to sound onset. We then calculated the mean post-stimulus activity in each trial. This yielded a single value for evoked HFB activity per trial/condition. For each condition, 99% confidence intervals on mean evoked activity across all trials were obtained by bootstrapping. Electrodes whose lower bound (bootstrapped .5th percentile) were greater than 0 in response to unfiltered speech were considered Speech-R electrodes.

To test for differences in mean HFB activity between conditions, the difference in time-varying HFB activity in each trail was calculated and then averaged across trials to obtain a single "difference time-varying HFB activity pattern" per electrode. Significance (the null hypothesis being no difference AFTER – BEFORE) was estimated using a cluster-based permutation test that corrects for multiple comparisons and computes statistics at the cluster level (Maris & Oostenveld, 2007).

Between-condition Coherence

The similarity in HFB activity between each filtered speech condition (BEFORE/AFTER) and the unfiltered speech condition (MIDDLE) was assessed by the measure of coherence. Coherence was chosen instead of the cross-correlation coefficient because of its robustness to high-frequency noise and invariance to systematic phase delays between signals. Similar results were obtained with correlation coefficient analysis (results not shown) but the correlation coefficient calculation requires additional assumptions on the relevant temporal scale of analysis related to the low-pass filtering needed to extract the lower frequency signals from the higher frequency noise. In the coherence calculation, the estimation of this relevant time scale is implicitly performed in a data driven fashion, as the signal to noise is estimated for each frequency. The integral of the coherence (expressed here in bits/s and shown in **Figure 22**)

yields then a measure of overall similarity for two time varying signals (Hsu, Borst, & Theunissen, 2004). For each stimulus, the time varying coherence between the BEFORE/MIDDLE conditions, and between the AFTER/MIDDLE conditions was estimated using a multi-taper windowing function (Slepian, 1978) and all the trials. The coherence was calculated for a sliding window of 400ms moving in 200ms steps from -500ms to 2500ms, relative to stimulus onset. Unbiased estimates of the coherence for each window were obtained using a jackknife method. To compare overall coherence across electrodes, coherence was converted to normal mutual information (in bits/second), an information theoretic representation that allows for the integration of the coherence across frequency bands (Hsu et al., 2004), which takes the following form:

$$MI_{norm}(f) = -log2(1 - coh(f))$$

The mean +/- standard error time-varying integrated coherence (in bits/s) was calculated across electrodes for each pair of conditions (BEFORE/MIDDLE and AFTER/MIDDLE). Statistics are performed for windows of interest on the mean difference in coherence between AFTER/MIDDLE and BEFORE/MIDDLE (See **Figure 22**). Code for performing the trial-to-trail coherence can be found in the Data Availability section.

eSTRF Model Formulation

Three eSTRFs were fit from the data obtained from each electrode: one using audio from the BEFORE trials, one using MIDDLE trials, and one using AFTER trials. This allowed for the comparison of eSTRFs coefficients from one trial type to the next.

The eSTRF is an encoding model that describes the linear mapping between the speech spectrogram and the HFB activity. It models the HFB signal as a weighted sum of the amplitude at each frequency band and for a range of points in time as follows:

$$\hat{R}(t,n) = \sum_{\tau}\sum_{p} g(\tau,p,n)S(t-\tau,p)$$

where $\mathrm{S}(t-\tau,p)$ is the estimated speech representation for the frequency band p at time lag $(t-\tau)$, with τ being a time lag ranging between 0ms and 400ms. $\widehat{\mathrm{R}}(t,n)$ is the estimated HFB neuronal response of electrode *n* at time t. Finally, $\mathrm{g}(\tau,p,n)$ is the linear transformation matrix (or set of eSTRFs), which depends on the time lag, feature of interest, and the electrode being predicted.

To obtain the eSTRF, a regularized linear regression algorithm was used. Linear regression attempts to find parameter values that capture the relationship between the input and output (in this case, stimulus features, and brain activity). It accomplishes this by finding parameters that minimize the squared difference between model fit and training data, the minimum square error (MSE). The MSE solution is the solution that maximizes the likelihood for Gaussian noise distributions. However, when the number of model parameters is large in comparison to the fitting data size, the MSE solution can yield parameter values that are determined by the particular data set rather than the underlying relationship (called overfitting). To control for this, regression is paired with regularization, a technique that minimizes the tendency of a model to overfit data by effectively shrinking the magnitude of parameters. Shrinkage is

obtained by implementing prior distributions on parameters centered at zero. This prior results in an additional penalty term that is added to the MSE.

In the case of linear Ridge Regression, a single parameter (here referred to as the Ridge parameter) controls the penalty incurred by large parameter values. Specifically, Ridge Regression includes a penalty term for the L2-norm of parameter weights. This type of penalty corresponds to a Gaussian prior centered at zero for the model parameters, with the Ridge Parameter specifying the variance of this distribution. To choose a value of the Ridge parameter, experimental trials were repeatedly split into training and test sets using a jackknife approach. On each iteration, one trial was left out for model validation. Models were fit on the training data for multiple values of the Ridge parameter. All training inputs/outputs were standardized to zero mean and unit standard deviation (i.e. z-scored) before model fitting. For each model, the goodness of fit was calculated using the coefficient of determination (R^2) between the predicted HFB response and the actual response in the validation trial. This cross-validation was performed for all electrodes / conditions, and repeated until all trials had been used in the test set, resulting in a distribution of selected ridge parameters yielding the maximum R^2.

To ensure that the prior over model coefficients was the same in all conditions, the mode of the distribution of ridge parameters for active electrodes was selected, and all models were re-fit with this single value for the ridge parameter using the same cross-validation described above. Model coefficients were averaged across all splits for final coefficient estimates. The cross-validation procedure was also used to calculate t-values of model coefficients by taking the mean divided by the standard deviation across CV splits. Code for performing encoding model fitting and cross-validation across trials can be found in the Data Availability section.

It should be noted that eSTRFs reported in this study visually have a slightly greater temporal extent than those in a recently published article that used a different (but related) approach to electrode receptive field analysis using maximally informative dimensions (Hullett et al., 2016). Receptive fields derived from models are sensitive to the assumptions and constraints of that model, and one would expect differences in STRF shape when using different models. This paper used $L2$ regularization (Ridge regression) due to its interpretability, computational efficiency, and robustness and prevalence in the literature. Other alternatives such as maximally informative dimensions, boosting, or L1 (Lasso) regularization may yield sparser STRFs (David et al., 2007; Hastie, Tibshirani, & Friedman, 2009; Hullett et al., 2016).

All model fitting was performed with custom code that relied on the Python libraries scikit-learn (Pedregosa, Grisel, Weiss, Passos, & Brucher, 2011) and MNE-python (Gramfort et al., 2013), which are built on top of the scipy/numpy stack (Van Der Walt, Colbert, & Varoquaux, 2011).

Modulation Transfer Function of eSTRFs

To investigate whether the eSTRF gain was tuned for spectrotemporal features found in speech stimuli, the modulation power spectrum (MPS) of the sounds was compared to the modulation transfer function (MTF) estimated for each eSTRF. Similar to a frequency power density spectrum, the MPS is obtained from the amplitude of the 2d Fourier transform of the spectrogram (Singh & Theunissen, 2003). It shows the spectro-temporal modulations (in cycles

per log-kHz for spectral modulation and in Hz for temporal modulations) that have high and low power in a given signal (corresponding to high occurrence and low occurrence). The MPS is invariant to translation and, unlike a spectrogram, can be averaged across samples of a signal to describe average properties. In a similar fashion, the Modulation Transfer Function (MTF) can be obtained from the 2d Fourier Transform of a spectrotemporal filter (here the STRF) and, without averaging, shows the tuning gain of the filter in the same space as the MPS.

STRF-Responsive Electrode Selection

We focused our analyses on electrodes located on the temporal lobe (particularly covering the Superior Temporal Gyrus (STG) and Superior Temporal Sulcus (STS). These regions of the brain respond to acoustic and linguistic features, and represent the best candidates for detecting a shift in spectrotemporal tuning. Responses then underwent several steps to exclude electrodes based on their non-significant and/or poorly fit responses.

The predictive score of all eSTRF models fit on unfiltered (MIDDLE) speech trials was calculated as the coefficient of determination (R^2) between predicted and actual HFB amplitude on held-out test data. For each electrode, we calculated the 99th percentile of model score across cross-validation splits. Electrodes whose lower-bound was greater than 0 were considered spectrotemporally-responsive (STRF-R, see **Figure 24** and Supplementary Figure 2).

To be included in the analysis, an electrode had to be Speech-R and STRF-R, and had to be located on the temporal lobe and in perisylvian regions (see Supplementary Figure 2).

eSTRF Comparisons

To detect an eSTRF tuning shift from the BEFORE to the AFTER condition, several analyses were carried out. The goal behind each was to compare the eSTRF properties in the BEFORE and AFTER conditions to the neural response in the MIDDLE speech condition. The primary aim of all analyses is to determine whether the subjective perceptual enhancement effect corresponds to a shift in spectrotemporal tuning to filtered speech.

MIDDLE Condition Coefficients Generalization

To assess the extent to which eSTRF plasticity improved the response to the speech signal, we estimated the extent to which coefficients fit in the MIDDLE condition (on unfiltered speech) generalized to the BEFORE and AFTER conditions. The eSTRF estimated in the MIDDLE condition was used to make predictions about the HFB activity in the BEFORE and AFTER condition. In this manner, one can determine the extent with which the spectrotemporal tuning estimated from unfiltered speech was a valid characterization of the tuning in each filtered condition. Predictions were obtained by convolving the eSTRF filter with the spectrogram in each filtered speech condition. We then compared the goodness of fit (R^2) between the predicted and actual HFB activity in the BEFORE and AFTER conditions.

eSTRF Unfiltered Speech Output Power Analysis

Next, the extent to which eSTRFs in the BEFORE and AFTER condition are responsive to spectrotemporal features of unfiltered speech was assessed. For this purpose, the predicted HFB response to unfiltered speech using the eSTRF in the BEFORE and AFTER conditions, and

calculated the output power of these predictions. This power reflects the extent to which the eSTRF overlapped with unfiltered speech features and, thus, is able at extract unfiltered speech sounds. To account for possible changes in the overall signal-to-noise ratio of eSTRFs in all condition, each eSTRF was standardized to zero mean and unit standard deviation before convolution. Then, the unfiltered speech spectrograms were passed through the eSTRF of the BEFORE and AFTER condition. Finally, the root-mean-squared amplitude of the output was calculated for each, and compared between the BEFORE and AFTER conditions (See **Figure 27**).

eSTRF Linear Overlap Partial Correlation Analysis

For each electrode in included in this analysis, the ensemble spectrotemporal receptive field (eSTRF) in each condition was calculated as the z-score for each spectro-temporal feature across CV splits. To test the hypothesis that spectrotemporal tuning shifts after hearing the unfiltered context speech, the partial correlation was calculated between sets of conditions. Partial correlation allows one to determine the correlation between two variables, conditioned on one or more other variables. It reflects the extent to which two variables are related in a manner that is linearly orthogonal to the conditioned variables, and can be represented in the following equation:

$$pcorr(a, b|c) = corr(\widehat{a_c} - a, b)$$

where $\widehat{a_c}$ is the predicted value of variable a regressed against variable c. In other words, one calculates partial correlation by regressing out all conditional variables, and then calculating the correlation between the residuals and the variable of interest. Partial correlation was calculated between the BEFORE/MIDDLE models and the AFTER/MIDDLE models with the following convention:

$$sim_{(bef,mid)} = pcorr(before, middle \mid after)$$
$$sim_{(aft,mid)} = pcorr(after, middle \mid before)$$

One would expect to find many similarities between the brain activity in response to the BEFORE and AFTER condition. As such, one wants to control for these similarities when calculating the correlation between before/after and the unfiltered condition. Partial correlation allows one to determine the extent to which one condition is correlated with the unfiltered condition, after removing the linear relationship with the other condition.

Between Condition Permutation Test Statistics

The following permutation-based procedure was used to compare a statistic for the difference between conditions. For each electrode, the statistic of choice (e.g., HFB amplitude or partial correlation) was computed for each condition. Then, the difference between conditions for each electrode was calculated. A null condition effect of condition would have values distributed around 0. Because of the paired nature of this design, one may simulate a permuted null distribution by randomly flipping the sign of all difference values, effectively randomizing values to condition A or B. The mean of the permuted difference vector was calculated as a single point in the null distribution. This procedure was repeated 10, 000 times to construct a null distribution against which the “true” difference vector mean is compared. Reported p-

values are the quantile for the difference vector mean with respect to this null distribution. All statistical tests are two-sided.

