Human Speech Processing and Its Clinical Applications

by

Connie Cheung

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

AND

UNIVERSITY OF CALIFORNIA, BERKELEY

[redacted]

Acknowledgments

This dissertation would not have been possible without the generous support of the UC Berkeley–SF Bioengineering community. I would especially like to thank the following people:

My advisor, *Eddie Chang*, for his mentorship, patience, and unwavering support.

My dissertation committee, *Flip Sabes, Nelson Morgan, John Houde,* and *Keith Johnson*, and my faculty advisors, *Christoph Schreiner* and *Sri Nagarajan*, for guiding my academic development.

Nima Mesgarani, for his kindness, encouragement, understanding, and mentorship.

Miranda Babiak, for always lending an ear –for both transcriptions and moral support.

The *Chang lab*, especially *Angela, David, Kris* and *Matt*, for providing a welcoming environment to discuss and improve upon my work.

My undergraduate mentors, *Cindee Madison, Lavi Secundo, and John Verboncoeur*, for encouraging and preparing me to pursue a graduate education.

The lesBEAST family, *Katie Fink, Jo Rys,* and *Paul Johnson*, for being a constant source of laughs, and adventure.

My family, especially *Mom, Dad, Karin,* and *Kellie*, for a lifetime of continuous support, motivation, and wisdom.

Parts of the text of this dissertation are reprints of the material as it appears in these publications:

1. Mesgarani, N., Cheung, C., Johnson, K., Chang, E. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science, 343*(6174), 1006-1010.

2. Cheung, C., & Chang, E. F. (2012). Real-time, time-frequency mapping of event-related cortical activation. *Journal of Neural Engineering, 9*(4), 046018.

Human Speech Processing and Its Clinical Applications

by

Connie Cheung

Doctor of Philosophy in Bioengineering

University of California, San Francisco and Berkeley

Professor Edward F. Chang, Chair

The human brain contains a remarkable sensory system that allows us to effortlessly process speech. The processing of speech begins at our ears, where speech sounds are converted into electrical signals that propagate up to our brain. How these signals are transformed from acoustic information into meaningful speech in the human cortex is still unknown. In this dissertation, high-density human direct cortical recordings were used to systematically detail the human speech processing system and address central linguistic theories of speech perception. In Chapter 2, we examine how the superior temporal gyrus, a brain area critically implicated in speech perception, encodes phonetic features of speech. We found single site response selectivity to distinct phonetic features and distributed population encoding mediated by acoustic properties, including pitch and voice-onset-time. In Chapter 3, we examine the representation of speech sounds in human motor cortex, a region controversially hypothesized to process articulatory gestures during perception. We found evidence that motor cortex does not represent articulatory representations of perceived actions in speech, but rather, auditory vocal information. These results are consistent with linguistic feature hierarchies organized around acoustic, rather than articulatory, features. Finally, built upon the principles of human speech electrophysiology, we developed a clinical cortical mapping tool to aid in the preservation of eloquent cortex, the details of which are described in Chapter 4. All together, this work lays a foundation for understanding the human speech processing system and developing clinical applications to aid the lives of thousands with neurological disorders, including epilepsy and brain tumors.

Contents

List of tables

List of figures

CHAPTER 1

Introduction

Whether we are conversing with a friend or listening to a podcast, we are constantly using our neural systems to accurately process speech sounds in order to navigate the physical world. Speech sounds, which are complex spectrotemporal patterns of air pressure created by human vocal tract configurations and rapid movements of articulators (Ladefoged et al., 2010), enter the ears and are converted into an electrical signal that propagates all the way up to the human cortex (Kandel et al., 2013). But, how exactly these sounds are represented in the human cortex is still unknown.

This dissertation seeks to understand the representation of speech sounds in human cortex by systematically detailing the human speech processing system. Additionally, this knowledge will be applied to the design of an innovative bioengineering tool. Insight into human speech processing will shed light on a fundamental process of the human neural system, and provide a physiologic framework to shape linguistic theories of speech perception. Its clinical application will also significantly impact the lives of thousands who undergo neurosurgical cortical resection procedures every year.

1.1 Human speech processing

How do humans convert acoustic waveforms to meaningful speech sounds? This fundamental question has been a topic of intense academic research for decades. Yet despite extensive behavioral, lesion, and neuroimaging studies, understanding the neurological basis of speech processing has largely eluded investigators.

The first hypothesis was proposed by Carl Wernicke, a German physician who specialized in neurological speech and language deficits, in the 1870s. Wernicke proposed that speech was supported by peri-Sylvian cortex (Fig. 1.1, reprinted from Wernicke, 1874). In this classic Wernicke-Lichtheim model, area *a* in the left posterior temporal lobe is a sensory area for speech sound that is connected by the arcuate fasciculus fiber bundle to area *b* in the left inferior frontal region for speech production (Wernicke, 1874).

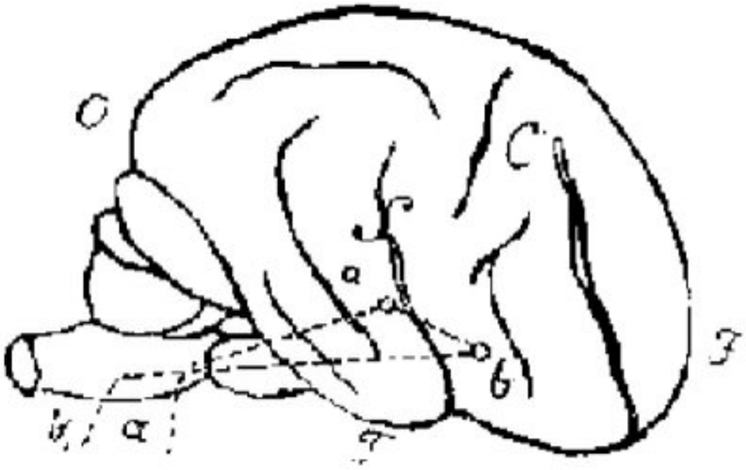


Fig 1.1. Wernicke's proposed speech system, where F denotes the frontal, O the occipital, and T the temporal lobe. C is the central sulcus, and S is the Sylvian fissure. From Wernicke, 1874.

Importantly, Wernicke claimed area *a*, formally called the left superior temporal gyrus (STG), was critical in speech perception, and cited patients with lesions in left STG presenting with auditory language comprehension disorders (Wernicke's aphasia) as evidence of this.

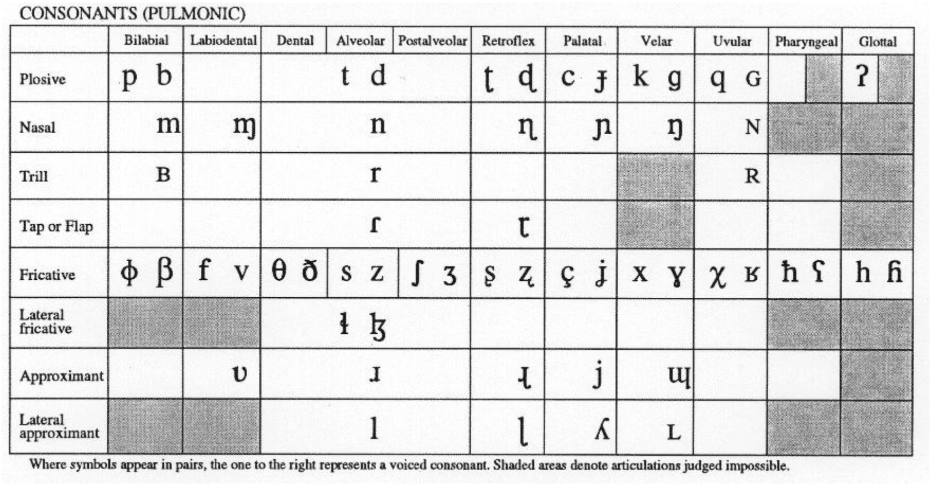
Though the Wernicke-Lichtheim model continues to have to strong influence in neuroscience and clinical teachings today, several discoveries in recent literature has since challenged the notion that left STG solely supported speech perception. First, it was observed that the inability to perceive speech minimally affected auditory comprehension deficits in

Wernicke's aphasia (Basso et al., 1977; Blumstein et al., 1977a; Blumstein et al., 1977b; Hickok et al., 2007; Miceli et al., 1980). Secondly, it was noted that left STG lesions caused deficits in speech production (Damasio et al., 1980). Thirdly, disruption of motor cortex via stimulation has been shown to bias identification of speech sounds (Fadiga et al., 2002; Meister et al., 2007). These findings bring into question how exactly left STG represents speech, and how motor cortex might contribute in the same process (Hickok et al., 2007). These issues are directly addressed in the following two chapters.

1.2 Linguistic theories of speech perception

Decades of linguistic research have culminated in a complete description of the representations of speech sounds, and provide clues to the speech representations we might expect to find at the cortical level.

At the most fundamental level, linguists have described speech to be made of distinctive features that are present in all oral languages (Chomsky et al., l968; Ladefoged et al., 2010). Distinctive features describe the speech sound's major class (e.g. syllabic, consonantal), glottal state (e.g. voicing), manner of articulation (e.g. plosive, fricative, nasal), and place of articulation (e.g. labial, dorsal, coronal, high, low, back) (Ladefoged et al., 2010). All together, these features provide a comprehensive set of basic features that describe the components of speech sound. But, no single distinctive feature is enough to create meaningful speech. To form speech, a combination of distinctive features is needed to create a phoneme (denoted by International Phonetic Alphabet (IPA) symbols, Fig 1.2), the smallest contrastive unit that changes a word's meaning (e.g. /b/ and /d/ as in bad versus dad) (Chomsky et al., l968; Ladefoged et al., 2010). Phonemes are further bundled into groups of syllables to create a word (Ladefoged et al., 2010).



CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Fig 1.2. IPA consonants chart organized by manner of articulation (rows) and place of articulation (columns). When symbols appear in pair, the right symbol is a voiced consonant. Shaded areas denote articulations judged impossible.

Though linguists have described these varying levels of speech, it remains unknown how the human cortex extracts the linguist elements (whether they be distinctive features, phonemes, or some higher level unit) from an acoustic signal to represent speech. Linguists have put forth two major theories of speech perception to address this issue.

One major theory, called the motor theory of speech perception, was first introduced by Liberman and colleagues. This motor-centric view of speech perception was a dominant view among linguists in the 1970s, and has found more recent proponents since the discovery of mirror neurons – neurons that fire both to the action and observation of a task – in motor cortex. The motor theory was inspired by Liberman et al.'s observation that speech sounds spoken in different contexts result in ambiguous mappings between spectrotemporal acoustics and perceived phonemes (Delattre et al., 1962; Delattre et al., 2005; Liberman et al., 1954). To overcome this acoustic variability, they reasoned that articulatory representations could provide disambiguating robust information, and proposed "the objects of speech perception are the intended phonetic gestures of the speaker" (Liberman et al., 1967; Liberman et al., 1985). The motor theory hypothesizes that the neural system is specialized to represent speech with the motor commands required to produce the sound (Liberman et al., 1967; Liberman et al., 1985, 1989).

Another major approach presents an alternative to the motor theory, and proposes a more acoustic-centric view of speech perception. This "general approach" was propelled by the findings that perception of speech stimuli could be replicated with non-speech stimuli (Pisoni, 1977; Stevens et al., 1974), contrary to the predictions of motor theory. Researchers hypothesized that speech sounds were perceived with the same auditory mechanisms for non-speech sounds by maintaining a running estimate of the acoustic context, instead of representing motor articulators (Diehl et al., 1989; Diehl et al., 2004; Holt et al., 2008; Massaro et al., 2008; Stevens, 2002). This approach therefore suggests that auditory cortex extract the necessary acoustic elements of speech, and is not dependent on motor cortex for speech perception.

In summary, linguistic research of speech perception has thus far led to the characterization of multiple levels of representation of speech, and complementary theoretical frameworks to guide the study of speech processing in the human neural system. In Chapters 2 and 3, this linguistic framework will be used to examine the representation of speech in both human STG and motor cortex.

1.2 Clinical mapping of eloquent cortex

Understanding the neurological basis for speech processing has significant implications for clinical practice. One such clinical application, which will be the topic of Chapter 4, is a neurosurgical tool for the preservation of eloquent cortex during resection of malignant cortical tissue.

The ability to identify eloquent regions of the brain through functional mapping is invaluable to neurosurgeons facing the challenge of removing abnormal cortical tissue without rendering a patient incapacitated. The current standard of care is called electrocortical stimulation (ECS), where electrical pulses are directly applied to the patient's brain to elicit or stop activity, such as speech(Haglund et al., 1994; Keles et al., 2004; Penfield et al., 1954; Penfield et al., 1959). As surgeons map the exposed cortex, regions critical for speech and movement are identified. These regions are avoided when resecting malignant cortical tissue (Fig. 1.3).

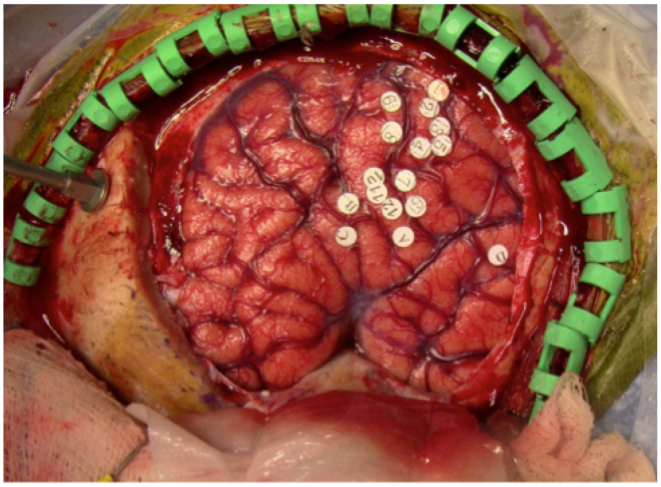


Fig 1.3. Patients undergo an awake craniotomy for intraoperative mapping of eloquent cortex. Cortex is electrically stimulated and marked for preoperative surgical planning.

Although considered the gold-standard, ECS is inefficient and dangerous (Lesser et al., 1984; Luders et al., 1991; Nii et al., 1996; Ojemann et al., 1989). Currently, a safe and efficient alternative to ECS does not exist in practice. Thus with knowledge of the physiologic representation of speech and other eloquent functions, a clinical tool can be built to allow for rapid, safe cortical mapping. The accuracy and benefits of this clinical tool is demonstrated in Chapter 4.

1.3 Electrocorticography

A major challenge in addressing the neural representation of speech is that cortical processing of speech sounds is exceptionally spatially discrete and temporally fast (Chang et al., 2010a; Formisano et al., 2008; Obleser et al., 2010; Steinschneider et al., 2011). To address this challenge, this dissertation employs the use of direct cortical recordings from patients implanted with chronic subdural electrode arrays undergoing intracranial monitoring for the localization of epileptic seizure foci.

This clinical practice of subdural grid implantation presents a unique opportunity to study an electrophysiologic signal called electrocorticography (ECoG). ECoG provides high spatial resolution (4mm pitch) (Bouchard et al., 2013; Chang et al., 2010a; Mesgarani et al., 2012; Mesgarani et al., 2014) with the absence of noise artifacts caused by the skull and cranial muscle often seen in other forms of cortical monitoring, such as scalp electroencephalography (EEG) (Pfurtscheller et al., 1975). Recent studies have revealed that high gamma frequencies (70-150Hz) in ECoG are an extremely discrete spatial and temporal marker of neural activity (Crone et al., 1998a; Edwards et al., 2009). Furthermore, high-gamma activity has been shown to be a strong correlate of multi-unit firing rate (Ray et al., 2011; Steinschneider et al., 2008). These properties make high-gamma ECoG signal an ideal candidate to study human speech processing, and will be the focus of this dissertation.

1.4 Chapter previews

Studies on human speech processing have resulted in questions regarding the representation of speech in human STG and motor cortex. In Chapter 2 and 3, this representation is described at an unprecedented spatiotemporal resolution, and presents a physiologic approach to validating linguistic theories of speech perceptions. In Chapter 4, a clinical cortical mapping tool designed to aid in the preservation of eloquent cortex is described. Together, this work details the human speech processing system and the benefits of physiologic design for the creation of clinical tools.

CHAPTER 2

Phonetic feature encoding in human superior temporal gyrus

During speech perception, linguistic elements such as consonants and vowels are extracted from a complex acoustic speech signal. Superior temporal gyrus (STG) participates in high-order auditory processing of speech, but how it encodes phonetic information is poorly understood. In this chapter, we used high-density direct cortical surface recordings in humans while they listened to speech to reveal the STG representation of the entire English phonetic inventory. At single electrodes, we found response selectivity to distinct phonetic features. Encoding of acoustic properties was mediated by a distributed population response. Phonetic features were directly related to tuning for spectrotemporal acoustic cues, some of which were encoded in a non-linear fashion or by integration of multiple cues. These findings demonstrate the acoustic-phonetic representation of speech in human STG.

This work was done in collaboration with Nima Mesgarani, Keith Johnson, and Edward F. Chang, and was published in Science (Mesgarani et al., 2014). This author contributed largely to Figures 2.1, 2.2, 2.3B, E, F, 2.4A, D, 2.6, S2.1, S2.2, S2.4-6, S2.10, S2.12, along with the corresponding analyses, data collection, and text.

2.1 Introduction

Phonemes—and the distinctive features composing them— are hypothesized to be the smallest contrastive units that change a word's meaning (e.g., /b/ and /d/ as in bad vs. dad) (Chomsky et al., l968). Superior temporal gyrus (Brodmann area 22, STG) has a key role in acoustic-phonetic processing as it responds to speech over other sounds (Binder et al., 2000). Additionally, with focal electrical stimulation, subjects experience selective interruption with speech discrimination (Boatman et al., 1995). These findings raise fundamental questions about the representation of speech sounds, such as whether local neural encoding is specific for phonemes, acoustic-phonetic features, or low-level spectrotemporal parameters. A major challenge in addressing this in speech is that cortical processing of individual speech sounds is extraordinarily spatially discrete and rapid (Chang et al., 2010a; Formisano et al., 2008; Obleser et al., 2010; Steinschneider et al., 2011).

We recorded direct cortical activity from seven human subjects implanted with high-density multi-electrode arrays as part of their clinical evaluation for epilepsy surgery. These recordings provide simultaneous high spatial and temporal resolution while sampling population neural activity from temporal lobe auditory speech cortex. We primarily analyzed high gamma (75-150Hz) cortical surface field potentials (Crone et al., 1998a), which correlate with neuronal spiking (Ray et al., 2011; Steinschneider et al., 2008). Other frequency bands including theta (4-7Hz), alpha (7-14Hz), beta (15-30Hz), and gamma (30-70Hz) were also examined.

2.2 Results

2.2.1 Neural activity when listening to syllables

Subjects listened to a randomly ordered set of 48 different consonant-vowel (CV) syllables, spoken by 6 different speakers (3 females and 3 males). A broad and representative sample of syllables commonly spoken in American English was selected to be played, and included 16 different consonants (/b d g k l m n p r s ʃ t v w y z/) and 3 different vowels (/a i u/). Neural responses (high-gamma activity, z-scored to a baseline rest period, further described in Chapter 2.4.3) in posterior and middle STG demonstrated a distributed spatiotemporal pattern of evoked activity when listening to syllables (Fig. 2.1A, comparing speech vs. silence, $p<0.01$, *t*-test). It should be noted that activity was also observed in supra-Sylvian sites, and was further examined in Chapter 3.

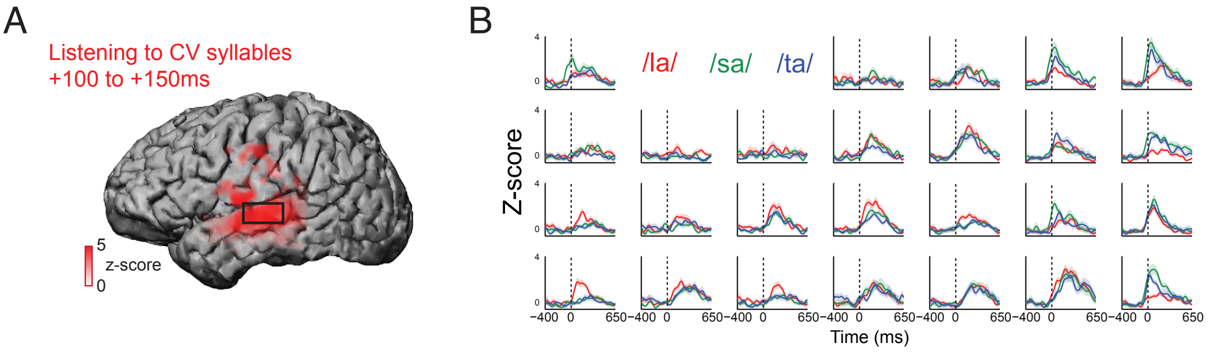


Fig 2.1. Human neural activity (A) Average activity when subjects are listening to CV syllables are plotted on a magnetic resonance image surface reconstruction of one subject's cerebrum. Opacity signifies the high-gamma z-score.

We segmented the responses to each syllable, and examined average STG responses to different speech sounds. Sites demonstrated significant high-gamma response selectivity to specific syllables (Fig. 2.1B, located in the black rectangle in Fig. 2.1A). For example, some electrodes showed a unique high response to approximants (/la/, red) over non-approximants sounds. This suggested the possibility of a strong phonetic representation in STG.

2.2.2 Superior temporal gyrus activity when listening to sentences

To further examine these responses in a more natural setting, subjects listened to continuous speech samples featuring a wide range of American English speakers (TIMIT, 500 sentences spoken by 400 people) (Garofolo, 1993). Again, most speech responsive sites were found in posterior and middle STG (Fig. 2.2A, 37-102 sites per subject, comparing speech vs. silence, *p*<0.01, *t*-test). Neural responses demonstrated a distributed spatiotemporal pattern evoked during listening (Fig. 2.2B-C, Fig. S2.1, S2.2).

We segmented the sentences into time-aligned sequences of phonemes to investigate whether STG sites show preferential responses. We estimated the mean neural response at each electrode to every phoneme, and found clear selectivity. For example, electrode e1 (Fig. 2.2D) showed large evoked responses to plosive phonemes: /p t k b d g/). Electrode e2 showed selective responses to sibilant fricatives: /s ʃ z/. The next two electrodes showed selective responses to subsets of vowels: low-back (electrode e3: e.g. /a aʊ/), high-front vowels and glides (electrode e4: e.g. /i/, /j/). Finally, neural activity recorded at electrode e5 was selective for nasals (/n m ŋ/).

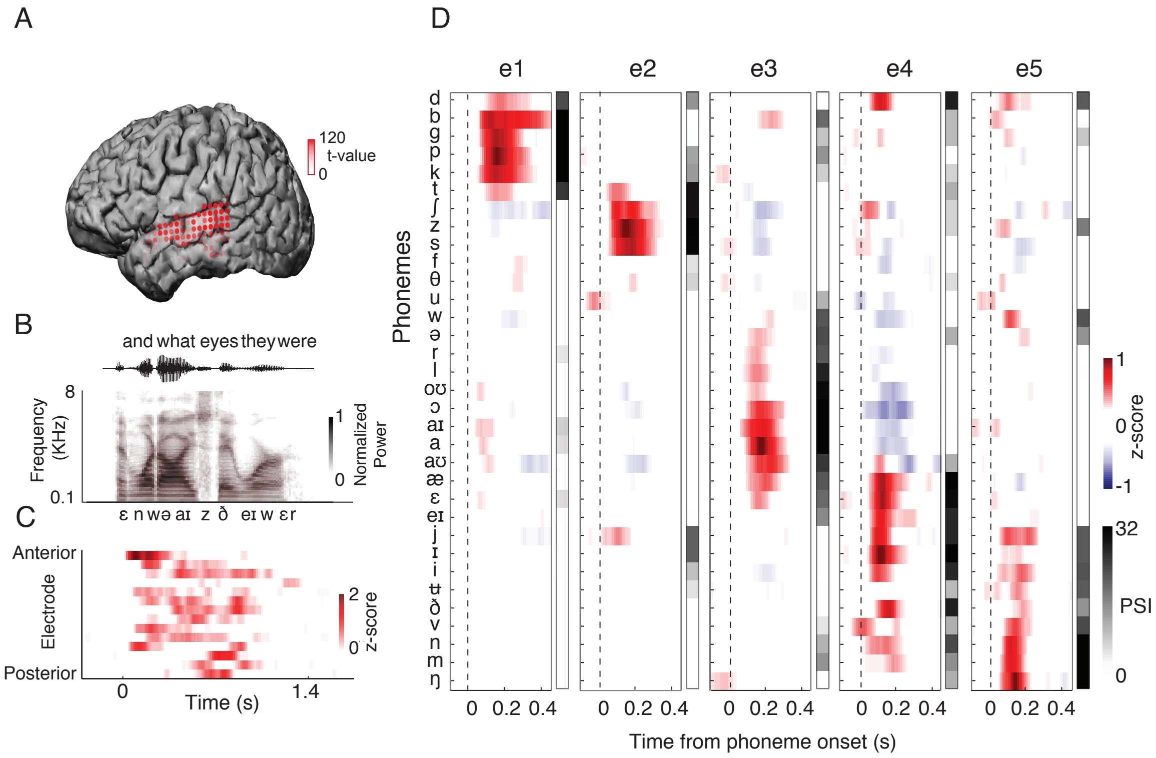


Fig. 2.2. Human STG cortical selectivity to speech sounds (A) Magnetic resonance image surface reconstruction of

one subject's cerebrum. Electrodes (red) are plotted with opacity signifying the t-test value when comparing responses to silence and speech (p<0.01, t-test). (B) Example sentence and its acoustic waveform, spectrogram and phonetic transcription. (C) Neural responses evoked by the sentence at selected electrodes. (D) Average responses at five example electrodes to all English phonemes, and their PSI vectors.

2.2.3 Phoneme selectivity

To quantify selectivity at single electrodes, we derived a metric indicating the number of phonemes with cortical responses statistically distinguishable from the response to a particular phoneme (phoneme selectivity index (PSI), dimension of 33 English phonemes; PSI=0 is nonselective, PSI=32 is extremely selective; Wilcox rank-sum test, *p*<0.01, Fig. 2.2D, methods shown in Fig. S2.3). We determined an optimal analysis time window of 50ms, centered 150ms after the phoneme onset using a phoneme separability analysis (f-statistic, Fig. S2.4A). The average PSI over all phonemes summarizes an electrode's overall selectivity. The average PSI was highly correlated to a site's response magnitude to speech over silence (r = 0.77, *p*<0.001, *t*-test, Fig. S2.5A), and the degree to which the response could be predicted with a linear spectrotemporal receptive field (STRF, *r* = 0.88, *p*<0.001, *t*-test, Fig. S2.5B (Theunissen et al., 2001)). Therefore, the majority of speech responsive sites in STG are selective to specific phoneme groups.

To investigate the organization of selectivity across the neural population, we constructed an array containing PSI vectors for electrodes across all subjects (Fig. 2.3A). In this array, each column corresponds to a single electrode, and each row corresponds to a single phoneme. Most STG electrodes are selective not to individual, but to specific groups of phonemes. To determine selectivity patterns across electrodes and phonemes, we used unsupervised hierarchical clustering analyses. Clustering across rows revealed groupings of phonemes based on the pattern of similarity of PSI values in the population response (Fig. 2.3B). Clustering across columns revealed single electrodes with similar PSI patterns (Fig. 2.3C). These two analyses revealed complementary local- and global- level organizational selectivity patterns. We also re-plotted the array using 14 phonetic features, defined in linguistics to contrast distinctive articulatory and acoustic properties (Fig. 2.3D, phoneme-feature mapping provided in Fig. S2.7) (Chomsky et al., l968; Sundberg et al., 1991).