Data Availability

Raw data is stored in the Collaborative Research in Computational Neuroscience (CRCNS) database at UC Berkeley (crcns.org). It can be accessed with a free CRCNS account at crcns.org/data-sets.

This manuscript relied heavily on the Python packages MNE-python, scikit-learn, numpy, scipy, pandas, and matplotlib. Analyses were conducted using these packages, and the large majority have been aggregated as a python package hosted on github.

Code for performing statistical permutation tests is found in the MNE-python statistics module. Code for model fitting, feature extraction, statistics, and visualization can be found at github.com/choldgraf/ecogtools.

Chapter 5 – Concluding remarks and future work

This thesis describes an attempt to join methodology from the single unit and animal literature to study questions that are well-suited for human neuroscience. The goal is to provide a framework for modeling and statistical analysis into the world of electrocorticography, in the hopes that this allows one to ask more complex questions about the relationship between the human brain and how it interacts with the world around it. This approach uncovers a subtle relationship between the brain and sensory information, making it possible to construct a mathematical model that can generate predictions about new datapoints in our quest to understand how we communicate.

The thesis first described the considerations that one must take in applying these methods to cognitive neuroscience, as there are many nuances to doing predictive modeling properly. This section was written up as a publication in the open access journal *Frontiers in Systems Neuroscience* along with a collection of jupyter notebooks that are aimed at making these techniques more accessible to the scientific community and easy to implement on real data. Next, this predictive modeling approach was used to investigate its potential for speech decoding, a technique in which we attempt to predict the speech features that are currently represented in the brain, using only neural activity. We showed that using decoding models of neural activity, it is possible to reconstruct spectral features of sound as the brain responds to speech features. Moreover, we show that it may be possible to reconstruct the spectral content of *imagined* speech features that did not result from any external stimulus. These are the critical first steps necessary for future developments in the field of neural prosthetics. Finally, the paper described a use of the predictive modeling approach toward studying the plasticity of feature representations in the brain. This is one of the first attempts at explicitly describing the relationship between human brain activity and spectral features, and how this relationship changes due to previous experience with sound. We showed that regions of the auditory cortex change the way that they parse spectral features of sounds, and that this may help the brain find meaningful information in noisy signals.

Taken together, this work is the beginning of what will hopefully become a broader line of questioning in cognitive neuroscience. The encoding / decoding approach is an extremely flexible and powerful method of asking questions about how the human brain processes perceptual information, and there are many ways in which these studies could be extended. The following section details areas for future study in this line of questioning.

Future work

There are many advances to be made in using encoding and decoding models for understanding human brain activity. First, as discussed above a key goal of the predictive modeling approach is that it allows one to use and understand naturalistic stimuli. However, what constitutes “natural” in the context of stimuli? This question is often answered on a subjective basis, and there is not a satisfying and rigorous definition of “naturalness” in stimuli. It would be beneficial to define the stimulus qualities that together define a spectrum from “artificial” to “natural” stimuli. Moreover, there is not currently a single set of stimuli that have the properties (e.g.,

naturalistic, well-annotated) necessary for encoding and decoding models. As a result, researchers tend to pick and choose from existing corpora, which introduces variability due to the choice of stimulus set. TIMIT is "near-natural", but it is far from the experience of listening to everyday speech. These sentences are disjointed and semantically unrelated to one another, making it a suboptimal stimulus set for exploring interactions between high-level features and low-level representation that unfold in natural speech. It will be important for the field to agree upon a standard collection of stimuli for certain kinds of tasks in order to allow for comparisons between experiments.

It is also important to consider how models that are fit with different stimulus features can be compared with one another (ideally on the same dataset). The work discussed in this thesis focused on one particular representation of stimuli (spectro-temporal features), but there are many other options for use in encoding models. Future work should more explicitly address how a single acoustic stimulus may be represented in parallel at many different levels of complexity in the brain. This will make it possible to reveal the hierarchical representation of auditory features across a wide range of cortex.

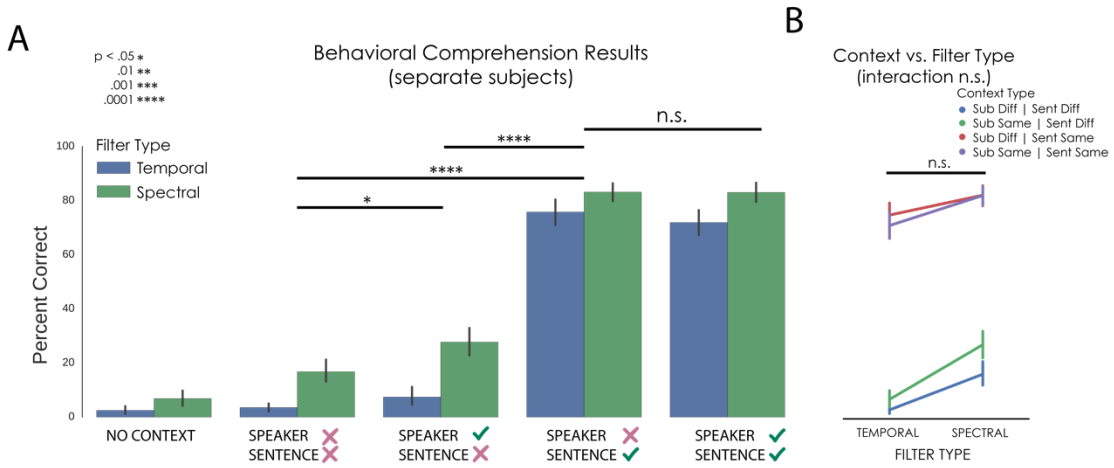
Another useful line of study will be to consider relationships *between* neural units instead of treating them as independent sources of information. Most encoding models are fit one per electrode, and the model parameters have no concept of interactions between electrodes. Future work should explicitly study how interactions between electrodes or brain regions underlie the plasticity in sensory representations reported in this manuscript. This may reveal details about the neural mechanism that underlies top-down sensory plasticity. It may also be possible to treat patterns of neural activity not as individual units (e.g., electrodes), but as coordinated patterns of activity across multiple regions simultaneously. Studying how these network-level patterns of activity relate to different stimulus features will help clarify how information is distributed throughout the brain, particularly with respect to more abstract, high-level features.

Finally, while the advances above address basic research in studying auditory perception in humans, they may also improve attempts at leveraging neural activity for the purposes of neuroprosthetic devices. There are many diseases and conditions which render an individual unable to communicate with the outside world, or which impair their ability to manipulate the world around them. For example, locked-in syndrome is a condition in which subjects maintain consciousness, but lack the ability to control any muscle movement, making traditional forms of communication impossible. Advances in relating neural activity to the outside world via predictive modeling may make it possible to generate actions in the world (for example, displaying a word on a screen) using neural signals generated by the patient and recorded with electrodes implanted in the brain. This will be an exciting new area of research with the potential to impact tens of thousands of lives.

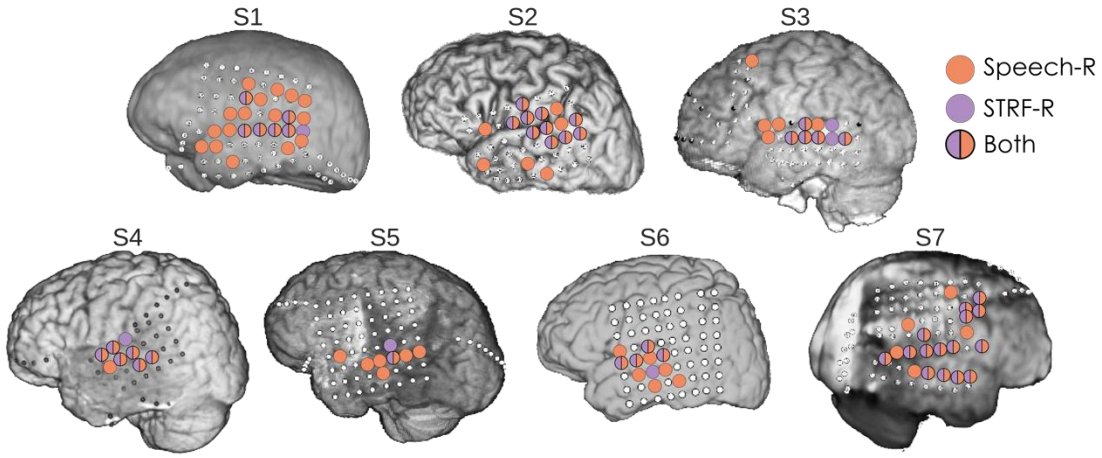
Supplemental Material

This material was created for the work described in Chapter 4. Among other things, it contains preliminary attempts at investigating interactions between regions of the brain during the reported speech enhancement. It also investigates the role of spectral content outside of the HFA region, and gives more detail about the findings described in Chapter 4.