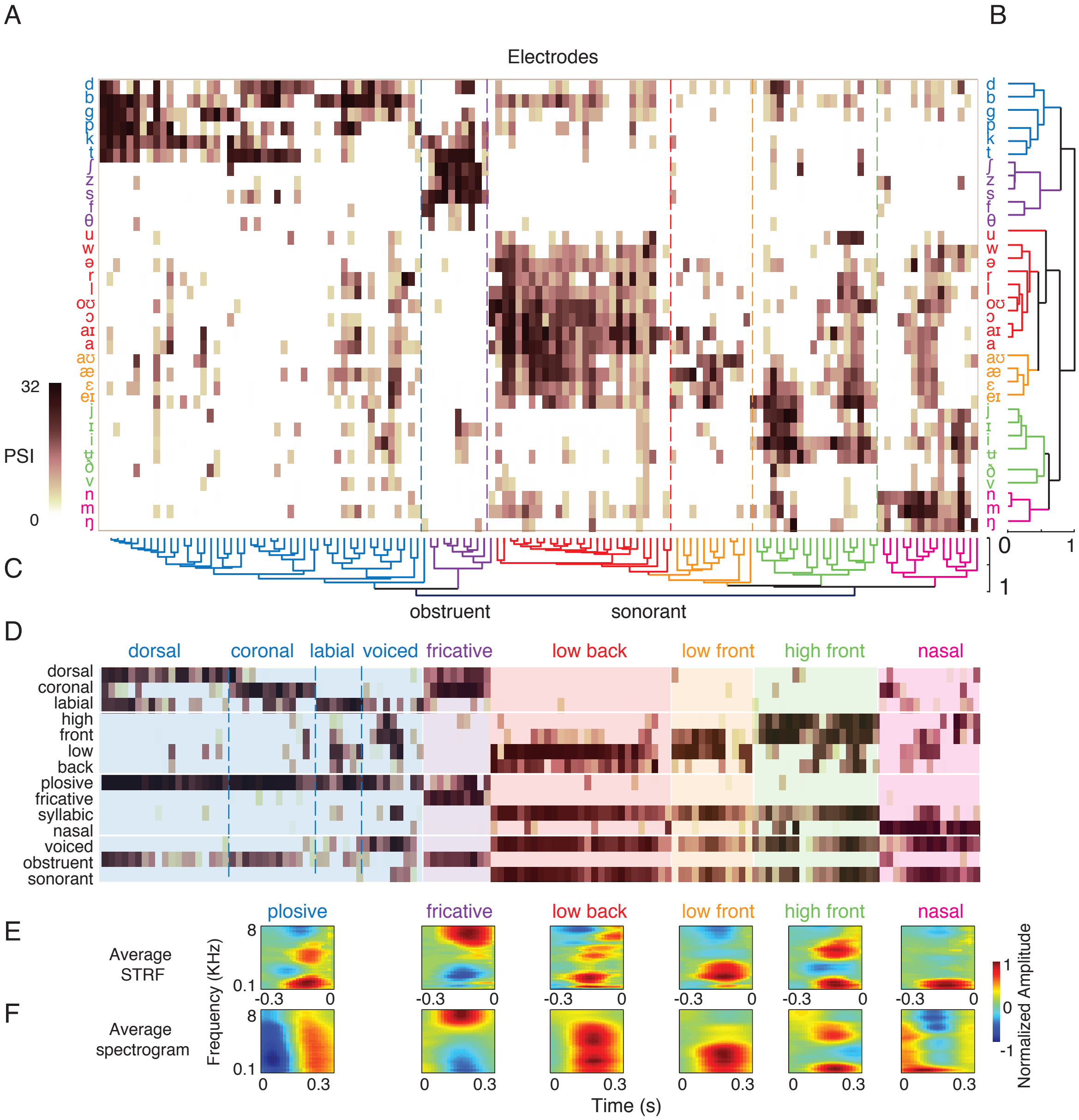


Fig. 2.3. Hierarchical clustering of single electrode and population responses (A) PSI vectors of selective electrodes across all subjects. Rows correspond to phonemes, and columns correspond to electrodes. (B) Clustering across population PSIs (rows). (C) Clustering across single electrodes (columns). (D) Alternative PSI vectors using rows now corresponding to phonetic features, not phonemes. (E) Weighted average STRFs of main electrode clusters. (F) Average acoustic spectrograms for phonemes in each population cluster. Correlation between average STRFs and average spectrograms: r = 0.67, p<0.01, t-test. (r = 0.50, 0.78, 0.55, 0.86, 0.86, 0.47 for plosives, fricatives, vowels and nasals respectively, p<0.01, t-test).

The first tier of the single electrode hierarchy analysis (Fig. 2.3C) divides STG sites into two distinct groups: obstruent- and sonorant- selective electrodes. The obstruent-selective group is divided into two subgroups: plosive and fricative electrodes (similar to electrodes e1-2 in Fig. 2.2D) (Ladefoged et al., 2010). Among plosive electrodes (blue), some were responsive to all plosives, whereas others were selective to place of articulation (dorsal /g k/ vs. coronal /d t/ vs. labial /p b/, labeled in Fig. 2.3D) and voicing (separating voiced /b d g/ from unvoiced /p t k/, labeled voiced in Fig. 2.3D). Fricative-selective electrodes (purple) showed weak, overlapping selectivity to coronal plosives (/d t/). Sonorant-selective cortical sites, in contrast, were partitioned into four partially overlapping groups: low-back vowels (red), low-front vowels (orange), high-front vowels (green) and nasals (magenta)(labeled in Fig. 2.3D, similar to e3-5 in Fig. 2.2D).

Both clustering schemes (Fig. 2.3B-C) revealed similar phoneme grouping based on shared phonetic features, suggesting a significant portion of the population-based organization can be accounted for by local tuning to features at single electrodes (similarity of average PSI values for the local and population subgroups of both clustering analyses is shown in Fig. S2.8, overall $r = 0.73$, $p<0.001$). Furthermore, selectivity is primarily organized by manner of articulation distinctions, and secondarily by place of articulation, corresponding to the degree and location of constriction in the vocal tract, respectively (Ladefoged et al., 2010). This systematic organization of speech sounds is consistent with auditory perceptual models positing that distinctions are most affected by manner contrasts (Clements, 1985; Stevens, 2002), compared to other feature hierarchies (articulatory/gestural theories) (Fowler, 1986).

We next determined what spectrotemporal tuning properties accounted for phonetic feature selectivity. We first determined the weighted average STRFs of the six main electrode clusters identified above, weighting them proportionally by their degree of selectivity (average PSI). These STRFs show well-defined spectrotemporal tuning (Fig. 2.3E) highly similar to average acoustic spectrograms of phonemes in corresponding population clusters (Fig. 2.3F, average correlation = 0.67, p<0.01, t-test). For example, the first STRF in Fig. 2.3E shows tuning for broadband excitation followed by inhibition, similar to the acoustic spectrogram of plosives. The second STRF is tuned to high frequency, which is a defining feature of sibilant fricatives. STRFs of vowel electrodes show tuning for characteristic formants that define low-back, low-front, and high-front vowels, respectively. Finally, STRF of nasal-selective electrodes is tuned primarily to low acoustic frequencies generated from heavy voicing and damping of higher frequencies (Ladefoged et al., 2010). The average spectrogram analysis requires *a priori* phonemic segmentation of speech, but is model-independent. The STRF analysis assumes a linear relationship between spectrograms and neural responses, but is estimated without segmentation. Despite these differing assumptions, the strong match between these confirms that phonetic feature selectivity results from tuning to signature spectrotemporal cues.

2.2.4 Vowels

We have thus far focused on local feature selectivity to discrete phonetic feature categories. However, we next wanted to address the encoding of continuous acoustic parameters that specify phonemes within vowel, plosive, and fricative groups. For vowels, we measured fundamental (F0) and formant (F1-F4) frequencies (Ladefoged et al., 2010). The first two formants (F1-F2) play a major perceptual role in distinguishing different English vowels (Ladefoged et al., 2010), despite tremendous variability within and across vowels (Fig. 2.4A) (Peterson et al., 2005). The optimal projection of vowels in formant space was the difference of F2 and F1 (first principal component, dashed line, Fig. 2.4A), which is consistent with vowel perceptual studies (Miller, 1989; Syrdal et al., 1986). Using partial correlation analysis, we quantified the relationship between electrode response amplitudes and F0-F4. On average, we observed no correlation between the sensitivity of an electrode to F0 with its sensitivity to F1 or F2. However, sensitivity to F1 and F2 was negatively correlated across all vowel-selective sites (Fig. 2.4B, r = -0.49, *p*<0.01, *t*-test), meaning single STG sites show an integrated response to both F1 and F2. Furthermore, electrodes selective to low-back and high-front vowels (labeled in Fig. 2.3D) showed an opposite differential tuning to formants, thereby maximizing vowel discriminability in the neural domain. This complex sound encoding matches the optimal projection in Fig. 2.4A, suggesting a specialized higher-order encoding of acoustic formant parameters (Nelken, 2008; Sussman, 1986) and contrasts with studies of speech sounds in nonhuman species (Engineer et al., 2008; Mesgarani et al., 2008).

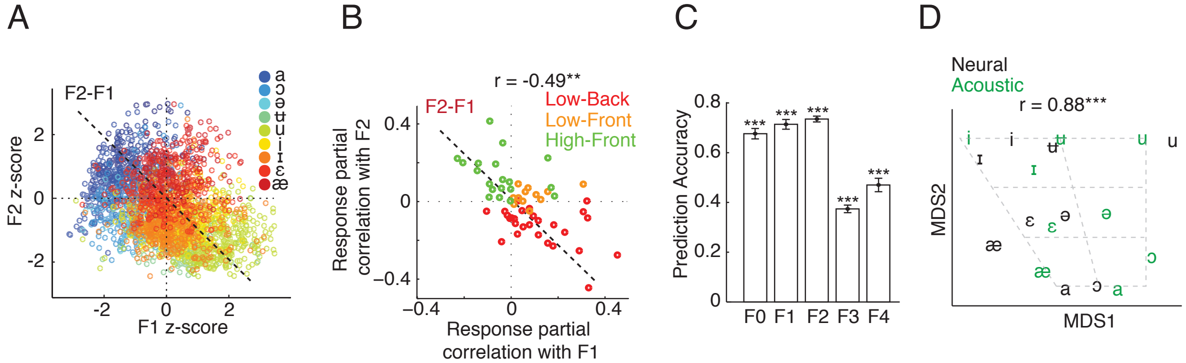


Fig. 2.4. Neural encoding of vowels (A) Formant frequencies, F1 and F2, for English vowels (F2-F1, dashed line, first principal component). (B) F1 and F2 partial correlations for each electrode's response (**p<0.01, t-test). Dots (electrodes) color-coded by their cluster membership. (C) Neural population decoding of fundamental and formant frequencies. (D) Multidimensional scaling of acoustic and neural space (***p<0.001, t-test).

To examine population representation of vowel parameters, we used linear regression to decode F0-F4 from neural responses. To ensure unbiased estimation, we first removed correlations between F0-F4 using linear prediction, and decoded the residuals. Relatively high decoding accuracies are shown in Fig. 2.4C (*p*<0.001, *t*-test), suggesting fundamental and formant variability is well represented in population STG responses (interaction between decoder weights with electrode STRFs shown in Fig. S2.9). Using multidimensional scaling, we found that the relational organization between vowel centroids in the acoustic domain is well preserved in neural space (Fig. 2.4D, $r = 0.88$, *p*<0.001).

2.2.5 Plosives and fricatives

For plosives, we measured three perceptually important acoustic cues (Fig. S2.10): Voice-Onset-Time (VOT) which distinguishes voiced (/b d g/) from unvoiced plosives (/p t k/), spectral peak (differentiating labials /p b/ vs. coronal /t d/ vs. dorsal /k g/), and second formant (F2) of the following vowel (Ladefoged et al., 2010). These acoustic parameters could be decoded from population STG responses (Fig. 2.5A, *p*<0.001, *t*-test). VOT in particular is a temporal cue that is perceived categorically, which suggests a non-linear encoding (Lisker et al., 1967). Figure 2.5B shows neural responses for three example electrodes plotted for all plosive instances (total of 1200), aligned to their release time and sorted by VOT. The first electrode responds to all plosives with same approximate latency and amplitude, irrespective of VOT. The second electrode responds only to plosive phonemes with short VOT (voiced), and the third electrode responds primarily to plosives with long VOT (unvoiced).

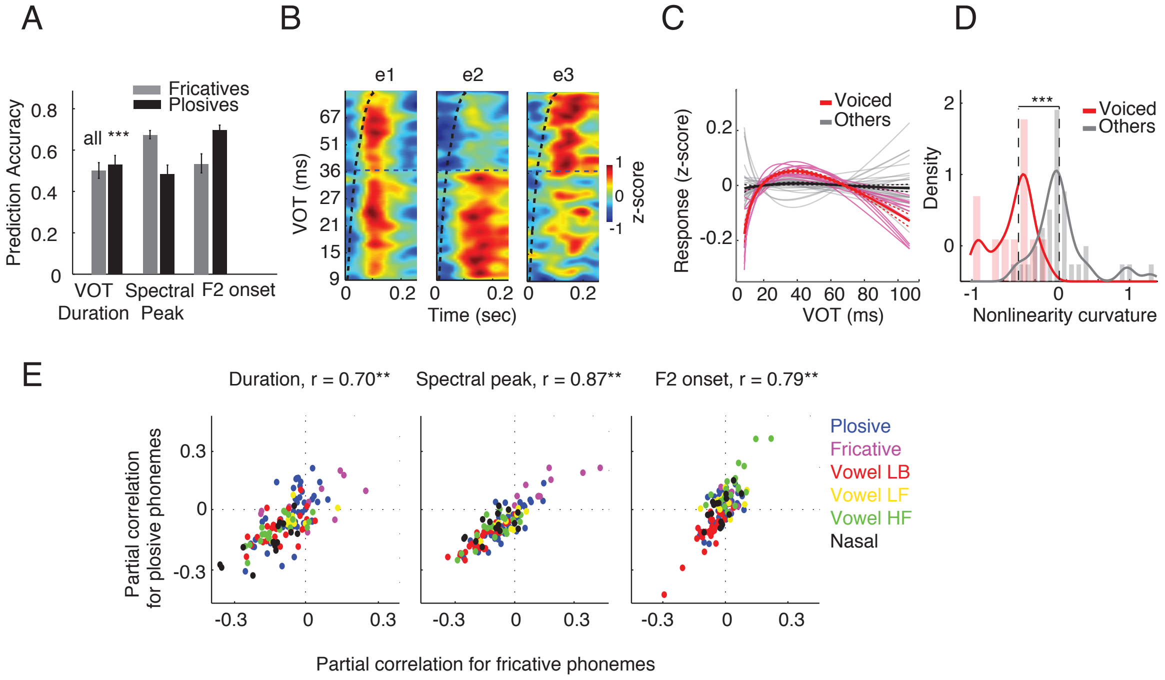


Fig. 2.5. Neural encoding of plosive and fricative phonemes (A) Prediction accuracy of plosive and fricative acoustic parameters from neural population responses. (B) Response of three example electrodes to all plosive phonemes sorted by VOT. (C) Nonlinearity of VOT-response transformation, and (D) distributions of nonlinearity for all plosive-selective electrodes identified in Figure 2D. Voiced plosive-selective electrodes are shown in pink, and the rest in gray. (E) Partial correlation values between response of electrodes and acoustic parameters shared between plosives and fricatives (**p<0.01, t-test). Dots (electrodes) color-coded by their cluster grouping from Fig. 2C.

To examine the nonlinear relationship between VOT and response amplitude for voiced-plosive electrodes (labeled voiced in Fig. 2.3D) compared to plosive electrodes with no sensitivity to voicing feature (labeled coronal, labial and dorsal in Fig. 2.3D), we fitted a linear and exponential function to VOT-response pairs (Fig. S2.11B). The difference between these two fits specifies the nonlinearity of this transformation, shown for all plosive electrodes in Fig. 2.5C. Voiced-plosive electrodes (pink) all show strong nonlinear bias for short VOTs, compared to all other plosive electrodes (gray). We quantified the degree and direction of this nonlinear bias for these two groups of plosive electrodes by measuring the average second-derivative of the curves in Fig. 2.5C. This measure maps electrodes with nonlinear preference for short VOTs (e.g. electrode e2 in Fig. 2.5B) to negative values, and electrodes with nonlinear preference for long VOTs (e.g. electrode e3 in Fig. 2.5B) to positive values. The distribution of this measure for voiced-plosive electrodes (Fig. 2.5D, red distribution) shows significantly greater nonlinear bias compared to the remaining plosive electrodes (Fig. 2.5D, gray distribution) (p<0.001, Wilcox rank-sum test). This suggests a specialized mechanism for spatially distributed, nonlinear rate encoding of VOT, and contrasts with previously described temporal encoding mechanisms (Engineer et al., 2008; Steinschneider et al., 2005).