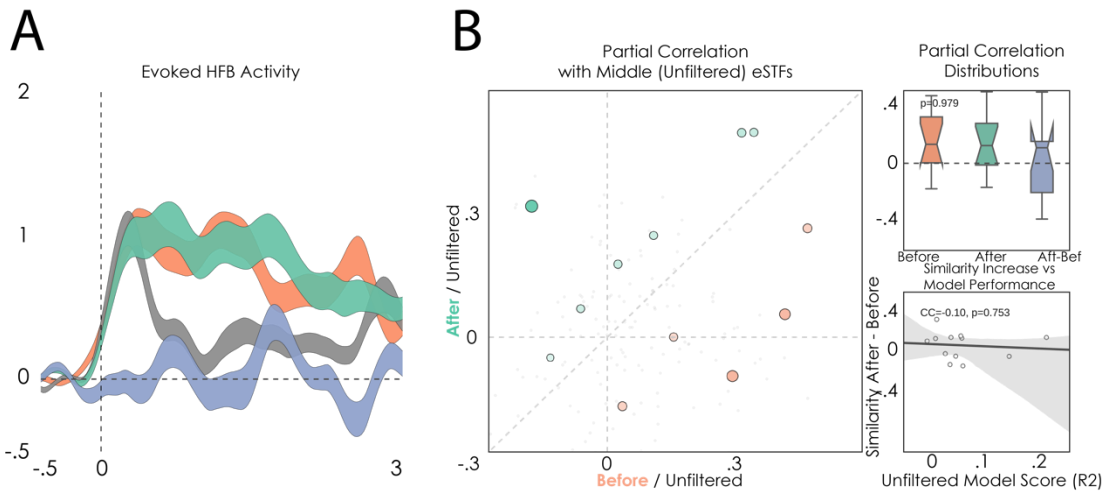
Supplementary Figures



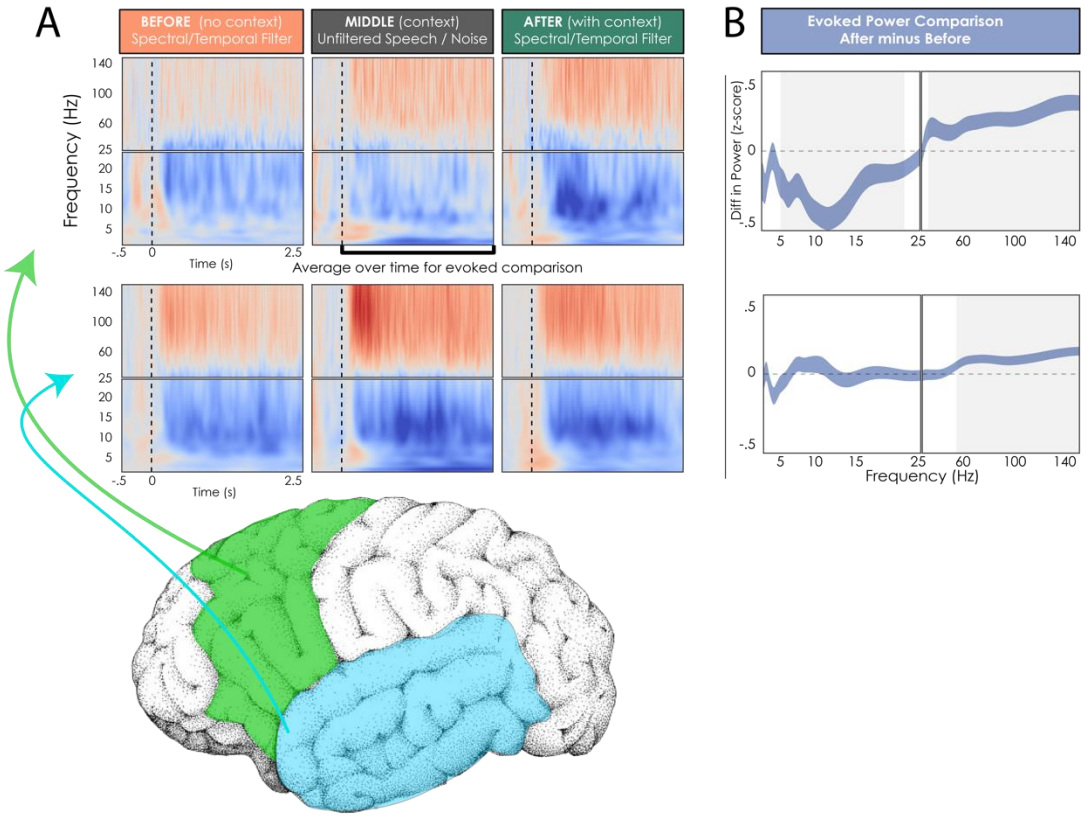
Supplementary Figure 1 - Behavioral task and controls
A behavioral task was conducted to assess the effect of filter type and context type on perceptual enhancement. We ran a sentence comprehension test with several groups of undergraduate students at UC Berkeley. All participants listened to combinations of filtered speech sentences and contextual sentences, and were asked to type out all words that they understood. Comparisons were made between the percentage of words correct in each group. (A) Bar plots show speech comprehension for various combinations of high/low context. First plot: a group of subjects responded to a single presentation of a filtered speech sentence, this is denoted the "no context" condition, and serves as a baseline for comprehension (corresponding to the BEFORE condition). Remaining plots, from left to right: different sentence/different speaker (to test the effect of repetition and potentially global arousal/activation of language network caused by speech), different sentence/same speaker (voice overlap that enhances the similarity with phonemes but not words and word order), same sentence/different speaker (phonemic and spectrally different, but with the same word identities spoken), and same sentence/same speaker (with spectral, phonemic, and word identity/rate overlap). Percentage reported correct are as follows. No Context: 4.5 +/- 0.8%. Diff Sentence, Diff Speaker: 10.4 +/- 1.2%. Diff Sentence, Same Speaker: 18.9 +/- 1.7%. Same Sentence, Different Speaker: 79.6 +/- 1.5%. Same Sentence, Same Speaker: 77.7 +/- 1.5%. Horizontal lines show significance values, and the following comparisons are all relative to the Different Speaker, Different Sentence condition: there was a main increase for Same Sentence, Different Speaker (t=9.65, df=22; showing the main perceptual enhancement effect). There is a small but significant increase of Same Speaker, Different Sentence (t=2.39, df=22; suggesting that acoustic properties of the speaker's voice is helpful in understanding the noisy stimulus). There is a much larger increase for the Same Sentence, Different Speaker condition (t=9.74, df=28; suggesting that linguistic properties of the speech are more important than acoustic properties in the speaker's voice). Finally, there is no significant difference between Same Sentence / Same Speaker and Same Sentence / Different Speaker (t=0.26, df=22; suggesting that the linguistic information shared between the two is responsible for the perceptual enhancement effect). (B) Shows the effect of filter type on perceptual enhancement in each context condition (means in each group +/- standard error). A linear mixed effects model (n observations=96 and n individuals=23) was used to calculate main effects of filter type and context on perceptual enhancement (as well as their interaction). There was a small and nonsignificant main effect of filter type (spectral > temporal, p=.082, confidence interval -41.61 to 2.47), and no significant interaction between filter type and context type, suggesting that the perceptual enhancement effect is similar across stimulus filters (interaction term, p=.61, confidence interval -5.64 to 10.73).



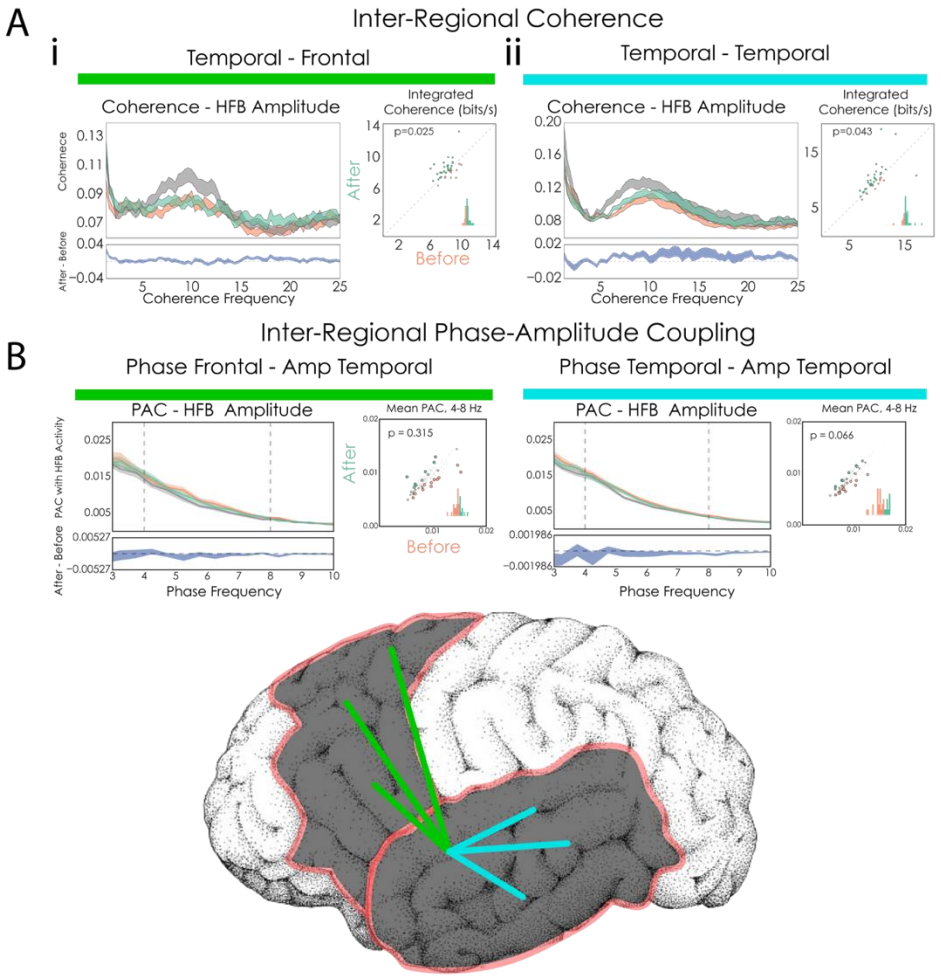
Supplementary Figure 2 - Comparison of Speech-R and STRF-R electrodes
Electrodes were characterized as responsive to speech (Speech-R) if their evoked HFB activity was significantly greater than 0 (z-score over baseline, confidence interval test across trials). Electrodes were characterized as well-modeled by spectro-temporal features (STRF-R) if their goodness of fit on held-out data was greater than 0 (confidence interval test across CV splits). Anatomical distribution of Speech-R electrodes (orange), STRF-R electrodes (purple) and electrodes responsive to both (split colors) are shown above.



Supplementary Figure 3 - Analysis during pink noise trials
A subset of subjects (n=3) performed a pink noise control on half of their trials. Instead of an unfiltered speech context in the MIDDLE condition, energy-matched pink noise was played. (A) HFB activity (mean +/- standard error across all electrodes included in the analysis) is shown. The difference (AFTER - BEFORE) is plotted in purple. See Figure 3 for details. (B) eSTRFs were calculated for electrodes during pink-noise conditions. The similarity between BEFORE/MIDDLE, and AFTER/MIDDLE eSTRFs was estimated using partial correlation coefficients (see methods and text) and these are shown in the middle scatterplot. There was no significant difference in partial correlation for BEFORE/MIDDLE vs. AFTER/MIDDLE conditions (permutation test, n=12). See Figure 10 for additional details.



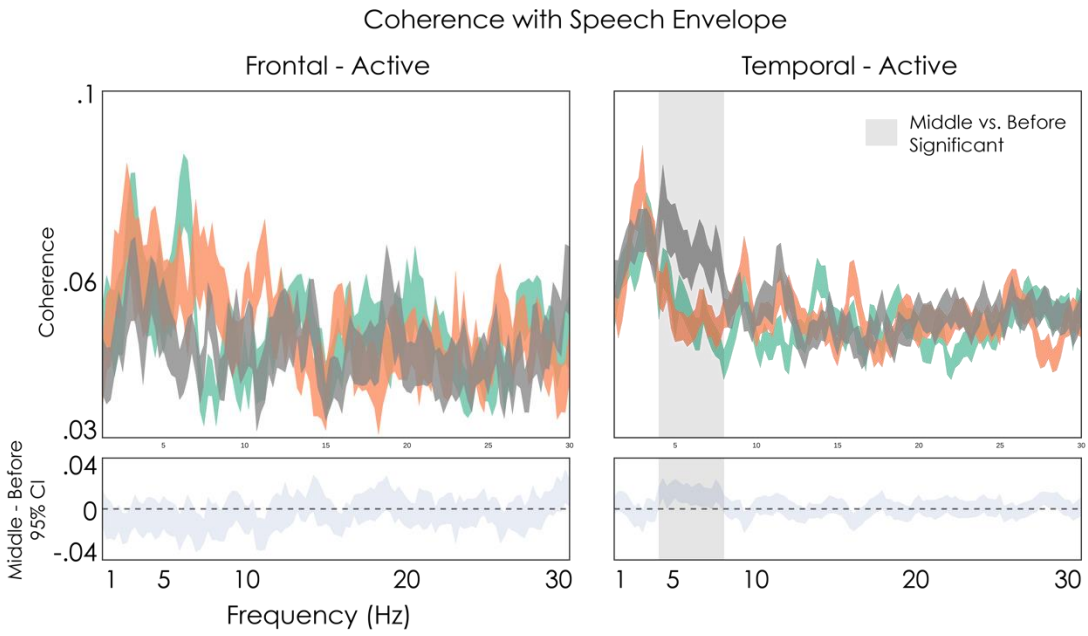
Supplementary Figure 4 - Time Frequency Response and Frequency Band Selection
For each electrode, the ECoG signal was convolved with 100 log-spaced Morlet wavelets (number of cycles fixed at 5) to create a time-varying power in each frequency band. Power was averaged within anatomical region and compared across conditions. (A) Mean TFR for Frontal (Top) and Temporal (Bottom) electrodes across conditions. (B) Evoked TFR in post-stimulus time points was averaged for frequency band selection. The mean +/- standard error is shown for each frequency. A cluster-based permutation test was used to find frequency-specific differences in power between the BEFORE and AFTER condition. Power in the high-frequency broadband (HFB) range significantly increased in the AFTER condition in both frontal and temporal electrodes (frontal, p=.001, n=75; temporal, p=.002, n=217). There was also a decrease in frequencies below 25 Hz in frontal electrodes (p=.001, n=75). This prompted further investigation of the HFB signal.



Supplementary Figure 5 - Connectivity Analyses

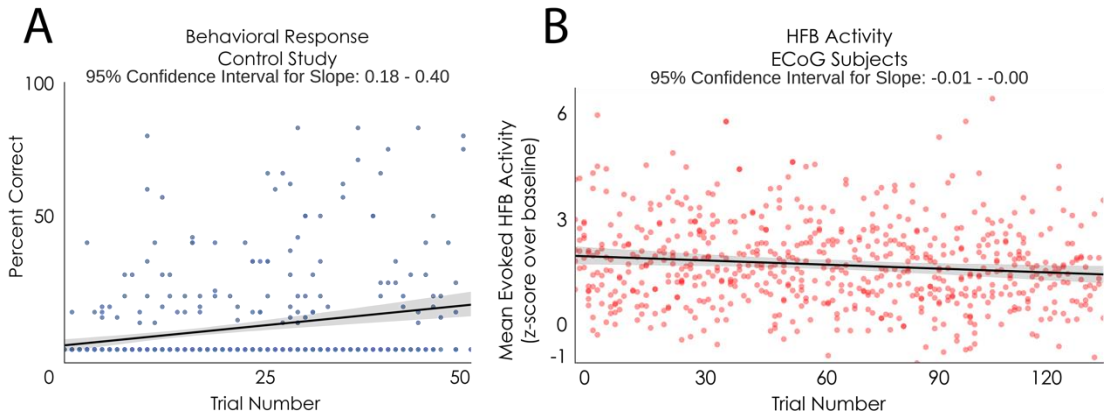
(A) Coherence was calculated between electrodes in regions of interest. Mean +/- standard error are shown across electrodes. We defined a set of "seed" electrodes in the temporal lobe that were Speech-R and STRF-R. For each seed, the coherence in the HFB amplitude was calculated between it and all other electrodes in the frontal (left, i) or temporal (right, ii) lobe. For each seed electrode, coherence values were averaged across targets and, converted into bits/second, and integrated across frequencies. These values are plotted in the scatterplots to the right along with p-values for the difference between AFTER and BEFORE (paired permutation test, n=39). We measured a small but significant increase in coherence for both temporal-frontal electrodes (p=.025), and temporal-temporal electrodes (p=.043). (B) Phase-Amplitude Coupling was calculated the regions described in A. The time-varying phase of theta-range frequencies (2 - 12Hz, .5Hz spacing) along with the mean amplitude of HFB frequencies (70 - 140Hz, 10Hz spacing) was calculated using band-pass filters followed by a Hilbert transform (and averaging frequency bands together in the case of the high-frequency amplitude). The strength of coupling was calculated using the Phase Amplitude Coupling measure defined in Ozkurt and Schnitzler, 2011 (see supplemental methods). The large line plots on the left show the mean +/- standard error PAC as a function of the frequency of the phase in the theta-range region. The smaller line plots below show the mean +/- standard error of the difference in PAC (AFTER - BEFORE). The scatterplots on the right show the mean PAC for the 4-8Hz phase in each condition. There was no significant change in PAC

between frontal and temporal electrodes in the AFTER condition relative to the BEFORE condition (paired permutation t-test).