We performed a similar analysis for fricatives, measuring duration, which aids the distinction between voiced (/z v/) and unvoiced fricatives (/s ʃ θ f/); spectral peak, which differentiate /f v/ vs. coronal /s z/ vs. dorsal / ʃ /; and second formant (F2) of the following vowel (Ladefoged et al., 2010) (Fig. S2.12). These parameters can be decoded reliably from population responses (Fig. 2.5A. p<0.001, t-test).

Since plosives and fricatives can be sub-specified using similar acoustic parameters, we determined whether the response of electrodes to these parameters depends on their phonetic category (i.e. fricative or plosive). We compared the partial correlation values of neural responses with spectral peak, duration, and F2-onset of fricative and plosive phonemes (Fig. 2.5E), where each point corresponds to an electrode color-coded by its cluster grouping in Fig. 2.3D. High correlation values (r = 0.70, 0.87, 0.79, p<0.001, t-test) suggest that electrodes respond to these acoustic parameters independent of their phonetic context. The similarity of responses to these isolated acoustic parameters suggests that electrode selectivity to a specific phonetic features (shown with colors in Fig. 2.5E) emerges from combined tuning to multiple acoustic parameters that define phonetic contrasts (Mesgarani et al., 2008; Nelken, 2008).

2.2.6 Neural correlates of speech perception

We have thus far explored the representation of acoustic-phonetic speech features in human STG. To evaluate how well this representation correlated with human psychoacoustic behavior, ten normal subjects were recruited for a psychoacoustic labeling experiment. Subjects were asked to identify the consonant from the original CV tokens (16 consonants, 3 vowels, 6 speakers) embedded in white noise (0dB SNR). We hypothesized that the neural representation of speech would correlate well with the ability of humans to accurately perceive phonemes.

The classic way to examine phonetic representations of speech behaviorally has been to induce confusability by embedding speech in noise and measuring the amount of perceptual errors (Miller et al., 1955). Phonemes commonly mistaken for one another are thought to be closer in perceptual space. We found that errors aligned along linguistic features, such as manner of articulation (Fig. 2.6A), consistent with previous investigations (Miller et al., 1955). On average, subjects were able to identify the correct consonant 61% of the time.

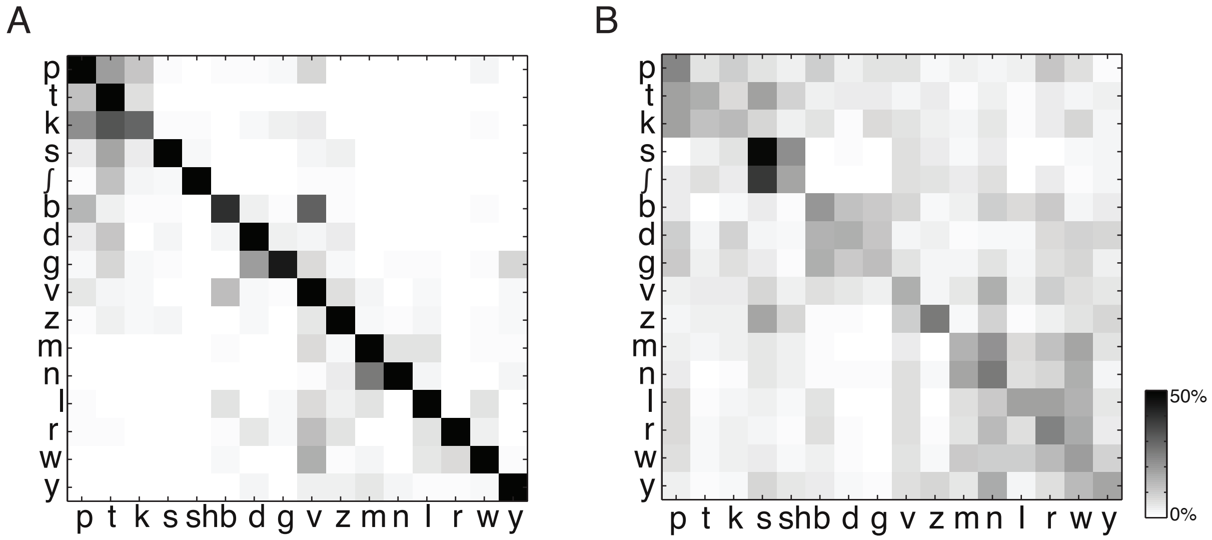


Fig. 2.6. (A) Psychophysics data collected during perception of CV syllables embedded in white noise. (B) Neural confusion matrix from consonant decoding using trained GMMs.

To obtain a probabilistic model of neural processing during speech perception, we trained multivariate Gaussian mixture models on a subset of the neural activity when listening to CV syllables. Decoding of consonants from the remaining trials showed high accuracy (Fig. 2.6B) Across subjects, consonants were accurately decoded 35% of the time; 6.25% is chance performance. The neural confusion matrix showed a significant correlation with normal phoneme identification errors (r=0.44, p<0.001). Our results suggest that human perceptual confusions can be accounted for by phonetic selectivity properties found in human STG.

2.3 Discussion

We have characterized the STG representation of the entire American English phonetic inventory. We used direct cortical recordings with high spatial and temporal resolution to determine how selectivity for phonetic features is correlated to acoustic spectrotemporal receptive field properties in STG. We found evidence for both spatially local, and distributed selectivity to perceptually relevant aspects of speech sounds, which, together, appear to give rise to our internal representation of a phoneme and accounts for human perceptual accuracies.

We found selectivity for some higher-order acoustic parameters, such as clear examples of non-linear, spatial encoding of VOT, which could have important implications for the categorical representation of this temporal cue. Furthermore, we observed a joint differential encoding of F1 and F2 at single cortical sites, suggesting evidence of spectral integration previously speculated in theories of combination-sensitive neurons for vowels (Chechik et al., 2012; Mesgarani et al., 2008; Nelken, 2008; Sussman, 1986)

Our results are consistent with previous single unit recordings in human STG which have not demonstrated invariant, local selectivity to single phonemes (Chan et al., 2013; Creutzfeldt et al., 1989). Instead, our findings suggest a multidimensional feature space for encoding the acoustic parameters of speech sounds (Mesgarani et al., 2008). Phonetic features defined by distinct acoustic cues for manner-of-articulation were the strongest determinants of selectivity, whereas place-of-articulation cues were less discriminable. This explains patterns of perceptual confusability between phonemes (Miller et al., 1955), and is consistent with feature hierarchies organized around acoustic cues (Stevens, 2002), where phoneme similarity space in STG is driven more by auditory-acoustic properties than articulatory ones (Liberman et al., 1967; Liberman et al., 1985). A featural representation has greater universality across languages, minimizes the need for precise unit boundaries, and can account for co-articulation and temporal overlap over phoneme-based models for speech perception (Stevens, 2002).

2.4 Methods

We recorded cortical activity from seven human subjects implanted with a multi-electrode array as part of their clinical evaluation for epilepsy surgery. Subjects listened to speech samples of individual syllables and continuous. We used phonetic transcriptions to segment the sentences into time-aligned sequences of phonemes. To quantify the selectivity, a phoneme selectivity index (PSI) was derived which indicates the total number of phonemes with cortical responses statistically distinguishable from the cortical response to a single particular phoneme (Wilcox rank-sum test, $p<0.01$). To determine selectivity patterns across electrodes and phonemes, we used unsupervised hierarchical clustering analyses across electrodes (rows) and phonemes (columns). We also explored an organizational alternative to single phonemes, by examining selectivity to a set of 14 phonetic features that have been defined previously in linguistics to contrast distinctive articulatory and acoustic properties of phonemes. We examined the neural encoding for several perceptually important acoustic parameters for three major phoneme groups: vowels, plosives, and fricatives using linear regression and partial correlation analysis. For the group of vowels, we estimated the instantaneous fundamental frequency (F0) and the first four formants (F1-F4). To measure the spectral peak of plosive and fricative phonemes, we used the acoustic spectrogram of phonemes to locate the maximum energy along the frequency axis. Finally, we used the phoneme transcriptions boundaries to extract Voice-Onset-Time of plosives and the duration of fricatives.

The experimental protocol was approved by the Committee for Human Research at the University of California, San Francisco.

2.4.1 Subjects

Seven human subjects underwent placement of a high-density subdural electrode array as part of routine clinical treatment of epilepsy, of which 3 listed to CV syllables and 6 listened to TIMIT sentences. The array contained 256 electrodes with 4 mm pitch. Subjects gave their written informed consent before surgery. All subjects had self-reported normal hearing and underwent neuropsychological language testing. The intracarotid sodium amobarbital (Wada) test was used for language dominance assessment. Our intracranial recordings were performed from 6 language dominant and 1 language non-dominant hemisphere. Nearly all patients were taking pain medication related to the craniotomy (acetaminophen and oxycodone, as needed). Half of patients were also taking stool softeners. Most patients were taken off seizure medication in order to lower their seizure threshold during intracranial monitoring.

Ten normal subjects were randomly recruited to participate in a psychophysical labeling experiment. Subjects had self-reported normal hearing.

2.4.2 Task

Subjects implanted with a subdural array listened to 48 consonant-vowel (CV) syllables (16 consonants, 3 vowels, 6 speakers), which represented a broad sample of syllables commonly spoken in American English. Stimuli were presented randomly and played between 15 to 18 times with an interstimulus interval of 1-2 seconds. Subjects passively listened to the speech sounds played monaurally from a loud speaker at a comfortable level.

Subjects were also asked to listen to speech samples from a publically available database commonly used in speech communication research, TIMIT (Garofolo, 1993). TIMIT is a corpus of phonemically transcribed speech of American English where each transcribed element has been delineated in time. We chose 500 unique sentences from 400 different male and female speakers representing a rich variety of context and speakers. The sentences were divided into 4 sessions of approximately 124 sentences, with a one second silence between consecutive sentences. Each sentence was repeated at most two times for which the responses were averaged. We chose a subset of TIMIT phonemes with an occurrence frequency of more than 30 in the 500 sentences.

In the psychophysical labeling experiment, normal subjects were asked to identify the same CV tokens embedded in white noise with a 0dB signal to noise ratio. Stimuli were randomly presented via headphones, and subjects were allowed to adjust the volume to a comfortable level. Responses were recorded for offline analyses.

2.4.3 Data preprocessing

Electrocorticography signal was recorded with a multichannel amplifier optically connected to a digital signal processor (Tucker-Davis Technologies). Each channel was visually and quantitatively inspected for artifacts. Channels with high-frequency noise more than twice the standard deviation were removed from analysis (less than 9% of channels were rejected using this criteria). The data were then segmented with a 100-ms pre-stimulus baseline and a 400-ms post-stimulus interval. The common mode signal was removed by subtracting the average response of channels that were fed into the same amplifier from each channel (Mesgarani et al., 2012). The common mode referencing did not have a significant effect on the selectivity analysis (Fig. S2.5D).

The analytic amplitude of the high gamma response (70-150Hz), gamma (30-70Hz), beta (15-30Hz), alpha (7-14Hz), and theta (4-7Hz) were then extracted with the Hilbert transform. We focused our study mainly on high-gamma because it has been shown to correlate well with neural spiking activity and provides the most consistent auditory response from the electrocorticogram compared to other frequency bands on a single trial basis, and has the temporal resolution to resolve individual phonemes. Example responses to other frequency bands are shown in Fig. S2.2.

2.4.4 Electrode selection

Cortical sites on superior and middle temporal gyri with reliable evoked responses to speech stimuli were selected for all the subsequent analysis. Our inclusion criteria consisted of a t-test between responses to randomly selected time frames during passive speech presentation and in silence (p<0.01), resulting in 73, 92, 37, 102, 80, 82, and 95 electrodes for subjects one to seven, respectively.

2.4.5 Acoustic spectrogram and parameter estimation, and response

The TIMIT phonetic transcriptions were used to align responses at each electrode to all instances of a given phoneme and then averaged to compute the peri-stimulus time histogram PSTH response to that phoneme (Fig. S2.3A)(Mesgarani et al., 2008). Responses were aligned to the onset of the phoneme. To minimize the preprocessing effects, we did not try to normalize the variation in phoneme length.

We estimated various acoustic parameters for different classes of phonemes which have been identified to have a perceptually important function in speaker and speech perception (Ladefoged et al., 2010). For the group of vowels, we estimated the instantaneous fundamental frequency (F0) and the first four formants (F1-F4) using. The values of these features were segmented by the transcribed vowel boundaries, and reduced to one number by taking the median value over the duration of the vowel. To measure the spectral peak of plosive and fricative phonemes, we used the acoustic spectrogram of phonemes to locate the maximum energy along the frequency axis. Finally, we used the phoneme transcriptions boundaries to extract Voice-Onset-Time of plosives and the duration of fricatives.

2.4.6 Phoneme selectivity index

To characterize the phoneme selectivity of different electrodes, we used a statistical framework to test whether responses of an electrode to all instances of a phoneme pair are significantly different. We created two distributions for a given electrode and phoneme pair by measuring response amplitudes of that electrode to all the occurrences of the two phonemes 110ms after the onset of the phoneme (average response latency = 110ms). We then used a non-parametric statistical hypothesis test (Wilcox rank-sum test) to assess whether these distributions have different medians (p<0.01, corrected for multiple comparisons). PSI is an index of the number of phonemes that a particular phoneme has a statistically distinguishable neural response from, at a given electrode. This is a value between 0 and 32 (there are 33 phonemes). If a phoneme has a PSI of 0, no other phonemes have a distinguishable response; if a phoneme has a PSI of 32, all other phonemes have a distinguishable response. The steps in estimation of PSI are shown graphically in Fig. S2.3. The phoneme selectivity map is obtained by stacking the phoneme selectivity indices of all electrodes, resulting in a two dimensional matrix of 33 phonemes and 131 total number of electrodes for all subjects (Fig. 2.3A).

Phoneme selectivity was also estimated for frequency bands other than high gamma, where responses showed much lower selectivity to phonemes (Fig. S2.5C). We also measured the similarity of PSI vectors with respect to the distance between them. We found a small, but significant dependence, suggesting a weak spatial organization of electrode selectivity to phonetic features (Fig. S2.6B).

2.4.7 Spectrotemporal receptive fields

We estimated spectrotemporal receptive fields (STRFs) of sites from the passive listening to TIMIT using the normalized reverse correlation algorithm (STRFLab software package: strflab.berkeley.edu (Theunissen et al., 2001)). To prevent overfitting, we used ridge regression with L2 regularization and optimized hyperparameters to maximize the mutual information between actual and predicted responses.

2.4.8 Phoneme separability analysis

We measured the separability of phonemes in acoustic and neural domain using the ratio of between-class to within-class variability (f-statistic) (Fig. S2.4A). This statistic can be estimated as a function of time lag with respect to the onset of phonemes and, therefore, shows the time at which the neural representation of different phonemes is most separable. In addition, the same statistics can be obtained from the first and second derivative of neural responses to measure the phoneme class specificity in the dynamics of the responses (Fig. S2.4B).

2.4.9 Phoneme to feature transformation

We used a subset of distinctive features defined by Chomsky and Halle to describe the articulatory and acoustic properties of each phoneme (Chomsky et al., l968). The list of these features and their included phonemes are shown in Figure S2.7. It should be noted that although phonemes form a mutually exclusive set of descriptors (each segment of speech signal corresponds to only one phoneme), the features do not. For example, a single segment of speech can have multiple features. Therefore, a phoneme can be considered a bundle of features. To map phoneme selectivity to feature selectivity (Fig. 2.3A to 2.3D), we summed the PSI values over all phonemes that shared a particular feature.

2.4.10 Linear regression analysis

We used linear regression analysis with least square cost function to estimate an optimal linear mapping between neural responses and acoustic parameters. We used cross-validation (20 non-overlapping subsets) to estimate an unbiased prediction for each of the values; thus, the predictions were always tested on held-out subsets not used in fitting the function.

2.4.11 Multidimensional scaling analysis

To examine the relational organization of the neural responses to vowels, we applied unsupervised multidimensional scaling (MDS) to the distance matrix of the mean neural responses. Thus, vowel tokens placed closed together in the MDS space elicited similar neural response patterns, whereas vowel tokens placed far apart elicited dissimilar responses. To calculate the distance between a pair of mean neural responses, the two responses were linearly correlated, and the resulting correlation coefficient was subtracted from 1. The final distance matrix was the average matrix across all subjects. A similar MDS analysis of the acoustic stimuli was completed using the mean acoustic spectrogram (described above).

2.4.12 Voice-Onset-Time analysis

We fitted a linear and exponential function to VOT-response pairs using nonlinear regression (intermediate steps shown in Fig. S2.11).

$$y_{lin}(x) = ax + b$$

$$y_{nonlin}(x) = a + b\,e^{cx}$$

where the coefficients are fitted to VOT-response pairs using least squares estimation. We measured the difference between the linear and nonlinear fits ($y_{nonlin}(x) - y_{lin}(x)$), as shown in

Fig. 2.5C) and quantified the degree of nonlinearity by averaging the second derivative of the difference between the fits (Fig. 2.5D):

$$crv = \sum_{x} \frac{d^2\left(y_{nonlin}\left(x\right) - y_{lin}\left(x\right)\right)}{dx^2}$$

We chose the exponential function as our nonlinear fit, because it showed a slightly more accurate fit to VOT-response pairs (comparison shown in Fig. S2.11D), however the observed nonlinear relationship shown in Fig. 2.5C,D does not depend on the nonlinear fitting function, and can also be obtained with a second order polynomial (Fig. S2.11E).

2.4.13 Gaussian mixture models

To obtain a probabilistic model of neural responses to speech sounds, we trained multivariate Gaussian mixture models (GMM) on neural responses for each consonant heard. First, responses were projected into a lower dimensional space to maximize class discrimination between the 16 consonants using linear discriminate analysis (LDA). This resulted in 15 different neural features. Next, the derivative of these features were calculated, and included as additional features for a total of 30 features. Using this feature space, a GMM was trained on four-fifths of the single trial neural responses to each consonant. The remaining responses were used to evaluate the log likelihood that a single trial was generated from a GMM. The responses were classified as the consonant with the GMM that yielded the highest log likelihood. This process was repeated five times, and the decoding accuracies were averaged. Final decoding accuracies were defined following a parameter grid search to optimize the number of LDA features and the number of Gaussians.

2.5 Supplementary materials

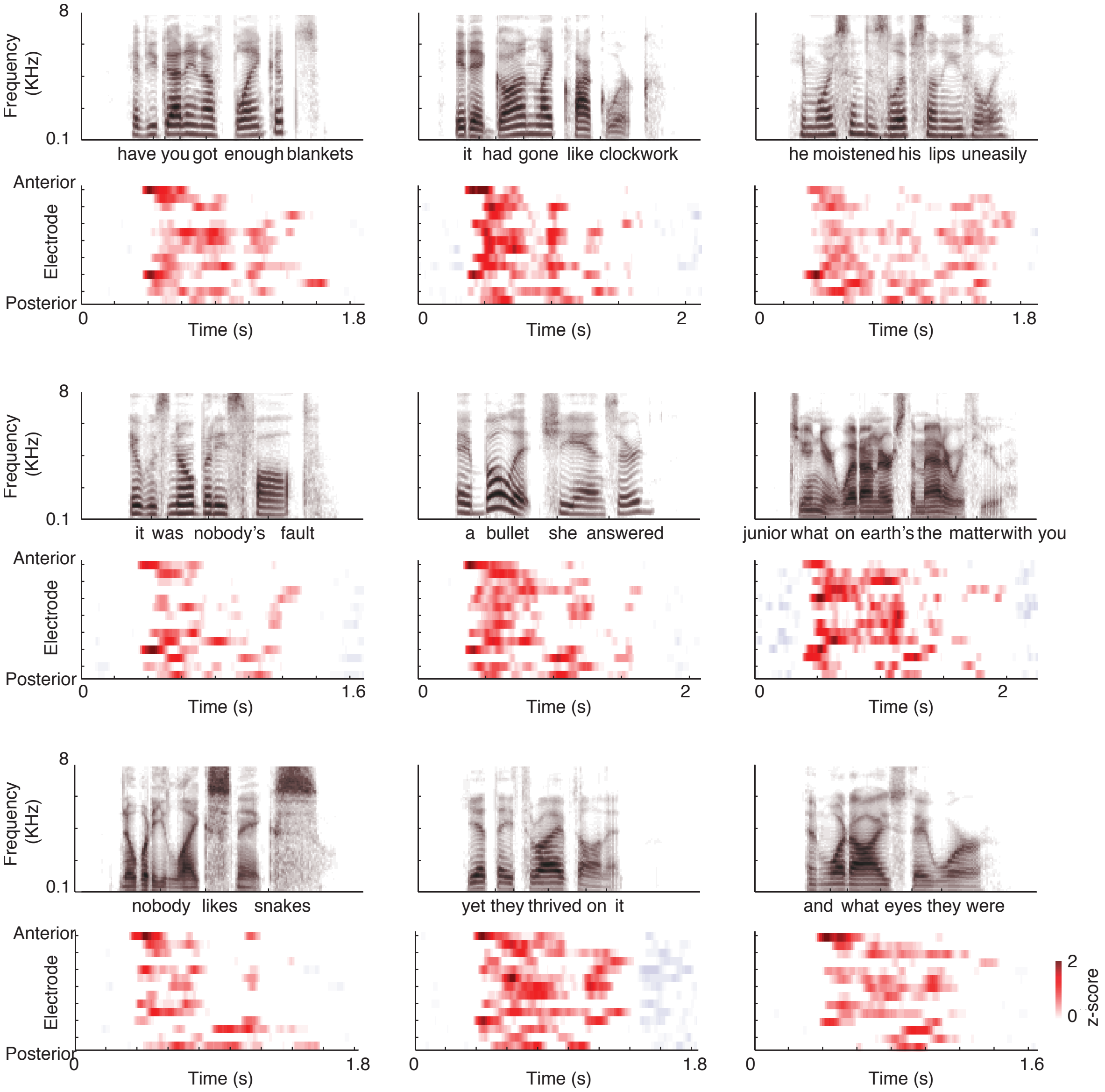


Fig. S2.1. Examples of electrode responses to speech sentences. High-gamma amplitude responses of selected electrodes from one subject while listening to nine different speech samples show a diverse spatiotemporal representation of speech in STG. These selected sites were most responsive to speech in comparison to silence.

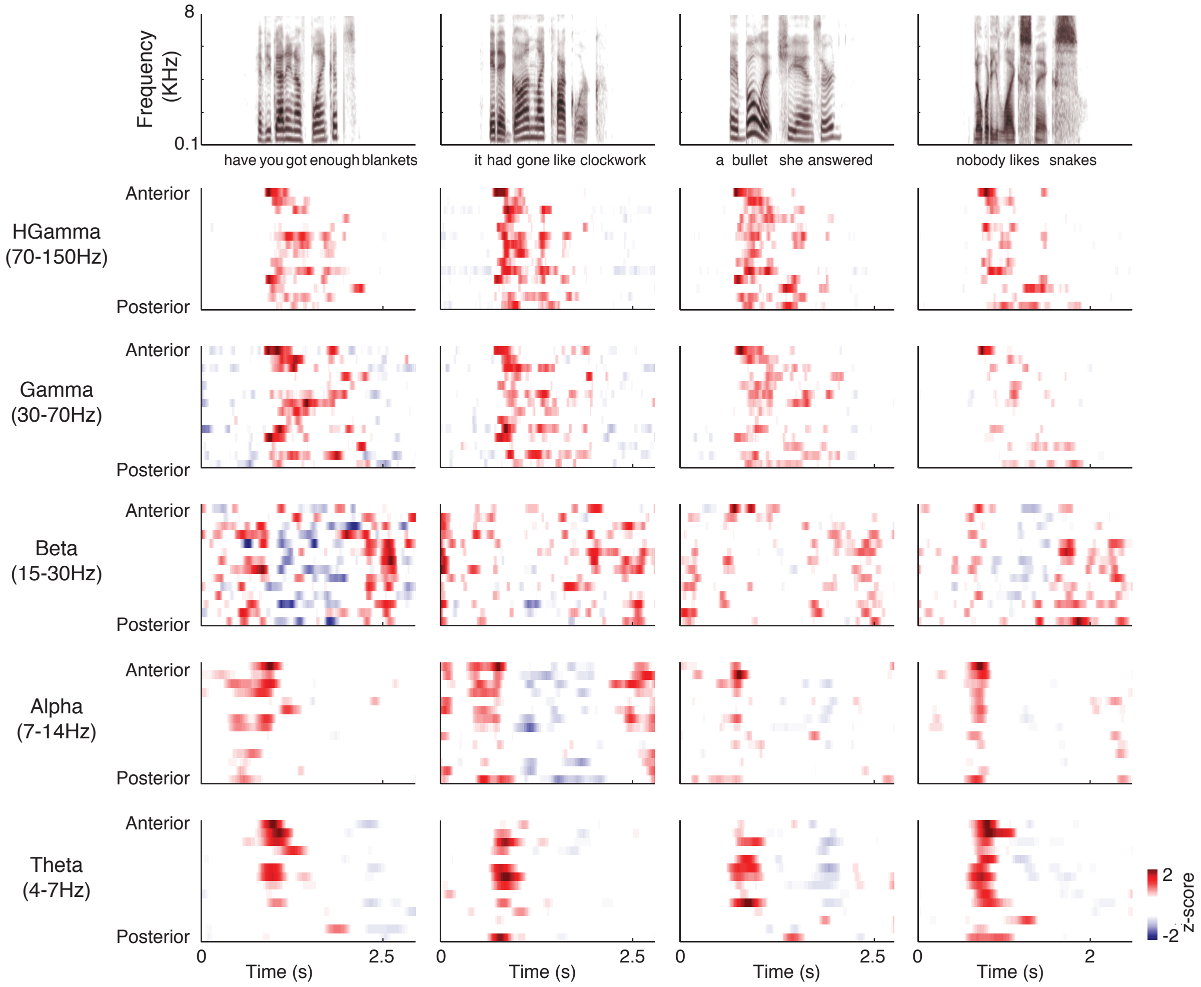


Fig. S2.2. Examples of electrode neural responses to speech samples. Amplitude responses of high gamma, gamma, beta, alpha, and theta bands for selected electrodes from one subject while listening to 4 different speech sentences.

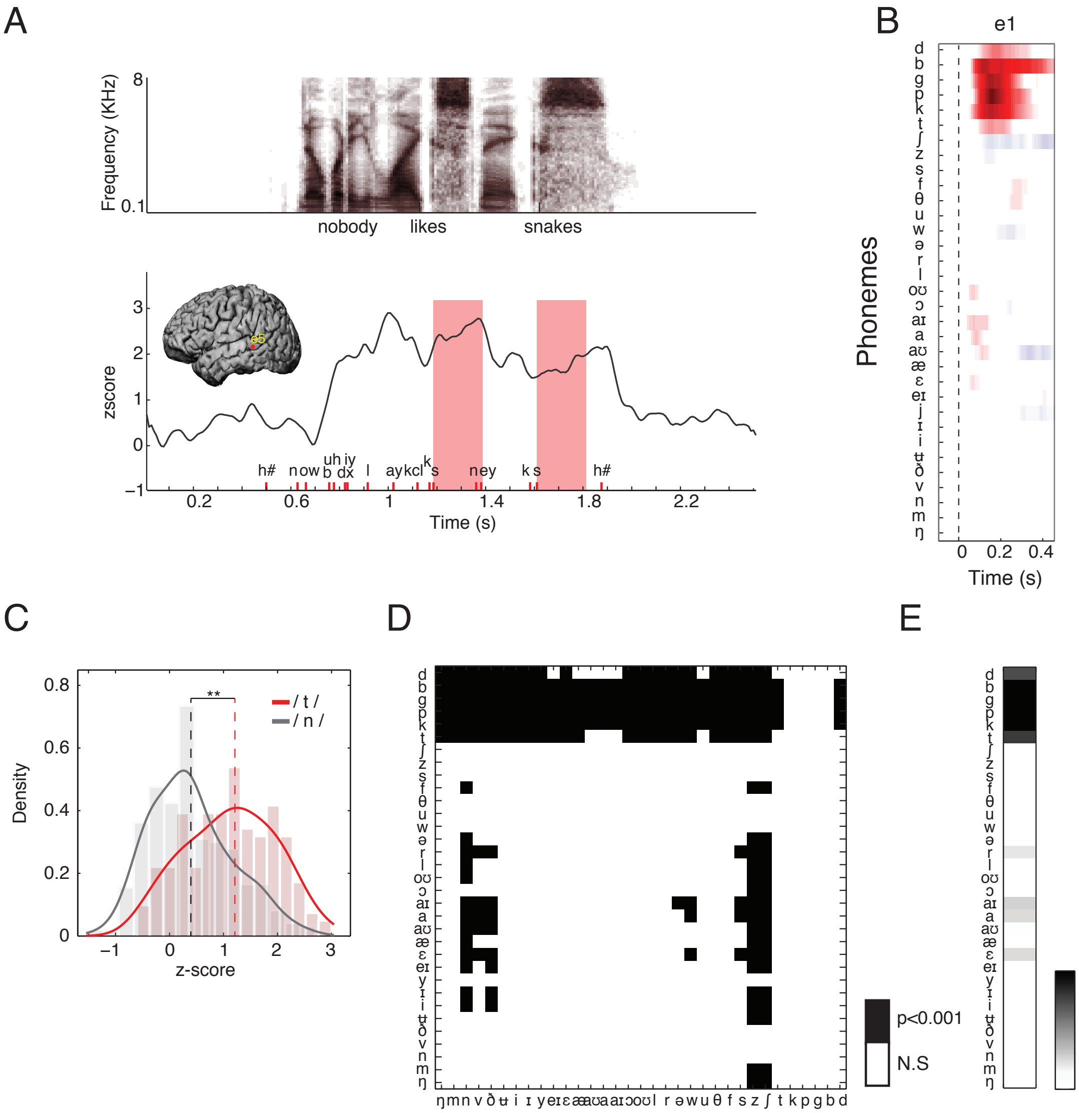


Fig. S2.3. Estimation steps of Phoneme Selectivity Index (PSI). (A) Spectrogram of a speech sample and the corresponding neural response of an example electrode (B) Average phoneme response of the example electrode to all phonemes show a strong preference for plosive phonemes. (C) Distribution of neural responses to all instances of phonemes /t/ and /n/ shows a significant median difference (**p<0.01, Wilcox rank-sum test). (D) Result of the hypothesis test for all phoneme pairs show that the plosive phonemes resulted in a significant change in the neural response compare to most other phonemes. (E) PSI vector is estimated by summing the phoneme selectivity matrix across all phonemes. Each element therefore indicates the total number of phonemes whose cortical responses were statistically distinguishable from responses to a single particular phoneme (0 is nonselective, 32 is extremely selective).