Supplementary Figure 6 - Coherence with Speech Envelope

Since prior studies have shown that low frequencies in the ECoG signal tracked the speech envelope and that this tracking was modulated by attentional processes(Zion Golumbic et al., 2013), we also examined whether we could detect a similar effect in our study. The envelope of speech in each condition was calculated in each condition by averaging across power from 64 wavelets log-spaced from frequencies 500 to 2000Hz. Post-stimulus coherence (0s to 3s) between the raw ECoG signal and the speech envelope was calculated for all electrodes with HFB activity (Speech-R). Top row: mean +/- standard error coherence is plotted for frontal (left) and temporal (right) electrodes. Bottom row: the difference in condition (MIDDLE - BEFORE) is plotted (mean +/- standard error) across active electrodes. In the temporal lobe, there was a significant increase in theta coherence during unfiltered speech relative to the BEFORE condition (permutation cluster test, p=.001, n=72), but no significant difference between BEFORE and AFTER conditions. There were no significant effects in frontal electrodes. Other neurophysiological studies have correlated the signals detected in the lower frequencies of ECoG, MEG, or EEG with the envelope of human speech(Peelle et al., 2013). While our study confirms that the neural encoding of speech features changes with intelligibility, we found no change in coherence between theta activity and the speech envelope. This may be due to the different spectrotemporal properties of our filtered speech stimuli, and the fact that ECoG records signal from a different distribution of neural sources than MEG. It should be noted that these effects have generally been described in premotor/frontal regions, which did not have extensive electrode coverage in this study.



Supplementary Figure 7 - Behavioral and HFB change over trials

(A) In behavioral control subjects, the percent correct words are plotted as a function of trial number for subjects that heard either no context, or a Different Subject, Different Sentence context (testing the effect of repeated exposure to the filtered speech stimuli, n=14). We used a bootstrap technique to calculate confidence intervals on the slope of the line relating the percent words correct to the trial number. Bootstrapped regression coefficients found a slightly positive relationship between percent correct and trial number, suggesting a small effect of session duration on perceptual enhancement. (B) The same bootstrapped regression approach was applied to mean HFB activity in ECoG subjects, using electrodes that showed an increase in HFB activity to speech (Speech-R). Bootstrapped coefficients show a non-significant relationship between HFB activity and trial number.

Supplementary Methods

Connectivity between frontal/temporal electrodes

To assess putative top-down signals underlying the reported eSTRF plasticity, we conducted several connectivity analyses between frontal/premotor cortex and temporal cortex. We tested whether there was an increase in frontal/temporal coherence in the AFTER condition, whether delta-theta activity in the frontal and temporal lobes was phase-locked to the speech envelope, and whether directional Phase-Amplitude Coupling between delta-theta and HFB activity was increased in the AFTER condition. Connectivity results did not yield any conclusive findings. As electrode coverage for this study was based on temporal lobe coverage, not frontal/motor coverage, this data is not well suited for answering questions about intra-cortical connectivity and how it pertains to eSTRF plasticity. Further details for these connectivity analysis and results are found below.

Inter- and Intra-regional Coherence Analysis

To investigate putative higher-level regions that may be involved in eSTRF plasticity, we conducted connectivity analyses between anatomical regions of interest. Analyses were conducted with the HFB activity of electrodes included in eSTRF analysis as seeds (Speech-R and STRF-R, located on the temporal lobe, hereafter called Temporal seeds). We performed a separate analysis for two groups of target electrodes: all other temporal lobe electrodes (Temporal targets), and electrodes located on frontal/premotor regions (hereafter, Frontal targets). It should be noted that there was generally sparse electrode coverage in frontal/premotor regions, as grid cases were primarily selected for temporal lobe coverage.

To calculate the coherence between electrodes, we used a multi-taper windowing method similar to that described above. For each trial, the coherence was calculated between pairs of electrodes with seeds/targets based on anatomical regions of interest (and only including electrodes included in the eSTRF analysis as seeds). For each seed, coherence was averaged across target electrodes. Next, the mean coherence was converted to normal mutual information by integrating across frequencies. The difference in condition (AFTER – BEFORE) was calculated for each electrode, and a permutation t-test was conducted across electrodes to test for a difference in condition. Coherence calculation was performed with the MNE-toolbox (see Data Availability in the main manuscript for information about how to access this code).

The coherence was first calculated between Temporal seeds and Frontal targets. The results are shown in **Supplementary Figure 5**A. There was a frequency peak around 9-10Hz, and a weak general increase in coherence in the AFTER condition over the BEFORE condition. For Temporal seeds and Temporal targets, there was a broadband increase in coherence across frequencies greater than 5Hz. (permutation test, see **Supplementary Figure 5**A for discussion).

Theta Connectivity Phase-Amplitude Coupling Analysis

Next, we investigated the role of the theta band in modulating HFB activity. It has been suggested that the phase of theta activity may modulate the amplitude of HFB activity, representing a neural mechanism by which associative cortical areas influence the processing

occurring in sensory cortex(Canolty et al., 2006; Voytek et al., 2015). We calculated the Phase-Amplitude Coupling (PAC) between electrodes included in the eSTRF analysis and groups of electrodes either in temporal or frontal/pre-motor regions.

To test for cross-frequency effects between theta phase and HFB amplitude, we split electrodes into groups to be analyzed for phase and for amplitude. To calculate phase, we band-pass filtered the signal from each electrode, then calculated the time-varying phase of the Hilbert transform of each of the band-limited signals. We calculated the phase for frequencies from 3Hz to 10Hz in increments of 0.5Hz. To calculate the HFB amplitude of each signal, we again performed a band-pass filter of the raw signal for 10 logarithmically-spaced bands from 70-140Hz. For each band we calculated the modulus of the Hilbert transform and averaged the bands together.

We calculated the Phase-Amplitude Coupling using the method described in Ozkurt and Schnitzler, 2011. This is form of the Modulation Index that is normalized by the amplitudes of the two filtered signals(Ozkurt and Schnitzler, 2011). For each condition, we concatenated the phases/amplitudes of the pair of electrodes across trials, then calculated a single value for PAC between electrodes. To test for a difference in condition (AFTER – BEFORE), we took the average PAC for phases from 3-8Hz, and the amplitude from the HFB signal, and calculated the difference AFTER – BEFORE. For each HFB amplitude electrode, we averaged the PAC value across all other theta phase electrodes to calculate a single value of PAC for each HFB amplitude electrode. Significance for the difference in condition (AFTER – BEFORE) was assessed with a permutation t-test for a difference from 0 (see Data Availability in the main manuscript for information about how to access this code).

As previously mentioned this study did not include dense coverage over premotor/frontal regions, and the reported effect sizes are small, precluding a conclusive result. Future studies should investigate the putative link between frontal and temporal electrodes in top-down mechanisms of speech perception.

Electrode coherence with speech envelope

There have been several studies suggesting a role of theta band activity in parsing speech utterances, especially in noise(Ding and Simon, 2014; Peelle et al., 2013). Previous studies have suggested that coherence between neural activity and the speech envelope increases in the theta band during intelligible speech. This entrainment has been interpreted as the tracking of the rhythmic structure of an attended speech stimulus by associative cortex, which facilitates speech processing. To investigate whether these effects are modulated by experience with intact speech, we calculated the coherence between the raw electrode signal and the speech envelope.

For each trial, we calculated the envelope of the speech stimulus by first performing a time-frequency decomposition of the audio waveform using 64 log-spaced frequency Gabor wavelets with center frequencies from 500 to 2000Hz. The 64 band-passed signals were rectified and then averaged across frequencies.

To calculate the coherence between ECoG activity and the envelope of each speech stimulus, we again used a multi-taper windowing method. For each trial, we calculated the coherence between electrodes of interest and the speech envelope. We converted this value to normal mutual information. To find particular frequencies that showed a difference in electrode-envelope coherence between conditions, we calculated the difference between conditions for each frequency band. We tested the difference in condition (AFTER – BEFORE, and MIDDLE – BEFORE) by conducting a cluster-based permutation t-test for a difference from 0 (see **Supplementary Figure 6**).

References

Introduction

Atencio, C. A., Sharpee, T. O., & Schreiner, C. E. (2012). Receptive field dimensionality increases from the auditory midbrain to cortex. *Journal of Neurophysiology*, *107*(10), 2594–2603. https://doi.org/10.1152/jn.01025.2011

Atiani, S., David, S. V., Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S. A., & Fritz, J. B. (2014). Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. *Neuron*, *82*(2), 486–99. https://doi.org/10.1016/j.neuron.2014.02.029

Block, N., & Siegel, S. (2013). Attention and perceptual adaptation. *The Behavioral and Brain Sciences*, *36*(3), 205–6. https://doi.org/10.1017/S0140525X12002245

Coull, J. T., Frith, C. D., Büchel, C., & Nobre, a C. (2000). Orienting attention in time: behavioural and neuroanatomical distinction between exogenous and endogenous shifts. *Neuropsychologia*, *38*(6), 808–19. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10689056

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology. Human Perception and Performance*, *30*(4), 643–56. https://doi.org/10.1037/0096-1523.30.4.643

David, S. V., Fritz, J. B., & Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(6), 2144–9. https://doi.org/10.1073/pnas.1117717109

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–47. https://doi.org/10.1016/j.heares.2007.01.014

Depireux, D. a, Simon, J. Z., Klein, D. J., & Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology*, *85*(3), 1220–1234.

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(13), 5728–35. https://doi.org/10.1523/JNEUROSCI.5297-12.2013

Eggermont, J. J. (2001). Between sound and perception: reviewing the search for a neural code. *Hearing Research*, *157*(1–2), 1–42. https://doi.org/10.1016/S0378-5955(01)00259-3

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., & Giraud, A.-L. (2014). The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nature Communications*, *5*(May), 4694. https://doi.org/10.1038/ncomms5694

Fritz, J. B., Shamma, S. A., Elhilali, M., & Klein, D. J. (2003). Rapid task-related plasticity of

spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, *6*(11), 1216–1223. https://doi.org/10.1038/nn1141

Gilbert, C. D., & Sigman, M. (2007). Brain states: top-down influences in sensory processing. *Neuron*, *54*(5), 677–96. https://doi.org/10.1016/j.neuron.2007.05.019

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Gross, J., Hoogenboom, N., Thut, G., Schyns, P. G., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLoS Biology*, *11*(12). https://doi.org/10.1371/journal.pbio.1001752

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *36*(6), 2014–26. https://doi.org/10.1523/JNEUROSCI.1779-15.2016

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron*, *77*(4), 750–61. https://doi.org/10.1016/j.neuron.2012.11.034

Martin, S., Brunner, P., Holdgraf, C. R., Heinze, H.-J., Crone, N. E., Rieger, J. W., … Pasley, B. N. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering*, *7*(May), 14. https://doi.org/10.3389/fneng.2014.00014

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–6. https://doi.org/10.1038/nature11020

Mesgarani, N., Slaney, M., & Shamma, S. A. (2006). Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(3), 920–930. https://doi.org/10.1109/TSA.2005.858055

Miller, K. J., Zanos, S., Fetz, E. E., den Nijs, M., & Ojemann, J. G. (2009). Decoupling the cortical power spectrum reveals real-time representation of individual finger movements in humans. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(10), 3132–7. https://doi.org/10.1523/JNEUROSCI.5506-08.2009

Miller, L. M., Escabí, M. a, Read, H. L., & Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology*, *87*(1), 516–527.

Moore, R. C., Lee, T., & Theunissen, F. E. (2013). Noise-invariant Neurons in the Avian Auditory Cortex: Hearing the Song in Noise. *PLoS Computational Biology*, *9*(3). https://doi.org/10.1371/journal.pcbi.1002942

Nelken, I., Chechik, G., Mrsic-Flogel, T. D., King, A. J., & Schnupp, J. W. H. (2005). Encoding stimulus information by spike numbers and mean response time in primary auditory

cortex. *Journal of Computational Neuroscience*, *19*(2), 199–221. https://doi.org/10.1007/s10827-005-1739-3

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Nathan, E., … Chang, E. F. (2012). Reconstructing Speech from Human Auditory Cortex. *PLoS Biology*, *10*(1). https://doi.org/10.1371/journal.pbio.1001251

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex (New York, N.Y. : 1991)*, *23*(June), 1378–87. https://doi.org/10.1093/cercor/bhs118

Rabinowitz, N. C., Willmore, B. D. B., King, A. J., & Schnupp, J. W. H. (2013). Constructing Noise-Invariant Representations of Sound in the Auditory Pathway. *PLoS Biology*, *11*(11). https://doi.org/10.1371/journal.pbio.1001710

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., & King, A. J. (2011). Contrast gain control in auditory cortex. *Neuron*, *70*(6), 1178–91. https://doi.org/10.1016/j.neuron.2011.04.030

Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J., & Hsiao, S. S. (2008). Neural correlates of high-gamma oscillations (60-200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *28*(45), 11526–36. https://doi.org/10.1523/JNEUROSCI.2848-08.2008

Ray, S., & Maunsell, J. H. R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biology*, *9*(4), e1000610. https://doi.org/10.1371/journal.pbio.1000610

Schroeder, C. E., Wilson, D. a, Radman, T., Scharfman, H., & Lakatos, P. (2010). Dynamics of Active Sensing and perceptual selection. *Current Opinion in Neurobiology*, *20*(2), 172–6. https://doi.org/10.1016/j.conb.2010.02.010

Shamma, S. A., & Fritz, J. B. (2014). Adaptive auditory computations. *Current Opinion in Neurobiology*, *25C*, 164–168. https://doi.org/10.1016/j.conb.2014.01.011

Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, *15*(6), 355–366. https://doi.org/10.1038/nrn3731

Theunissen, F. E., Sen, K., & Doupe, a J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *20*(6), 2315–2331. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10704507

Vouloumanos, a, Kiehl, K. a, Werker, J. F., & Liddle, P. F. (2001). Detection of sounds in the auditory stream: event-related fMRI evidence for differential activation to speech and nonspeech. *Journal of Cognitive Neuroscience*, *13*(7), 994–1005. https://doi.org/10.1162/089892901753165890

Wodlinger, B., Degenhart, A. D., Collinger, J. L., Tyler-Kabara, E. C., & Wang, W. (2011). The impact of electrode characteristics on electrocorticography (ECoG). *Conference*

Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, *2011*(1), 3083–6. https://doi.org/10.1109/IEMBS.2011.6090842

Woolley, S. M. N., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, *8*(10), 1371–9. https://doi.org/10.1038/nn1536

Yin, P., Fritz, J. B., & Shamma, S. A. (2014). Rapid Spectrotemporal Plasticity in Primary Auditory Cortex during Behavior. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(12), 4396–408. https://doi.org/10.1523/JNEUROSCI.2799-13.2014

Chapter 2

Aertsen, A. M. H. J., & Johannesma, P. I. M. (1981). The spectro-temporal receptive field. *Biological Cybernetics*, *42*, 133–143. Retrieved from http://www.springerlink.com/index/N1J6Q78Q6N7843H4.pdf

Ahrens, M. B., Paninski, L., & Sahani, M. (2008). *Inferring input nonlinearities in neural encoding models. Network (Bristol, England)* (Vol. 19). https://doi.org/10.1080/09548980701813936

Andoni, S., & Pollak, G. D. (2011). Selectivity for Spectral Motion as a Neural Computation for Encoding Natural Communication Signals in Bat Inferior Colliculus. *Journal of Neuroscience*, *31*(46), 16529–16540. https://doi.org/10.1523/JNEUROSCI.1306-11.2011

Bell, A. J., & Sejnowski, T. J. (1995). An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, *7*(6), 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Bennett, C. M., Baird, A., Miller, M. B., & Wolfrod, G. L. (2009). Neural correlates of interspieces perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. *Human Brain Mapping*, *1*, 1995. https://doi.org/10.1016/S1053-8119(09)71202-9

Blakely, T., Miller, K. J., Rao, R. P. N., Holmes, M. D., & Ojemann, J. G. (2008). Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids. *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, *2008*, 4964–4967. https://doi.org/10.1109/IEMBS.2008.4650328

Bouchard, K. E., & Chang, E. F. (2014). Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography. *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2014*, 6782–6785. https://doi.org/10.1109/EMBC.2014.6945185

Bressler, S. L., & Seth, A. K. (2011). Wiener-Granger Causality: A well established methodology. *NeuroImage*, *58*(2), 323–329. https://doi.org/10.1016/j.neuroimage.2010.02.059

Brumberg, J. S., Nieto-Castanon, A., Kennedy, P. R., & Guenther, F. H. (2010). Brain-Computer Interfaces for Speech Communication. *Speech Communication*, *52*(4), 367–379. https://doi.org/10.1016/j.specom.2010.01.001

Brumberg, Wright, Andreasen, D. S., Guenther, & Kennedy. (2011). Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. *Frontiers in Neuroscience*. https://doi.org/10.3389/fnins.2011.00065

Campbell, D. T., & Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research.* Ravenio Books.

Chakrabarti, S., Krusienski, D. J., Schalk, G., & Brumberg, J. S. (2013). Predicting mel-frequency cepstral coefficients from electrocorticographic signals during continuous speech

production. *6th International IEEE/EMBS Conference on Neural Engineering (NER)*.

Chang, E. F., Edwards, E., Nagarajan, S. S., Fogelson, N., Dalal, S. S., Canolty, R. T., … Knight, R. T. (2011). Cortical spatio-temporal dynamics underlying phonological target detection in humans. *Journal of Cognitive Neuroscience*, *23*(6), 1437–46. https://doi.org/10.1162/jocn.2010.21466

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428–32. https://doi.org/10.1038/nn.2641

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887. https://doi.org/10.1121/1.1945807

Christianson, G. B., Sahani, M., & Linden, J. F. (2008). The Consequences of Response Nonlinearities for Interpretation of Spectrotemporal Receptive Fields. *Journal of Neuroscience*, *28*(2), 446–455. https://doi.org/10.1523/JNEUROSCI.1775-07.2007

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*(November), 604. https://doi.org/10.3389/fnhum.2016.00604

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, (April). https://doi.org/10.1038/nn.3381

Curran-Everett, D. (2000). Multiple comparisons: philosophies and illustrations. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology*, *279*(1), R1-8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10896857

David, S. V. (2004). Natural Stimulus Statistics Alter the Receptive Field Structure of V1 Neurons. *Journal of Neuroscience*, *24*(31), 6991–7006. https://doi.org/10.1523/JNEUROSCI.1422-04.2004

David, S. V., & Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network: Computation in Neural Systems*, *16*(2–3), 239–260. https://doi.org/10.1080/09548980500464030

David, S. V., & Shamma, S. A. (2013). Integration over Multiple Timescales in Primary Auditory Cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(49), 19154–66. https://doi.org/10.1523/JNEUROSCI.2270-13.2013

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *The Journal of Neuroscience*, 3267–16. https://doi.org/10.1523/JNEUROSCI.3267-16.2017

Degenhart, A. D., Sudre, G. P., Pomerleau, D. A., & Tyler-Kabara, E. C. (2011). Decoding semantic information from human electrocorticographic (ECoG) signals (pp. 6294–6298). IEEE. https://doi.org/10.1109/IEMBS.2011.6091553

Depireux, D. a, Simon, J. Z., Klein, D. J., & Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology*, *85*(3), 1220–1234.

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(8), E505-14. https://doi.org/10.1073/pnas.1113427109

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology*, *25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*, 412–431. https://doi.org/10.1016/0001-6918(69)90065-1

Eggermont, J. J. (1993). Wiener and Volterra analyses applied to the auditory system. *Hearing Research*, *66*(2), 177–201. https://doi.org/10.1016/0378-5955(93)90139-R

Eggermont, J. J. (2001). Between sound and perception: reviewing the search for a neural code. *Hearing Research*, *157*(1–2), 1–42. https://doi.org/10.1016/S0378-5955(01)00259-3

Eggermont, J. J., Johannesma, P. I. M., & Aertsen, A. M. H. J. (1983). Reverse-correlation methods in auditory research. *Quarterly Reviews of Biophysics*, *16*(3), 341. https://doi.org/10.1017/S0033583500005126

Elie, J. E., & Theunissen, F. E. (2016). The vocal repertoire of the domesticated zebra finch: a data-driven approach to decipher the information-bearing acoustic features of communication signals. *Animal Cognition*, *19*(2), 285–315. https://doi.org/10.1007/s10071-015-0933-6

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*, *5*(3), e1000302. https://doi.org/10.1371/journal.pcbi.1000302

Escabi, M. A., & Schreiner, C. E. (2002). Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *22*(10), 4114–31. https://doi.org/20026325

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierachical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature Neuroscience*, *8*(12), 1643–6. https://doi.org/10.1038/nn1608

Fitzgerald, J. D., Sincich, L. C., & Sharpee, T. O. (2011). Minimal models of multidimensional computations. *PLoS Computational Biology*, *7*(3). https://doi.org/10.1371/journal.pcbi.1001111

Friston, K. J. (2003). Introduction: Experimental design and Statistical Parametric Mapping. *SPM Introduction*. https://doi.org/10.1016/B978-012693019-1/50024-1

Fritz, J. B., Elhilali, M., & Shamma, S. A. (2005). Differential dynamic plasticity of A1 receptive

fields during multiple spectral tasks. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *25*(33), 7623–35. https://doi.org/10.1523/JNEUROSCI.1318-05.2005

Fritz, J. B., Shamma, S. A., Elhilali, M., & Klein, D. J. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, *6*(11), 1216–1223. https://doi.org/10.1038/nn1141

Frye, M., Micheli, C., Schepers, I. M., Schalk, G., Rieger, J. W., & Meyer, B. T. (2016). Neural responses to speech-specific modulations derived from a spectro-temporal filter bank. In *Proc Interspeech*.

Green, D. M., & Swets, J. A. (1988). *Signal Detection Theory and Psychophysics*. Peninsula Publishing. Retrieved from https://books.google.com/books?id=AjWwQgAACAAJ

Güçlü, U., & van Gerven, M. A. J. (2014). Unsupervised Feature Learning Improves Prediction of Human Brain Activity in Response to Natural Images. *PLoS Computational Biology*, *10*(8). https://doi.org/10.1371/journal.pcbi.1003724

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, *7*(1), 37–53. https://doi.org/10.1007/s12021-008-9041-y

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Haufe, S., Meinecke, F., G??rgen, K., D??hne, S., Haynes, J. D., Blankertz, B., & Bie??mann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Henniges, M., & Puertas, G. (2010). Binary Sparse Coding, 450–457.

Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 578–589. https://doi.org/10.1109/89.326616

Hickok, G., & Small, S. L. (2015). *Neurobiology of language*. Academic Press.

Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J., … Donoghue, J. P. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, *485*(7398), 372–375. https://doi.org/10.1038/nature11076

Holdgraf, C. R., de Heer, W., Pasley, B. N., Rieger, J. W., Crone, N., Lin, J. J., … Theunissen, F. E. (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature Communications*, *7*(May), 13654. https://doi.org/10.1038/ncomms13654

Hollmann, M., Rieger, J. W., Baecke, S., Lützkendorf, R., Müller, C., Adolf, D., & Bernarding, J. (2011). Predicting decisions in human social interactions using real-time fMRI and pattern classification. *PLoS ONE*, *6*(10). https://doi.org/10.1371/journal.pone.0025304

Hsu, A., Borst, A., & Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation in Neural*

Systems, *15*(2), 91–109. https://doi.org/10.1088/0954-898X/15/2/002

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *36*(6), 2014–26. https://doi.org/10.1523/JNEUROSCI.1779-15.2016

Huth, A. G., Heer, W. A. De, Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, *76*(6), 1210–1224. https://doi.org/10.1016/j.neuron.2012.10.014

Kay, K. N., & Gallant, J. L. (2009). I can see what you see. *Nature Neuroscience*, *12*(3), 245. https://doi.org/10.1038/nn0309-245

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–5. https://doi.org/10.1038/nature06713

Kellis, S., Miller, K. J., Thomson, K., Brown, R., House, P., & Greger, B. (2010). Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of Neural Engineering*, *7*(5), 56007. https://doi.org/10.1088/1741-2560/7/5/056007

Khalighinejad, B., Cruzatto da Silva, G., & Mesgarani, N. (2017). Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. *The Journal of Neuroscience*, *37*(8), 2176–2185. https://doi.org/10.1523/JNEUROSCI.2383-16.2017

Kiang, N. (1984). Peripheral neural processing of auditory information. *Comprehensive Physiology*, 639–674.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The Tracking of Speech Envelope in the Human Cortex. *PLoS ONE*, *8*(1), e53398. https://doi.org/10.1371/journal.pone.0053398

Leonard, M. K., Bouchard, K. E., Tang, C., & Chang, E. F. (2015). Dynamic Encoding of Speech Sequence Probability in Human Temporal Cortex. *Journal of Neuroscience*, *35*(18), 7203–7214. https://doi.org/10.1523/JNEUROSCI.4100-14.2015

Lescroart, M. D., Kanwisher, N., & Golomb, J. D. (2016). No Evidence for Automatic Remapping of Stimulus Features or Location Found with fMRI. *Frontiers in Systems Neuroscience*, *10*, 53. https://doi.org/10.3389/fnsys.2016.00053

Lescroart, M. D., Stansbury, D. E., & Gallant, J. L. (2015). Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. *Frontiers in Computational Neuroscience*, *9*(November), 135. https://doi.org/10.3389/fncom.2015.00135

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), 356–63. https://doi.org/10.1038/nn831

Lotte, F., Brumberg, J. S., Brunner, P., Gunduz, A., Ritaccio, A. L., Guan, C., & Schalk, G. (2015). Electrocorticographic representations of segmental features in continuous speech. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00097

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–90. https://doi.org/10.1016/j.jneumeth.2007.03.024

Marmarelis, P. Z., & Marmarelis, V. Z. (1978). *Analysis of Physiological Systems*. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4613-3970-0

Martin, S., Brunner, P., Holdgraf, C. R., Heinze, H.-J., Crone, N. E., Rieger, J. W., … Pasley, B. N. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering*, *7*(May), 14. https://doi.org/10.3389/fneng.2014.00014

Martin, S., Brunner, P., Iturrate, I., Millán, J. del R., Schalk, G., Knight, R. T., & Pasley, B. N. (2016). Word pair classification during imagined speech using direct brain recordings. *Scientific Reports*, *6*, 25803. https://doi.org/10.1038/srep25803

McFarland, J. M., Cui, Y., & Butts, D. A. (2013). Inferring Nonlinear Neuronal Computation Based on Physiologically Plausible Inputs. *PLoS Computational Biology*, *9*(7). https://doi.org/10.1371/journal.pcbi.1003143

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–6. https://doi.org/10.1038/nature11020

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2009). Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex. *Journal of Neurophysiology*, *102*(6), 3329–3339. https://doi.org/10.1152/jn.91128.2008

Mesgarani, N., Slaney, M., & Shamma, S. A. (2006). Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(3), 920–930. https://doi.org/10.1109/TSA.2005.858055

Meyer, A. F., Diepenbrock, J.-P., Ohl, F. W., & Anemüller, J. (2014). Temporal variability of spectro-temporal receptive fields in the anesthetized auditory cortex. *Frontiers in Computational Neuroscience*, *8*(DEC), 165. https://doi.org/10.3389/fncom.2014.00165

Meyer, A. F., Williamson, R. S., Linden, J. F., & Sahani, M. (2017). Models of Neuronal Stimulus-Response Functions: Elaboration, Estimation, and Evaluation. *Frontiers in Systems Neuroscience*, *10*(January), 1–25. https://doi.org/10.3389/fnsys.2016.00109

Miller, L. M., Escabí, M. a, Read, H. L., & Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology*, *87*(1), 516–527.

Moerel, M., De Martino, F., Santoro, R., Ugurbil, K., Goebel, R., Yacoub, E., & Formisano, E. (2013). Processing of natural sounds: characterization of multipeak spectral tuning in human auditory cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(29), 11888–98. https://doi.org/10.1523/JNEUROSCI.5306-12.2013

Moreno-Bote, R., Beck, J. M., Kanitscheider, I., Pitkow, X., Latham, P. E., & Pouget, A. (2014). Information-limiting correlations. *Nature Neuroscience*, *17*(10), 1410–1417. https://doi.org/10.1038/nn.3807

Mugler, E. M., Patton, J. L., Flint, R. D., Wright, Z. A., Schuele, S. U., Rosenow, J., … Slutzky, M. W. (2014). Direct classification of all American English phonemes using signals from functional speech motor cortex. *Journal of Neural Engineering*, *11*(3), 35015. https://doi.org/10.1088/1741-2560/11/3/035015

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–10. https://doi.org/10.1016/j.neuroimage.2010.07.073

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902–15. https://doi.org/10.1016/j.neuron.2009.09.006

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology : CB*, *21*(19), 1641–6. https://doi.org/10.1016/j.cub.2011.08.031

Norman, K. a, Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–30. https://doi.org/10.1016/j.tics.2006.07.005

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an incomplete basis set: a strategy employed by \protect{V1}. *Vision Research*. https://doi.org/http://dx.doi.org/10.1016/S0042-6989(97)00169-7

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–7. https://doi.org/10.1016/j.conb.2004.07.007

Paninski, L. (2003). Convergence properties of three spike-triggered analysis techniques. *Network: Computation in Neural Systems*, *14*(3), 437–464. https://doi.org/10.1088/0954-898X/14/3/304

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, *15*(4), 243–262. https://doi.org/10.1088/0954-898X/15/4/002

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Nathan, E., … Chang, E. F. (2012). Reconstructing Speech from Human Auditory Cortex. *PLoS Biology*, *10*(1). https://doi.org/10.1371/journal.pbio.1001251

Pasley, B. N., & Knight, R. T. (2012). Decoding speech for understanding and treating aphasia. *Progress in Brain Research*, *207*, 435–56. https://doi.org/10.1016/B978-0-444-63327-9.00018-7

Pedregosa, F., Grisel, O., Weiss, R., Passos, A., & Brucher, M. (2011). Scikit-learn : Machine Learning in Python, *12*, 2825–2830.

Pei, X., Barbour, D. L., Leuthardt, E. C., & Schalk, G. (2011). Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of Neural Engineering*, *8*(4), 46028. https://doi.org/10.1088/1741-2560/8/4/046028

Pillow, J. W., Ahmadian, Y., & Paninski, L. (2011). Model-based decoding, information estimation, and change-point detection techniques for multineuron spike trains. *Neural Computation*, *23*(1), 1–45. https://doi.org/10.1162/NECO_a_00058

Poeppel, D., Emmorey, K., Hickok, G., & Pylkkänen, L. (2012). Towards a new neurobiology of language. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(41), 14125–31. https://doi.org/10.1523/JNEUROSCI.3244-12.2012

Pulvermüller, F., Lutzenberger, W., & Preissl, H. (1999). Nouns and verbs in the intact brain: evidence from event-related potentials and high-frequency cortical responses. *Cerebral Cortex (New York, N.Y. : 1991)*, *9*, 497–506. https://doi.org/10.1093/cercor/9.5.497

Qiu, A., Schreiner, C. E., & Escabí, M. a. (2003). Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition. *Journal of Neurophysiology*, *90*(1), 456–476. https://doi.org/10.1152/jn.00851.2002

Quandt, F., Reichert, C., Hinrichs, H., Heinze, H.-J., Knight, R. T., & Rieger, J. W. (2012). Single trial discrimination of individual finger movements on one hand: A combined MEG and EEG study. *NeuroImage*, *59*(4), 3316–3324. https://doi.org/10.1016/j.neuroimage.2011.11.053

Ray, S., & Maunsell, J. H. R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biology*, *9*(4), e1000610. https://doi.org/10.1371/journal.pbio.1000610

Reichert, C., Fendrich, R., Bernarding, J., Tempelmann, C., Hinrichs, H., & Rieger, J. W. (2014). Online tracking of the contents of conscious perception using real-time fMRI. *Frontiers in Neuroscience*, *8*(8 MAY), 1–11. https://doi.org/10.3389/fnins.2014.00116

Rieger, J. W., Reichert, C., Gegenfurtner, K. R., Noesselt, T., Braun, C., Heinze, H.-J., … Hinrichs, H. (2008). Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. *NeuroImage*, *42*(3), 1056–1068. https://doi.org/10.1016/j.neuroimage.2008.06.014

Sahani, M., & Linden, J. F. (2003). How linear are auditory cortical responses? *Advances in Neural Information Processing Systems*, *15*, 109–116. Retrieved from http://books.google.com/books?hl=en&lr=&id=AAVSDw4Rw9UC&oi=fnd&pg=PA125&dq=How+Linear+are+Auditory+Cortical+Responses+?&ots=U5sfvDovuW&sig=jWDdamuI63UHcshFNToE9U3rcC4

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLoS Computational Biology*, *10*(1).

https://doi.org/10.1371/journal.pcbi.1003412

Schwartz, O., Pillow, J. W., Rust, N. C., & Simoncelli, E. P. (2006). Spike-triggered neural characterization. *Journal of Vision*, *6*(4), 484–507. https://doi.org/10.1167/6.4.13

Sen, K., Theunissen, F. E., & Doupe, A. J. (2001). Feature analysis of natural sounds in the songbird auditory forebrain. *Journal of Neurophysiology*, *86*(3), 1445–58. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11535690

Shamma, S. A. (2015). Spectrotemporal Receptive Fields. *Encyclopedia of Computational Neuroscience*, 2794–2798.

Sharpee, T. O. (2016). How invariant feature selectivity is achieved in cortex. *Frontiers in Synaptic Neuroscience*, *8*(AUG), 1–7. https://doi.org/10.3389/fnsyn.2016.00026

Sharpee, T. O., Atencio, C. A., & Schreiner, C. E. (2011). Hierarchical representations in the auditory cortex. *Current Opinion in Neurobiology*, *21*(5), 761–767. https://doi.org/10.1016/j.conb.2011.05.027

Sharpee, T. O., Rust, N. C., & Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Computation*, *16*(2), 223–50. https://doi.org/10.1162/089976604322742010

Shelton, J. A., Sheikh, A. S., Bornschein, J., Sterne, P., & Lücke, J. (2015). Nonlinear spike-And-Slab sparse coding for interpretable image encoding. *PLoS ONE*, *10*(5), 1–25. https://doi.org/10.1371/journal.pone.0124088

Slee, S. J., & David, S. V. (2015). Rapid Task-Related Plasticity of Spectrotemporal Receptive Fields in the Auditory Midbrain. *Journal of Neuroscience*, *35*(38), 13090–13102. https://doi.org/10.1523/JNEUROSCI.1671-15.2015

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network (Bristol, England)*, *12*(3), 289–316. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11563531

Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, *15*(6), 355–366. https://doi.org/10.1038/nrn3731

Theunissen, F. E., Sen, K., & Doupe, a J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *20*(6), 2315–2331. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10704507

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., & Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, *33*(4), 1104–16. https://doi.org/10.1016/j.neuroimage.2006.06.062

Thorson, I. L., Liénard, J., & David, S. V. (2015). The Essential Complexity of Auditory Receptive Fields. *PLoS Computational Biology*, *11*(12), 1–33. https://doi.org/10.1371/journal.pcbi.1004628

Touryan, J., Felsen, G., & Dan, Y. (2005). Spatial structure of complex cell receptive fields measured with natural images. *Neuron*, *45*(5), 781–91. https://doi.org/10.1016/j.neuron.2005.01.029

Varoquaux, G., Raamana, P., Engemann, D., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2016). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Retrieved from http://arxiv.org/abs/1606.05201

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., & Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, *110*, 48–59. https://doi.org/10.1016/j.neuroimage.2015.01.036

Willmore, B., & Smyth, D. (2003). Methods for first-order kernel estimation: simple-cell receptive fields from responses to natural scenes. *Network (Bristol, England)*, *14*(3), 553–77. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12938771

Woolley, S. M. N., Gill, P. R., Fremouw, T., & Theunissen, F. E. (2009). Functional groups in the avian auditory system. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(9), 2780–93. https://doi.org/10.1523/JNEUROSCI.2042-08.2009

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*, 477–505. https://doi.org/10.1146/annurev.neuro.29.051605.113024

Yamins, D. L. K., & DiCarlo, J. J. (2016). Eight open questions in the computational modeling of higher sensory cortex. *Current Opinion in Neurobiology*, *37*, 114–120. https://doi.org/10.1016/j.conb.2016.02.001

Yin, P., Fritz, J. B., & Shamma, S. A. (2014). Rapid Spectrotemporal Plasticity in Primary Auditory Cortex during Behavior. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(12), 4396–408. https://doi.org/10.1523/JNEUROSCI.2799-13.2014

Zhang, D., Gong, E., Wu, W., Lin, J., Zhou, W., & Hong, B. (2012). Spoken sentences decoding based on intracranial high gamma response using dynamic time warping. *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3292–3295. https://doi.org/10.1109/EMBC.2012.6346668

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., … Schroeder, C. E. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a “Cocktail Party.” *Neuron*, *77*(5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and Beyond. *Speech Communication*, *9*, 351–356. Retrieved from http://www.sciencedirect.com/science/article/pii/0167639390900107

Chapter 3

Aleman, A. (2004). The Functional Neuroanatomy of Metrical Stress Evaluation of Perceived and Imagined Spoken Words. *Cerebral Cortex*, *15*(2), 221–228. https://doi.org/10.1093/cercor/bhh124

Aziz-Zadeh, L., Cattaneo, L., Rochat, M., & Rizzolatti, G. (2005). Covert speech arrest induced by rTMS over both motor and nonmotor left hemisphere frontal sites. *Journal of Cognitive Neuroscience*, *17*(6), 928–938. https://doi.org/10.1162/0898929054021157

Basho, S., Palmer, E. D., Rubio, M. A., Wulfeck, B., & Müller, R.-A. (2007). Effects of generation mode in fMRI adaptations of semantic fluency: paced production and overt speech. *Neuropsychologia*, *45*(8), 1697–1706. https://doi.org/10.1016/j.neuropsychologia.2007.01.007

Bialek, W., Rieke, F., de Ruyter van Steveninck, R., & Warland, D. (1991). Reading a neural code. *Science*, *252*(5014), 1854–1857. https://doi.org/10.1126/science.2063199

Billingsley-Marshall, R., Clear, T., Mencl, W. E., Simos, P. G., Swank, P. R., Men, D., … Papanicolaou, A. C. (2007). A comparison of functional MRI and magnetoencephalography for receptive language mapping. *Journal of Neuroscience Methods*, *161*(2), 306–313. https://doi.org/10.1016/j.jneumeth.2006.10.020

Birk, D. (2013). cocor: Comparing correlations.