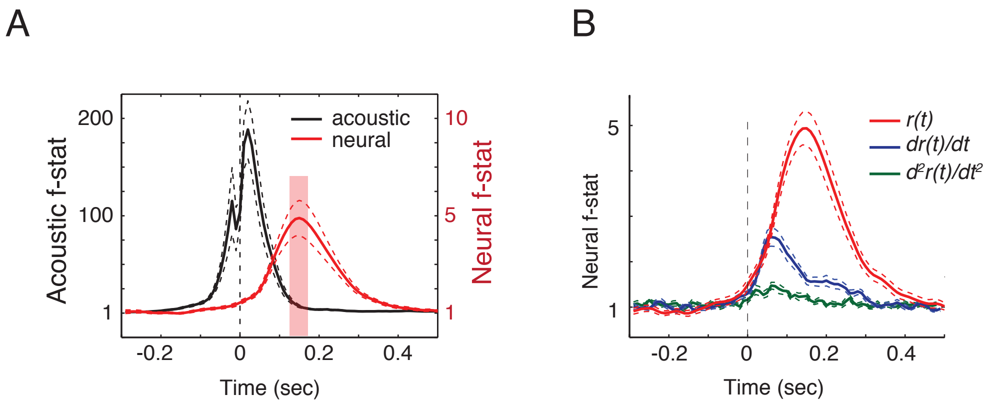


Fig. S2.4. Separability analysis of phonemes in acoustic and neural responses. (A) F-statistic defined as the ratio of between-phoneme to within-phoneme variability derived from acoustic spectrograms (black) and neural population responses (red) relative to the onset of the phonemes (vertical dashed-line). F-statistic peaks at 25ms after the onset of the phoneme in acoustic and 150ms in neural space. The shaded area shows the integration window used for selectivity analysis (B) F-statistic estimated from neural data and its first and second derivatives. The separation of phonemes is more evident in the amplitude of the neural responses compare to their temporal dynamics.

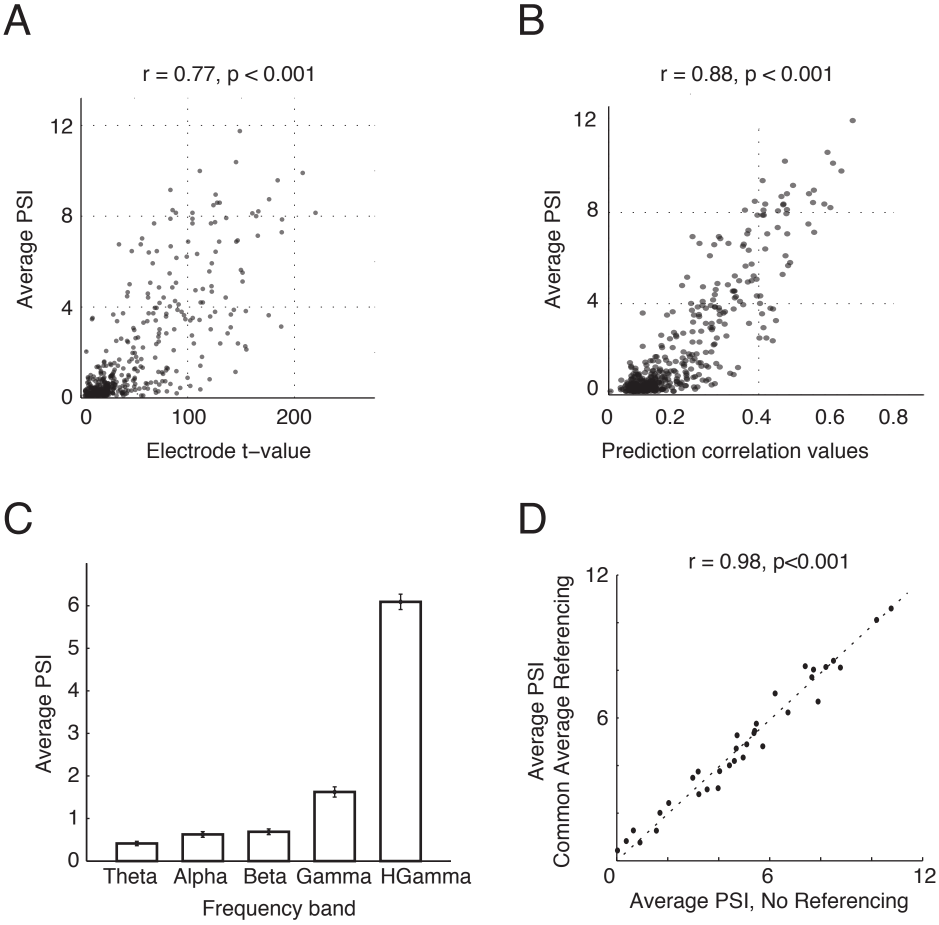


Fig. S2.5. Quantifying factors affecting phoneme selectivity. (A) Averaging the PSI over all phonemes summarizes the overall selectivity for a given electrode and is highly correlated to electrode responsiveness to speech (r = 0.77, ***p<0.001, t-test) and (B) linear STRF prediction values (r = 0.88, ***p<0.001, t-test). The high correlation values imply that phoneme selectivity is an inherent property of STG sites, where most speech responsive sites selectively respond to specific phoneme groups, and can be accounted for by a linear STRF model. (C) Average phoneme selectivity index (PSI) estimated from different neural frequency bands shows a much greater phoneme selectivity in high-gamma band compare to Gamma, Beta, Alpha, and Theta bands. (D) Average PSI for all electrodes in one subject with and without re-referencing the electrode responses. High correlation value (r = 0.98, ***p<0.001, t-test) shows that phoneme selectivity is unaffected by referencing.

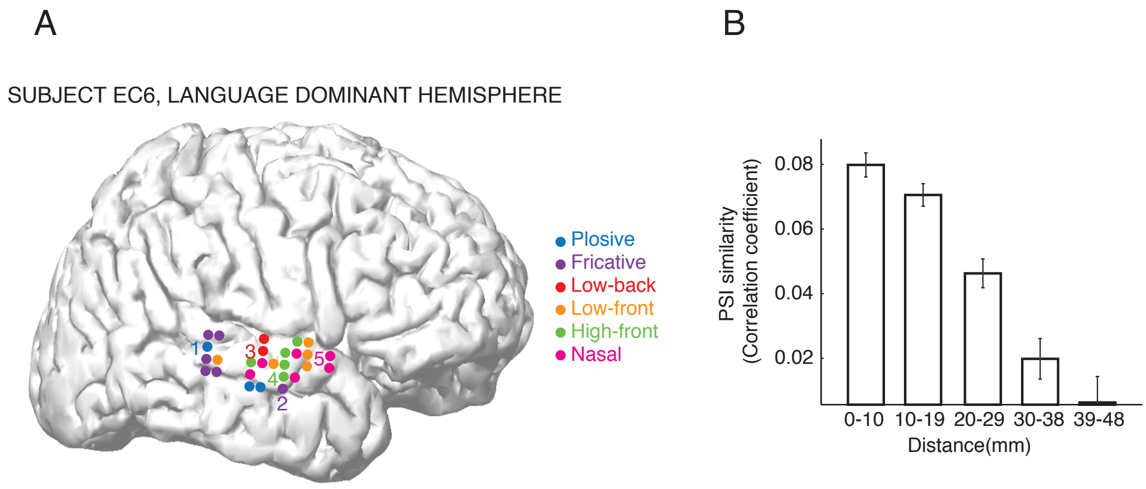


Fig. S2.6. Spatial organization of responses in STG. (A) Location of electrodes in one subject color coded by cluster membership in Fig. 2.3C shows a dispersed pattern (B) Correlation values of electrode PSIs plotted against their distance shows a small but significant relation between similarity of PSIs with distance between electrodes.

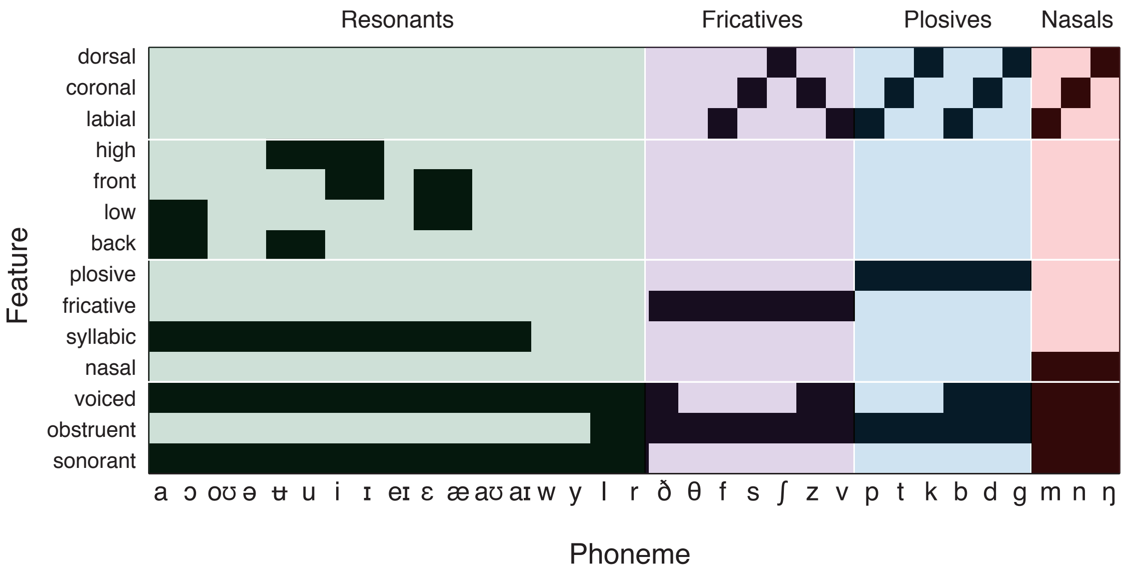


Fig. S2.7. Phoneme to feature mapping. Each phoneme can be deconstructed into a combination of binary features that describe its voicing, manner and place of articulation as shown in this binary matrix for the 33 phonemes in subgroups of vowels, fricatives, plosives and nasals.

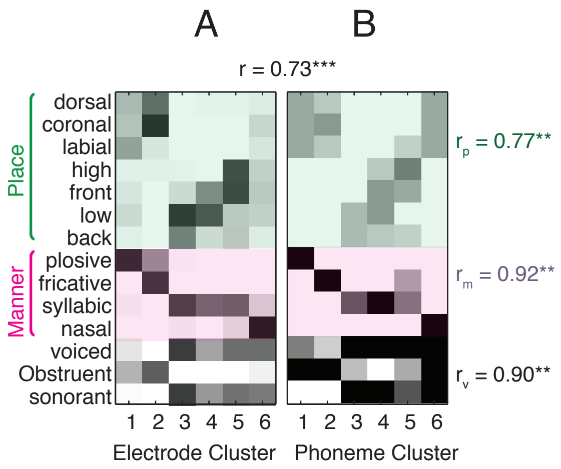


Fig. S2.8. Comparison of global and local selectivity and tuning properties**.** (A) Average PSI values for the six main single electrode clusters (from Fig. 2.3C). (B) Average PSI values for the main population-based phoneme clusters (Fig. 2B). Overall correlation of the two maps is r = 0.73, ***p<0.001, t-test. Correlation for subgroups of features is 0.90 for voicing, 0.92 for manner and 0.77 for place features (**p<0.01, t-test).

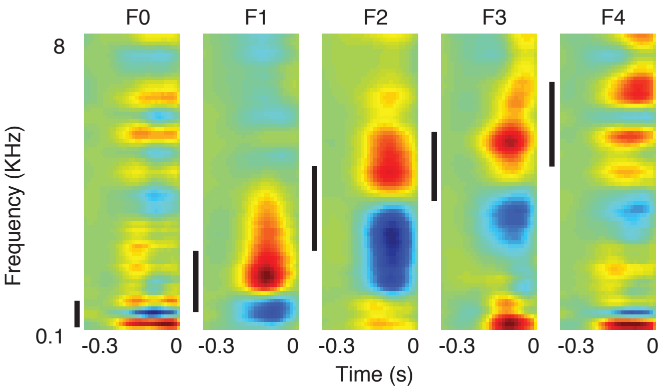


Fig. S2.9. Average weighted STRFs for fundamental and formant frequency decoding. Average weighted STRFs for each fundamental and formant frequencies, where each electrode's STRF was weighted by the regression beta values. These weighted STRFs reveal tuning for formant spectra, and show strong contrastive tuning edges within each formant range. Vertical black bars denote the acoustic spectral range for each parameter.