Boonstra, T. W., Houweling, S., & Muskulus, M. (2009). Does Asynchronous Neuronal Activity Average out on a Macroscopic Scale? *Journal of Neuroscience*, *29*(28), 8871–8874. https://doi.org/10.1523/JNEUROSCI.2020-09.2009

Brigham, K., & Kumar, B. V. K. V. (2010). Imagined Speech Classification with EEG Signals for Silent Communication: A Preliminary Investigation into Synthetic Telepathy (pp. 1–4). IEEE. https://doi.org/10.1109/ICBBE.2010.5515807

Carmena, J. M., Lebedev, M. A., Crist, R. E., O'Doherty, J. E., Santucci, D. M., Dimitrov, D. F., … Nicolelis, M. A. L. (2003). Learning to Control a Brain–Machine Interface for Reaching and Grasping by Primates. *PLoS Biology*, *1*(2), e2. https://doi.org/10.1371/journal.pbio.0000042

Chi, T., Gao, Y., Guyton, M. C., Ru, P., & Shamma, S. A. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *The Journal of the Acoustical Society of America*, *106*(5), 2719–2732.

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887. https://doi.org/10.1121/1.1945807

Creutzfeldt, O., Ojemann, J. G., & Lettich, E. (1989). Neuronal activity in the human lateral temporal lobe. II. Responses to the subjects own voice. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale*, *77*(3), 476–489.

Duffau, H., Capelle, L., Denvil, D., Gatignol, P., Sichez, N., Lopes, M., … Van Effenterre, R. (2003).

The role of dominant premotor cortex in language: a study using intraoperative functional mapping in awake patients. *NeuroImage*, *20*(4), 1903–1914.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*, *5*(3), e1000302. https://doi.org/10.1371/journal.pcbi.1000302

Feinberg, T. E., Gonzalez Rothi, L. J., & Heilman, K. M. (1986). "Inner speech" in conduction aphasia. *Archives of Neurology*, *43*(6), 591–593.

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, *10*(4), 507. https://doi.org/10.2307/2331838

Flinker, A., Chang, E. F., Kirsch, H. E., Barbaro, N. M., Crone, N. E., & Knight, R. T. (2010). Single-trial speech suppression of auditory cortex activity in humans. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(49), 16643–16650. https://doi.org/10.1523/JNEUROSCI.1809-10.2010

Geva, S., Bennett, S., Warburton, E., & Patterson, K. (2011). Discrepancy between inner and overt speech: Implications for post-stroke aphasia and normal language processing. *Aphasiology*, *25*(3), 323–343. https://doi.org/10.1080/02687038.2010.511236

Geva, Correia, & Warburton. (2011). Diffusion tensor imaging in the study of language and aphasia. *Aphasiology*, *25*(5), 543–558. https://doi.org/10.1080/02687038.2010.534803

Geva, Jones, Crinion, Baron, & Warburton, E. (2011). The neural correlates of inner speech defined by voxel-based lesion-symptom mapping. *Brain*, *134*(10), 3071–3082. https://doi.org/10.1093/brain/awr232

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, *31*(7). Retrieved from ftp://ftp.up.ac.za/pub/windows/CRAN/web/packages/dtw/vignettes/dtw.pdf

Guenther, F. H., Brumberg, J. S., Wright, E. J., Nieto-Castanon, A., Tourville, J. a, Panko, M., … Kennedy, P. R. (2009). A wireless brain-machine interface for real-time speech synthesis. *PloS One*, *4*(12), e8218. https://doi.org/10.1371/journal.pone.0008218

Guenther, F. H., Ghosh, S. S., & Tourville, J. a. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, *96*(3), 280–301. https://doi.org/10.1016/j.bandl.2005.06.001

Gunduz, A., Brunner, P., Daitch, A., Leuthardt, E. C., Ritaccio, A. L., Pesaran, B., & Schalk, G. (2012). Decoding covert spatial attention using electrocorticographic (ECoG) signals in humans. *NeuroImage*, *60*(4), 2285–2293. https://doi.org/10.1016/j.neuroimage.2012.02.017

Halpern. (1988). Mental scanning in auditory imagery for songs. *Journal of Experimental*

Psychology. Learning, Memory, and Cognition, *14*(3), 434–443.

Halpern, A. R. (1989). Memory for the absolute pitch of familiar songs. *Memory & Cognition*, *17*(5), 572–581.

Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, *42*(9), 1281–1292. https://doi.org/10.1016/j.neuropsychologia.2003.12.017

Heim, S., Opitz, B., & Friederici, A. D. (2002). Broca's area in the human brain is involved in the selection of grammatical gender for language production: evidence from event-related functional magnetic resonance imaging. *Neuroscience Letters*, *328*(2), 101–104.

Hickok. (2001). Functional anatomy of speech perception and speech production: psycholinguistic implications. *Journal of Psycholinguistic Research*, *30*(3), 225–235.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(May), 393–402. Retrieved from http://www.nature.com/nrn/journal/v8/n5/abs/nrn2113.html

Hinke, R. M., Hu, X., Stillman, A. E., Kim, S. G., Merkle, H., Salmi, R., & Ugurbil, K. (1993). Functional magnetic resonance imaging of Broca's area during internal speech. *Neuroreport*, *4*(6), 675–678.

Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J., … Donoghue, J. P. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, *485*(7398), 372–375. https://doi.org/10.1038/nature11076

Hochberg, L. R., Serruya, M. D., Friehs, G. M., Mukand, J. A., Saleh, M., Caplan, A. H., … Donoghue, J. P. (2006). Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, *442*(7099), 164–171. https://doi.org/10.1038/nature04970

Huang, J., Carr, T. H., & Cao, Y. (2002). Comparing cortical activations for silent and overt speech using event-related fMRI. *Human Brain Mapping*, *15*(1), 39–53.

Hubbard, T. L. (2013). Auditory Aspects of Auditory Imagery. In S. Lacey & R. Lawson (Eds.), *Multisensory Imagery* (pp. 51–76). New York, NY: Springer New York.

Intons-Peterson, M. J. (1980). The role of loudness in auditory imagery. *Memory & Cognition*, *8*(5), 385–393.

Jeannerod, M. (2003). Action Monitoring and Forward Control of Movements. In *in Arbib M., The handbook of brain theory and neural networks* (2nd ed, pp. 83–85). Cambridge, Mass: MIT Press.

Kaneoke, Y., Donishi, T., Iwatani, J., Ukai, S., Shinosaki, K., & Terada, M. (2012). Variance and Autocorrelation of the Spontaneous Slow Brain Activity. *PLoS ONE*, *7*(5), e38131. https://doi.org/10.1371/journal.pone.0038131

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–5. https://doi.org/10.1038/nature06713

Kennedy, J. F. (1961). “Inaugural Address.”

Kosslyn, S. M. (2005). Mental images and the Brain. *Cognitive Neuropsychology*, *22*(3), 333–347. https://doi.org/10.1080/02643290442000130

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews. Neuroscience*, *2*(9), 635–642. https://doi.org/10.1038/35090055

Kosslyn, S. M., & Thompson, W. L. (n.d.). Shared mechanisms in visual imagery and visual perception: Insights from cognitive neuroscience. *M. S. Gazzaniga (Ed.). The New Cognitive Neurosciences*.

Lachaux, J.-P., Axmacher, N., Mormann, F., Halgren, E., & Crone, N. E. (2012). High-frequency neural activity and human cognition: past, present and possible future of intracranial EEG research. *Progress in Neurobiology*, *98*(3), 279–301. https://doi.org/10.1016/j.pneurobio.2012.06.008

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., … Fox, P. T. (2000). Automated Talairach atlas labels for functional brain mapping. *Human Brain Mapping*, *10*(3), 120–131.

Leuthardt, E. C., Schalk, G., Wolpaw, J. R., Ojemann, J. G., & Moran, D. W. (2004). A brain–computer interface using electrocorticographic signals in humans. *Journal of Neural Engineering*, *1*(2), 63–71. https://doi.org/10.1088/1741-2560/1/2/001

McGuire, P. K., Silbersweig, D. A., Murray, R. M., David, A. S., Frackowiak, R. S., & Frith, C. D. (1996). Functional anatomy of inner speech and auditory verbal imagery. *Psychological Medicine*, *26*(1), 29–38.

Miller, K. J., Leuthardt, E. C., Schalk, G., Rao, R. P. N., Anderson, N. R., Moran, D. W., … Ojemann, J. G. (2007). Spectral changes in cortical surface potentials during motor movement. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(9), 2424–2432. https://doi.org/10.1523/JNEUROSCI.3886-06.2007

Mother Goose’s Nursery Rhymes. (1867). 1877. A Collection of Alphabets, Rhymes, Tales and Jingles.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–10. https://doi.org/10.1016/j.neuroimage.2010.07.073

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902–15. https://doi.org/10.1016/j.neuron.2009.09.006

Palmer, E. D., Rosen, H. J., Ojemann, J. G., Buckner, R. L., Kelley, W. M., & Petersen, S. E. (2001). An Event-Related fMRI Study of Overt and Covert Word Stem Completion. *NeuroImage*, *14*(1), 182–193. https://doi.org/10.1006/nimg.2001.0779

Partovi, S., Konrad, F., Karimi, S., Rengier, F., Lyo, J. K., Zipp, L., … Stippich, C. (2012). Effects of Covert and Overt Paradigms in Clinical Language fMRI. *Academic Radiology*, *19*(5), 518–525. https://doi.org/10.1016/j.acra.2011.12.017

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Nathan, E., … Chang, E. F. (2012). Reconstructing Speech from Human Auditory Cortex. *PLoS Biology*, *10*(1). https://doi.org/10.1371/journal.pbio.1001251

Pei, X., Leuthardt, E. C., Gaona, C. M., Brunner, P., Wolpaw, J. R., & Schalk, G. (2011). Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *NeuroImage*, *54*(4), 2960–72. https://doi.org/10.1016/j.neuroimage.2010.10.029

Pitt, M. A., & Crowder, R. G. (1992). The role of spectral and dynamic cues in imagery for musical timbre. *Journal of Experimental Psychology. Human Perception and Performance*, *18*(3), 728–738.

Price, C. J. (2011). A generative model of speech production in Broca's and Wernicke's areas. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00237

Price, C. J. (2012). A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847. https://doi.org/10.1016/j.neuroimage.2012.04.062

Rosen, H. J., Ojemann, J. G., Ollinger, J. M., & Petersen, S. E. (2000). Comparison of Brain Activation during Word Retrieval Done Silently and Aloud Using fMRI. *Brain and Cognition*, *42*(2), 201–217. https://doi.org/10.1006/brcg.1999.1100

Roth, M., Decety, J., Raybaudi, M., Massarelli, R., Delon-Martin, C., Segebarth, C., … Jeannerod, M. (1996). Possible involvement of primary motor cortex in mentally simulated movement: a functional magnetic resonance imaging study. *Neuroreport*, *7*(7), 1280–1284.

Roy, E., & Basler, P. (1955). Lincoln, A. 1863. "The Gettysburg Address." In "The Collected Works of Abraham Lincoln." New Brunswick, NJ: Rutgers UP.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49. https://doi.org/10.1109/TASSP.1978.1163055

Schalk, G. (2010). *A practical guide to brain-computer interfacing with BCI2000: general-purpose software for brain-computer interface research, data acquisition, stimulus presentation, and brain monitoring*. London ; New York: Springer.

Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., & Wolpaw, J. R. (2004). BCI2000: A General-Purpose Brain-Computer Interface (BCI) System. *IEEE Transactions on Biomedical Engineering*, *51*(6), 1034–1043. https://doi.org/10.1109/TBME.2004.827072

Shamma, S. A. (2003). Physiological foundations of temporal integration in the perception of speech. *Journal of Phonetics*, *31*(3–4), 495–501. https://doi.org/10.1016/j.wocn.2003.09.001

Shuster, L. I., & Lemieux, S. K. (2005). An fMRI investigation of covertly and overtly produced mono- and multisyllabic words. *Brain and Language*, *93*(1), 20–31. https://doi.org/10.1016/j.bandl.2004.07.007

Stevenson, R. J., & Case, T. I. (2005). Olfactory imagery: a review. *Psychonomic Bulletin & Review*, *12*(2), 244–264.

Tian, X., & Poeppel, D. (2012). Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Frontiers in Human Neuroscience*, *6*(November), 1–11. https://doi.org/10.3389/fnhum.2012.00314

Towle, V. L., Yoon, H.-A., Castelle, M., Edgar, J. C., Biassou, N. M., Frim, D. M., … Kohrman, M. H. (2008). ECoG gamma activity during a language task: differentiating expressive and receptive speech areas. *Brain*, *131*(8), 2013–2027. https://doi.org/10.1093/brain/awn147

Warland, D. K., Reinagel, P., & Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, *78*(5), 2336–2350.

Yetkin, F. Z., Hammeke, T. A., Swanson, S. J., Morris, G. L., Mueller, W. M., McAuliffe, T. L., & Haughton, V. M. (1995). A comparison of functional MR activation patterns during silent and audible language tasks. *AJNR. American Journal of Neuroradiology*, *16*(5), 1087–1092.

Chapter 4

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–8. https://doi.org/10.1016/j.tics.2012.05.003

Atencio, C. A., Sharpee, T. O., & Schreiner, C. E. (2012). Receptive field dimensionality increases from the auditory midbrain to cortex. *Journal of Neurophysiology*, *107*(10), 2594–2603. https://doi.org/10.1152/jn.01025.2011

Atiani, S., David, S. V., Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S. A., & Fritz, J. B. (2014). Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. *Neuron*, *82*(2), 486–99. https://doi.org/10.1016/j.neuron.2014.02.029

Block, N., & Siegel, S. (2013). Attention and perceptual adaptation. *The Behavioral and Brain Sciences*, *36*(3), 205–6. https://doi.org/10.1017/S0140525X12002245

Bornkessel-schlesewsky, I., & Schlesewsky, M. (2013). Brain & Language Reconciling time , space and function : A new dorsal – ventral stream model of sentence comprehension. *Brain and Language*, *125*(1), 60–76. https://doi.org/10.1016/j.bandl.2013.01.010

Bouchard, K. E., & Chang, E. F. (2014). Control of Spoken Vowel Acoustics and the Influence of Phonetic Context in Human Speech Sensorimotor Cortex. *Journal of Neuroscience*, *34*(38), 12662–12677. https://doi.org/10.1523/JNEUROSCI.1219-14.2014

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887. https://doi.org/10.1121/1.1945807

Coull, J. T., Frith, C. D., Büchel, C., & Nobre, a C. (2000). Orienting attention in time: behavioural and neuroanatomical distinction between exogenous and endogenous shifts. *Neuropsychologia*, *38*(6), 808–19. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10689056

Crone, N. E., Korzeniewska, A., & Franaszczuk, P. J. (2011). Cortical γ responses: searching high and low. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *79*(1), 9–15. https://doi.org/10.1016/j.ijpsycho.2010.10.013

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology. Human Perception and Performance*, *30*(4), 643–56. https://doi.org/10.1037/0096-1523.30.4.643

David, S. V., Fritz, J. B., & Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(6), 2144–9. https://doi.org/10.1073/pnas.1117717109

David, S. V., Mesgarani, N., & Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network (Bristol, England)*, *18*(3), 191–212. https://doi.org/10.1080/09548980701609235

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–47. https://doi.org/10.1016/j.heares.2007.01.014

Depireux, D. a, Simon, J. Z., Klein, D. J., & Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology*, *85*(3), 1220–1234.

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(8), E505-14. https://doi.org/10.1073/pnas.1113427109

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*(May), 311. https://doi.org/10.3389/fnhum.2014.00311

Eggermont, J. J. (2001). Between sound and perception: reviewing the search for a neural code. *Hearing Research*, *157*(1–2), 1–42. https://doi.org/10.1016/S0378-5955(01)00259-3

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*, *5*(3), e1000302. https://doi.org/10.1371/journal.pcbi.1000302

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752. https://doi.org/10.1037/a0017196

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., & Giraud, A.-L. (2014). The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nature Communications*, *5*(May), 4694. https://doi.org/10.1038/ncomms5694

Fritz, J. B., Shamma, S. A., Elhilali, M., & Klein, D. J. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, *6*(11), 1216–1223. https://doi.org/10.1038/nn1141

Gilbert, C. D., & Sigman, M. (2007). Brain states: top-down influences in sensory processing. *Neuron*, *54*(5), 677–96. https://doi.org/10.1016/j.neuron.2007.05.019

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Gramfort, A., Luessi, M., Larson, E., Engemann, D. a, Strohmeier, D., Brodbeck, C., … Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*(December), 267. https://doi.org/10.3389/fnins.2013.00267

Gross, J., Hoogenboom, N., Thut, G., Schyns, P. G., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLoS Biology*, *11*(12). https://doi.org/10.1371/journal.pbio.1001752

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York,

NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Horwitz, B., & Braun, A. R. (2004). Brain network interactions in auditory, visual and linguistic processing. *Brain and Language*, *89*(2), 377–84. https://doi.org/10.1016/S0093-934X(03)00349-3

Hsu, A., Borst, A., & Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation in Neural Systems*, *15*(2), 91–109. https://doi.org/10.1088/0954-898X/15/2/002

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *36*(6), 2014–26. https://doi.org/10.1523/JNEUROSCI.1779-15.2016

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron*, *77*(4), 750–61. https://doi.org/10.1016/j.neuron.2012.11.034

Leaver, A. M., & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(22), 7604–7612. https://doi.org/10.1523/JNEUROSCI.0296-10.2010

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–90. https://doi.org/10.1016/j.jneumeth.2007.03.024

Martin, S., Brunner, P., Holdgraf, C. R., Heinze, H.-J., Crone, N. E., Rieger, J. W., … Pasley, B. N. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering*, *7*(May), 14. https://doi.org/10.3389/fneng.2014.00014

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–6. https://doi.org/10.1038/nature11020

Mesgarani, N., Slaney, M., & Shamma, S. A. (2006). Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(3), 920–930. https://doi.org/10.1109/TSA.2005.858055

Miller, K. J., Zanos, S., Fetz, E. E., den Nijs, M., & Ojemann, J. G. (2009). Decoupling the cortical power spectrum reveals real-time representation of individual finger movements in humans. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(10), 3132–7. https://doi.org/10.1523/JNEUROSCI.5506-08.2009

Miller, L. M., Escabí, M. A., Read, H. L., & Schreiner, C. E. (2001). Functional convergence of response properties in the auditory thalamocortical system. *Neuron*, *32*(1), 151–160. https://doi.org/10.1016/S0896-6273(01)00445-7

Miller, L. M., Escabí, M. a, Read, H. L., & Schreiner, C. E. (2002). Spectrotemporal receptive

fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology*, *87*(1), 516–527.

Moore, R. C., Lee, T., & Theunissen, F. E. (2013). Noise-invariant Neurons in the Avian Auditory Cortex: Hearing the Song in Noise. *PLoS Computational Biology*, *9*(3). https://doi.org/10.1371/journal.pcbi.1002942

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Nathan, E., … Chang, E. F. (2012). Reconstructing Speech from Human Auditory Cortex. *PLoS Biology*, *10*(1). https://doi.org/10.1371/journal.pbio.1001251

Pedregosa, F., Grisel, O., Weiss, R., Passos, A., & Brucher, M. (2011). Scikit-learn : Machine Learning in Python, *12*, 2825–2830.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex (New York, N.Y. : 1991)*, *23*(June), 1378–87. https://doi.org/10.1093/cercor/bhs118

Peirce, J. W. (2008). Generating Stimuli for Neuroscience Using PsychoPy. *Frontiers in Neuroinformatics*, *2*(January), 10. https://doi.org/10.3389/neuro.11.010.2008

Rabinowitz, N. C., Willmore, B. D. B., King, A. J., & Schnupp, J. W. H. (2013). Constructing Noise-Invariant Representations of Sound in the Auditory Pathway. *PLoS Biology*, *11*(11). https://doi.org/10.1371/journal.pbio.1001710

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., & King, A. J. (2011). Contrast gain control in auditory cortex. *Neuron*, *70*(6), 1178–91. https://doi.org/10.1016/j.neuron.2011.04.030

Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J., & Hsiao, S. S. (2008). Neural correlates of high-gamma oscillations (60-200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *28*(45), 11526–36. https://doi.org/10.1523/JNEUROSCI.2848-08.2008

Schroeder, C. E., Wilson, D. a, Radman, T., Scharfman, H., & Lakatos, P. (2010). Dynamics of Active Sensing and perceptual selection. *Current Opinion in Neurobiology*, *20*(2), 172–6. https://doi.org/10.1016/j.conb.2010.02.010

Shamma, S. A., & Fritz, J. B. (2014). Adaptive auditory computations. *Current Opinion in Neurobiology*, *25C*, 164–168. https://doi.org/10.1016/j.conb.2014.01.011

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, *114*(6), 3394. https://doi.org/10.1121/1.1624067

Slepian, D. (1978). Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty-V: The Discrete Case. *Bell System Technical Journal*, *57*(5), 1371–1430. https://doi.org/10.1002/j.1538-7305.1978.tb02104.x

Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews*

Neuroscience, *15*(6), 355–366. https://doi.org/10.1038/nrn3731

Theunissen, F. E., Sen, K., & Doupe, a J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *20*(6), 2315–2331. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10704507

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, *13*, 22–30. https://doi.org/10.1109/MCSE.2011.37

Wassenhove, V. Van, & Schroeder, C. E. (2012). *The Human Auditory Cortex*. (D. Poeppel, T. Overath, A. N. Popper, & R. R. Fay, Eds.) (Vol. 43). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-2314-0

Wodlinger, B., Degenhart, A. D., Collinger, J. L., Tyler-Kabara, E. C., & Wang, W. (2011). The impact of electrode characteristics on electrocorticography (ECoG). *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, *2011*(1), 3083–6. https://doi.org/10.1109/IEMBS.2011.6090842

Woolley, S. M. N., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, *8*(10), 1371–9. https://doi.org/10.1038/nn1536

Woolley, S. M. N., Gill, P. R., Fremouw, T., & Theunissen, F. E. (2009). Functional groups in the avian auditory system. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(9), 2780–93. https://doi.org/10.1523/JNEUROSCI.2042-08.2009

Yin, P., Fritz, J. B., & Shamma, S. A. (2014). Rapid Spectrotemporal Plasticity in Primary Auditory Cortex during Behavior. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(12), 4396–408. https://doi.org/10.1523/JNEUROSCI.2799-13.2014

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46. https://doi.org/10.1016/S1364-6613(00)01816-7

Supplemental

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., … Knight, R. T. (2006). High Gamma Power Is Phase-Locked to Theta Oscillations in Human Neocortex. *Science*, *313*(5793), 1626–1628. https://doi.org/10.1126/science.1128115

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*(May), 311. https://doi.org/10.3389/fnhum.2014.00311

Ozkurt, T. E., & Schnitzler, A. (2011). A critical note on the definition of phase-amplitude cross-frequency coupling. *Journal of Neuroscience Methods*, *201*(2), 438–443. https://doi.org/10.1016/j.jneumeth.2011.08.014

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex (New York, N.Y. : 1991)*, *23*(June), 1378–87. https://doi.org/10.1093/cercor/bhs118

Voytek, B., Kayser, A. S., Badre, D., Fegen, D., Chang, E. F., Crone, N. E., … D'Esposito, M. (2015). Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. *Nature Neuroscience*, (July), 1–10. https://doi.org/10.1038/nn.4071

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., … Schroeder, C. E. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party." *Neuron*, *77*(5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037