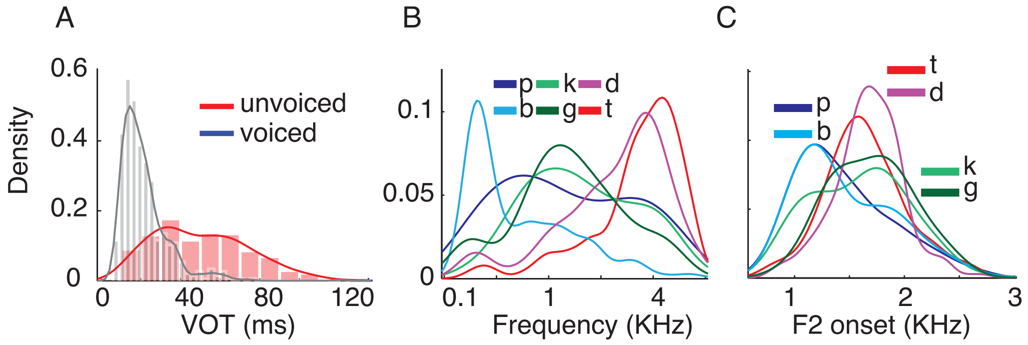


Fig. S2.10. Distribution of acoustic parameters of plosives. (A) Distribution of voice-onset-time (VOT) (B) spectral peak, and (C) second formant of the following vowel for all plosive phonemes.

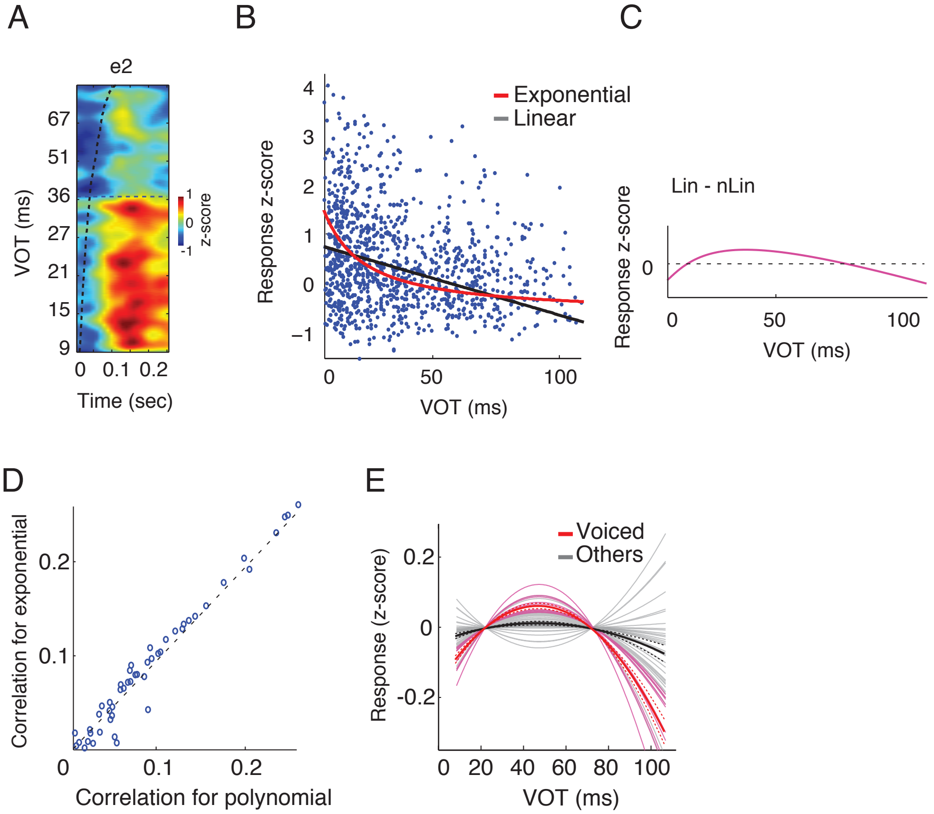


Fig. S2.11. Measuring the degree and shape of VOT-response relationship. (A) An example voiced-plosive electrode with strong preference for short VOTs. (B) Scatter plot of neural response amplitude vs. VOT for the same electrode was used to fit a linear (black plot) and an exponential (red plot) curves to VOT-Response pairs using linear and nonlinear regression. (C) The difference between linear and exponential fits (black and red plots) shows

the degree and direction of the nonlinearity in VOT-response transformation. Averaging the second derivative of the difference function with respect to VOT results in a scalar value proportional to the degree of nonlinearity, and its sign indicates preference for short vs. long VOT sounds. (D) Comparison of exponential vs. second order polynomial functions fitted to VOT-response pairs. Exponential function shows a slightly better fitting accuracy (E) Nonlinearity plots for all plosive electrodes estimated using a second-order polynomial function show similar degree and direction of nonlinearity as the exponential function used in Fig. 2.5C.

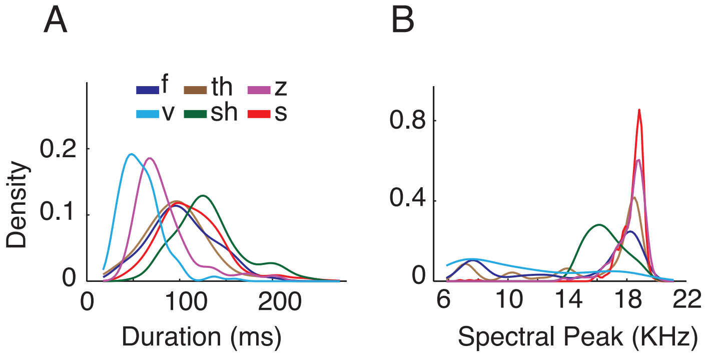


Fig. S2.12. Distribution of acoustic parameters of fricative phonemes. (A) Distribution of duration and (B) spectral peak for all fricative phonemes.

CHAPTER 3

Representation of speech sounds in human motor cortex

Perception engages the sensory system, but precisely if and how the motor system represents perceptual information is still poorly understood. Speech is an ideal context to address this issue in humans. In this chapter, we used direct human cortical surface recordings while subjects listened to speech sounds to determine how they are represented in the motor cortex. We found that the pattern of evoked responses in motor cortex during listening was completely different than those observed during production of the same sounds. Neural activity was found only in the ventral-most and dorsal-most regions of motor cortex during perception, whereas it was distributed throughout the sensorimotor cortex during production. In fact, the representational structure was organized along acoustic features rather than articulatory features. Furthermore, these evoked responses in motor cortex revealed definable spectrotemporal tuning, primarily to the feature of voicing. Our results provide evidence for an acoustic sensory representation of speech in motor cortex, and suggest that sensorimotor cortex does not represent articulation commands of perceived actions.

This work was done in collaboration with Keith Johnson and Edward F. Chang.

3.1 Introduction

Our motor and sensory systems are traditionally thought to be functionally separate systems (Fodor, 1983; Hubel, 1995). However, an accumulating number of studies have led their roles in action and perception to be thought of as increasingly integrated. For example, studies have demonstrated that both sensory and motor cortices are engaged during perception (Broca, 1861; Cogan et al., 2014; di Pellegrino et al., 1992; Gallese et al., 1996; Wilson et al., 2004).

The discovery of mirror neurons that fired when a monkey produced an action and observed a similar action in macaque F5 (di Pellegrino et al., 1992; Gallese et al., 1996)– human homologue to Broca's area (Brodmann 44) – has led to the hypothesis that integration of action and perception in motor cortex provides a unified basis for cognitive understanding (Pulvermuller et al., 2010). In humans, this hypothesis also provides a neural mechanism for perceiving speech. This idea, called the motor theory of speech perception, was first introduced by Liberman and colleagues who originally noted that speech sounds spoken in different contexts result in ambiguous mappings between spectrotemporal acoustics and perceived phonemes (Delattre et al., 1962; Delattre et al., 2005; Liberman et al., 1954). They reasoned that articulatory representations could provide disambiguating information, and hypothesized that "the objects of speech perception are the intended phonetic gestures of the speaker" (Liberman et al., 1967; Liberman et al., 1985). The motor theory predicts that speech sounds activate their neural motor articulator counterparts. Activation of motor cortex to speech sound has therefore been interpreted as evidence that the motor system is critical to active perception (Pulvermuller et al., 2010).

However, the interpretation of this cortical activity has been controversial.
There are several challenges to the notion that perception is dependent on motor cortex. For example, individuals with severely impaired speech production due to lesions in motor cortex have relatively intact speech perception abilities (Broca, 1861). Additionally, supra-Sylvian cortex surrounding the central sulcus is known to exhibit responses to a variety of motor and sensory behaviors (Hatsopoulos et al., 2011; Matyas et al., 2010; Tremblay et al., 2003), which raises the question of whether observed activity necessarily reflects motor representations. One alternative interpretation of motor cortex activity during speech perception is that it is a form of

auditory sensory feedback that directly interfaces with the motor system (Hickok et al., 2011; Hickok et al., 2007), much like how somatosensory feedback interfaces with the somatomotor system in tactile perception paradigms. This may be a result of learned associations made from the sensory consequences of speech production (Hickok et al., 2011). Such an interpretation would suggest motor cortex activity reflects an acoustic sensory representation of speech in the motor system, and is not a critical system for perception.

These two competing interpretations are still unresolved in part because, while it is known that motor areas are involved in speech perception, it is still unclear what information is actually encoded in these responses. Addressing the representation of speech sounds in motor cortex may provide more definitive information regarding three fundamental questions: (1) Which areas of motor cortex are active when listening to speech sounds, (2) what speech features are encoded in motor cortex, (3) and what phonetic organization emerges from the distributed pattern of activity?

To address these questions, we recorded direct neural activity spanning the ventral half of the lateral sensorimotor cortex (vSMC, Brodmann area 1, 2, 3, 6b) and superior temporal gyrus (STG, Brodmann area 22). The vSMC and STG provide an optimal setting to understand and compare the representation of speech in motor cortex since they are two prototypical regions that encode speech motor articulator and acoustic sensory features, respectively. As already outlined in Chapter 2, since cortical processing of speech sounds is spatially discrete and temporally fast (Creutzfeldt et al., 1989; Formisano et al., 2008; Mesgarani et al., 2014), we recorded activity from five human subjects implanted with high-density multi-electrode arrays as part of their clinical epilepsy surgery evaluation. These recordings allow for simultaneous high spatial and temporal resolution neural recordings from speech motor and auditory cortex.

3.2 Results

3.2.1 Motor cortex activity to speech sounds

Participants performed complementary speech tasks that sampled a wide range of speech sounds and articulatory gestures. First, subjects listened passively to consonant-vowel (CV) syllables (American English male speaking 8 consonants followed by the vowel /a/, Fig. S3.1). In a separate task, subjects said the same CV syllables aloud.

To understand which cortical regions showed evoked activity during listening and speaking, we measured the average evoked activity during the listening and speaking CV tasks. As with the previous chapter, we focused our analysis on high gamma (70-150Hz) cortical surface local field potentials (Crone et al., 1998a), which strongly correlate with extracellular multi-unit neuronal spiking (Ray et al., 2011; Steinschneider et al., 2008). We aligned neural responses to the onset of speech acoustics (t=0) to provide a common reference point across speech sounds. When listening, activity spanned STG and two distinct motor regions in the ventral-most and dorsal-most regions of vSMC (Fig. 3.1A). We observed a different distributed spatial activity pattern in vSMC when speaking. When speaking, neural activity was spread throughout vSMC (Fig. 3.1B). STG activity was also seen due to auditory feedback. We next examined the locations of the significant activity during speech perception and production. When listening, there were 115 electrodes significantly active above baseline ($p<0.01$, t-test, compared to pre-stimulus silence period) in motor regions (defined as electrodes located dorsal to the Sylvian fissure, Fig. 3.1C). These electrodes were primarily in dorsal and ventral vSMC, with a few in supramarginal, inferior-, and middle- frontal gyrus. When speaking, a total of 362 electrodes in motor cortex were found to be significantly active (Fig. 3.1C, $p<0.01$, t-test, compared to pre-stimulus silence period). Critically, there was a subset of sites in motor cortex (98 out of 362, ~30%) that was active during both listening and speaking (Fig. 3.1C).

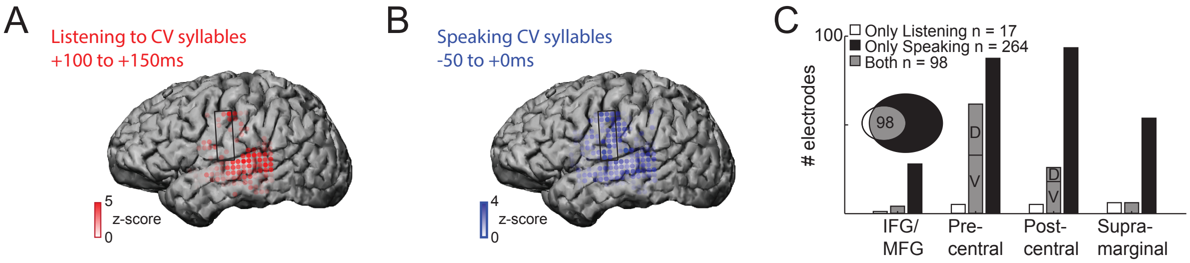


Fig. 3.1. Human neural activity. (A) Magnetic resonance image surface reconstruction of one subject's cerebrum. Average neural activity when subjects are listening to CV syllables are plotted with opacity signifying the high-gamma z-score. (B) Average activity when subjects are speaking CV syllables are plotted. (C) Number of electrodes and their locations are reported for sites that were significantly active in only the listening task, only the speaking task, and for both tasks (p<0.01, *t*-test, responses compared to silence and speech). D denotes dorsal vSMC sites and V denotes ventral vSMC sites.

3.2.2 Absence of motor articulatory features

Producing consonants with different constriction locations or places of articulation has been shown to evoke unique signatures in vSMC (Bouchard et al., 2013). For example, the plosive consonants /b/, /d/, and /g/ are produced by the closure of the vocal tract by the lips, front tongue, and back tongue, respectively (Fig. 3.2A, Fig. S3.1), and the vowel /a/ is produced by an open vocal tract and low, back-positioned tongue. These different constriction locations correspond to major articulatory features: labial, coronal tongue, and dorsal tongue, that can be robustly decoded from somatotopically organized neural responses in vSMC when speaking (Bouchard et al., 2013). To examine if activity between stimuli also exhibited somatotopic differences when listening to speech, we examined average cortical activity at single electrode sites distributed along the vSMC dorsoventral axis. Figure 3.2B shows activity across electrodes for speaking (blue lines) and listening (red lines) three CV syllables (/ba/, /da/, and /ga/). When speaking the labial syllable /ba/, electrodes in mid-vSMC were uniquely active (electrodes 5-6, blue lines). In contrast, more ventral electrodes were active when speaking /da/, a coronal tongue syllable (electrodes 8-10, blue lines). Even further ventral electrodes were active when speaking /ga/, a dorsal tongue syllable (electrodes 13, blue line). Other sites (electrodes 1-4, 11-12, blue lines) appeared to be active across all three syllables during speech vocalization. In contrast, when listening, the majority of the vSMC electrodes were not significantly active ($p>0.01$, t-test compared to silence, transparent red lines). For the few that were (electrodes 1, 2, 4, 11-12, solid red lines), responses were primarily active for all three syllables. In contrast to speech production neural signatures, somatotopically organized activity were not observed during speech perception. These examples suggest that somatotopically differentiable responses for different speech sounds cannot be qualitatively observed from individual vSMC sites.

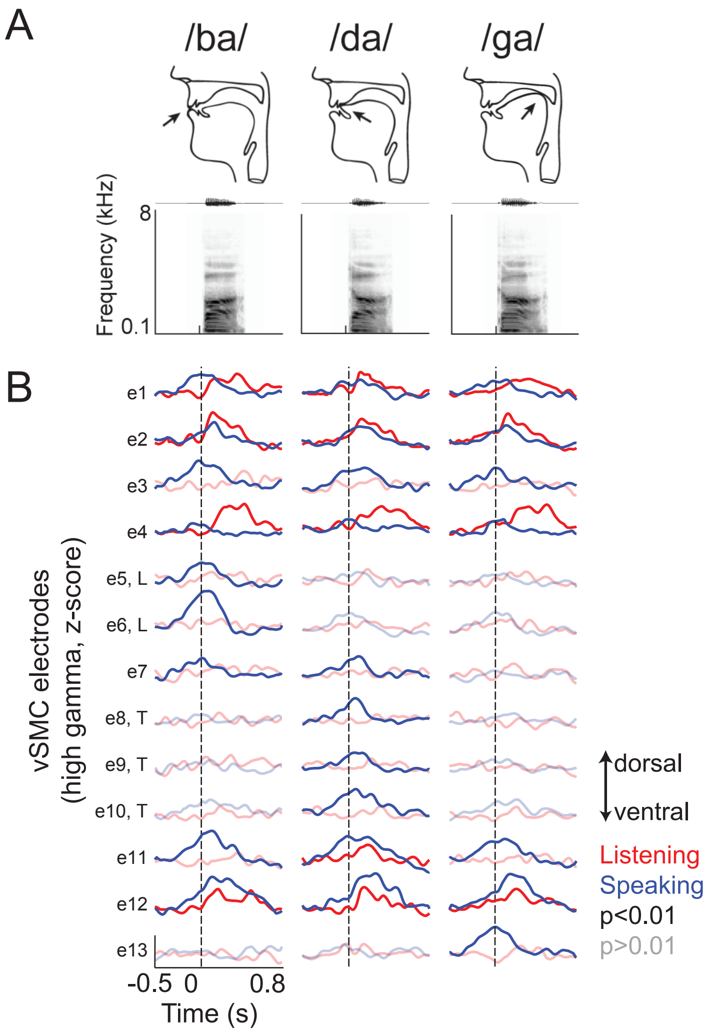


Fig. 3.2. vSMC activity when listening and speaking syllables. (A) Top, vocal tract schematics for three syllables (/ba/, /da/, /ga/) produced by the occlusion at the lips, tongue tip, and tongue body, respectively (arrow). Middle, acoustic waveforms of the spoken syllable. Bottom, spectrograms of syllables. (B) Average neural activity at electrodes for speaking (blue) and listening (red) syllables /ba/, /da/, and /ga/. Vertical dashed line indicates the onset of the syllable acoustics (t=0).

3.2.3 Distributed organization of speech

It is apparent from the activity across single vSMC electrodes that speaking requires coordination of multiple articulators, which is reflected in distributed spatial patterns of neural activity. While we did not observe any organization across individual vSMC electrodes during listening, it is possible that motor articulator representations emerge from the spatially distributed activity across motor cortex. We visualized the similarity of population activity evoked by different consonants using unsupervised multidimensional scaling (MDS), where the 2-dimensional Euclidean distances between stimuli markers correspond to the similarity of their neural responses (Chang et al., 2010b; Iverson et al., 2000). We determined an optimal analysis time window of 50 ms, centered at 50 ms and 180 ms after syllable onset for speaking and listening tasks, respectively, using a syllable separability analysis (f-statistic, Fig. S3.2). Visual inspection of MDS plots suggested that, during speech production, evoked activity in vSMC clustered along motor articulator features (Fig 3.3A): labials (/b/, /p/), coronal tongue (/s/, /sh/, /t/, /d/), and dorsal tongue (/g/, /k/) were represented by distinct patterns of population activity, consistent with previous investigations of vSMC organization (Bouchard et al., 2013). In contrast, neural responses during speech perception did not cluster along the same features (Fig. 3.3B). To confirm these results, we used unsupervised K-means clustering analysis to examine the grouping of neural responses, and its similarity to linguistically defined articulator feature categories. We measured the adjusted Rand Index (RI_{adj}) which quantifies the degree of agreement between two clustering schemes, where RI_{adj} = 1 denotes identical clustering and RI_{adj} = 0 denotes independent clustering (Hubert et al., 1985; Rand, 1971). While evoked activity during speaking showed strong clustering by motor articulator, activity during listening did not (Fig. 3.3C). Thus, motor responses during speech perception do not show a spatially distributed representation of speech motor articulators.

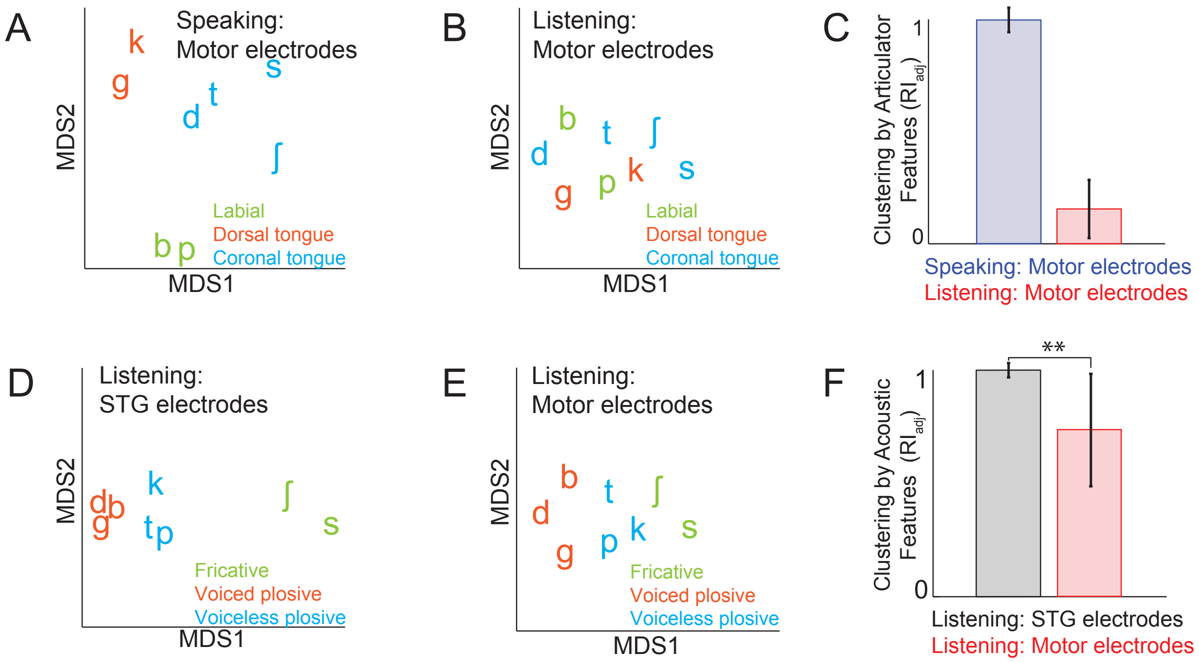


Fig. 3.3. Organization of neural activity patterns to speech. (A) Organization of mean motor responses using MDS when speaking. Token are colored by their main motor articulator. (B) Organization of mean motor responses using MDS when listening. (C) Organization by motor articulators. K-means clustering was used to assign mean neural responses to 3 groups (labial, dorsal tongue, coronal tongue) for both listening and speaking neural organizations (A,B). The similarity of the grouping to linguistically-defined major articulator groupings was measured by the adjusted Rand Index. An index of 1 indicates neural responses group by major motor articulator features. (D) Organization of mean STG responses using MDS when listening. Tokens are colored by their main acoustic feature. (E) Organization of mean motor responses using MDS when listening colored by their main acoustic feature. This panel is identical to (B), but recolored by their acoustic feature. (F) Organization by acoustic features (fricative, voiced plosive, voiceless plosive) for both STG and motor organizations when listening (D, E). The similarity of the grouping to known acoustic feature groupings was measured by the adjusted Rand Index.

We have thus far focused on the hypothesis that motor cortex activity during speech perception reflects the motor articulator counterparts of the speech sounds. Finding no evidence that major articulator features are either locally or spatially distributed, we next examined the hypothesis that motor activity during speech perception is an acoustic sensory representation of speech. To address this hypothesis, we compared the distributed spatial patterns of vSMC and STG neural activity, since STG was described in Chapter 2 to reflect an acoustic sensory representation of speech. Using multidimensional scaling, STG spatial patterns showed clustering according to three high-order acoustic features: voiced plosives (/b/, /d/, /g/), unvoiced plosives (/p/, /t/, /k/), and fricatives (/s/, /sh/) (Fig. 3.3D). This is consistent with the previous chapter's investigation of STG organization (Mesgarani et al., 2014), and similar to the relational organization derived by the stimuli acoustics (Fig. S3.3). With the same analyses, we observed that activity in motor cortex clustered along the same three acoustic features (Fig. 3.3E, note this panel is identical to Fig. 3.3B, simply re-colored). Unsupervised K-means clustering analysis confirmed that motor activity, when listening, organized into these linguistically defined acoustic feature groups, but was significantly weaker than the organization of STG (p<0.01, Wilcox rank-sum, Fig. 3.3F). This organization suggests that the motor activity during speech perception reflects an acoustic sensory representation of speech in the motor system that complements acoustic representations of speech in auditory cortex.

3.2.4 Acoustic sensory representation of speech

To further explore neural responses to a wider range of speech acoustics in a natural context, subjects were asked to listened passively to natural, continuous speech samples from a phonetically transcribed corpus with a range of American English speakers (TIMIT, 500 sentences, 400 speakers) (Garofolo, 1993). We estimated the mean cortical response at each motor site to every phoneme in English and found a diverse set of responses (Fig. S3.4A) that were notably weaker than STG responses (Fig. S3.4B, C). To characterize the responses, we measured the extent to which spectrotemporal properties of speech acoustics accounted for single site activity in motor cortex. A fraction of motor sites (14%) were strongly predicted with a linear spectrotemporal receptive field (STRF, r>=0.10, p<0.01) (Theunissen et al., 2001), indicating the existence of acoustic sensory sites in motor cortex. Significant STRFs across all subjects were localized to dorsal and ventral vSMC in addition to STG (Fig. 3.4A). Notably, the dorsal and ventral vSMC showed distinctly different weighted average STRFs, where each STRF was weighted by its prediction strength (Fig. 3.4B). Dorsal vSMC exhibited strong tuning to high frequencies and inhibition of lower frequencies (circled). In contrast, ventral vSMC showed strong tuning to both high and low frequencies (circled). Furthermore, the majority of STRFs in both regions showed strong low frequency tuning (100-200Hz) properties related to voicing (Fig. 3.4C), suggesting the acoustic sensory sites in motor cortex were sensitive to speech properties derived by movement of the larynx. These findings reveal that individual sites in motor cortex reflect sensory responses to spectrotemporal acoustics of speech, including voicing attributes, which give rise to the acoustic organization found in the distributed spatial patterns of neural activity.

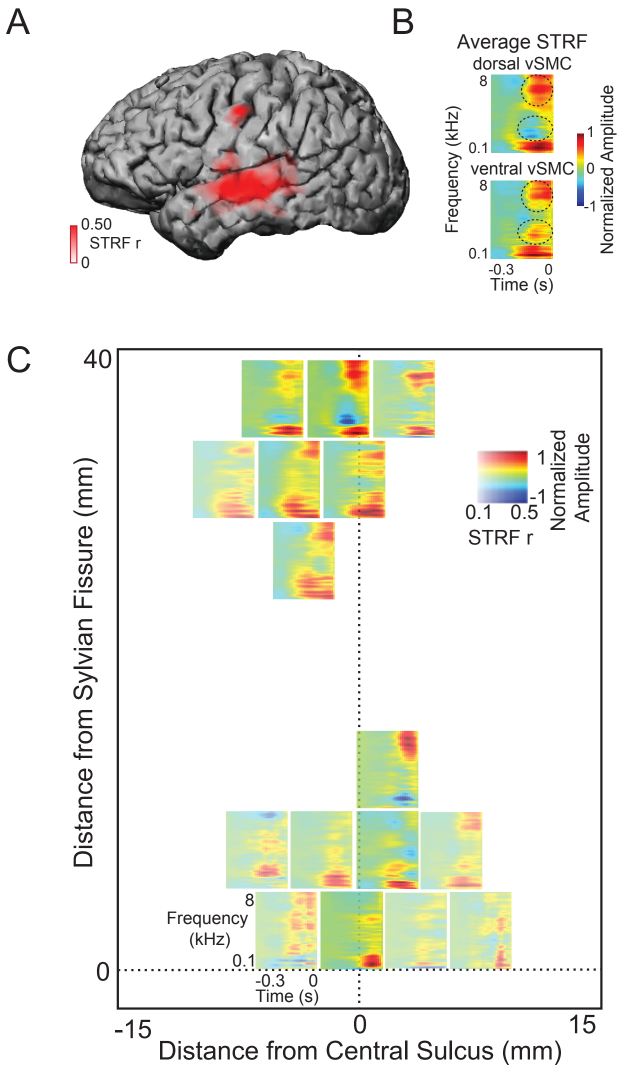


Fig. 3.4. Sensory representation of sound in vSMC. (A) STRF correlations are plotted with opacity signifying the strength of the correlation. (B) Average STRFs of two main electrode sites. (C) Individual STRFs plotted as a function of distance from the central sulcus and Sylvian fissure, with opacity signifying the strength of the STRF correlation.

3.3 Discussion

Using high-resolution cortical recordings and a wide array of American English speech sounds, we found that neural activity in vSMC during speech perception was spatially localized to dorsal and ventral vSMC during speech perception. A motor articulatory organization was absent in local and population neural activity. Furthermore, the activity reflected acoustic sensory properties of speech, with individual sites tuned for spectrotemporal acoustic features, suggesting a presence of acoustic sensory information in motor cortex.

Mirror neurons' unique property of firing during both action and observation of a task has been taken as evidence that motor cortex is critically involved in active perception. The motor theory of speech perception predicts that vSMC activity during perception reflects the same underlying representations that are active in these regions during speech production (i.e., speech articulators). Human neuroimaging studies have hinted at this by showing "mirror-like" activity in ventral premotor cortex (PMv) during passive listening to native and non-native syllables (Wilson et al., 2006; Wilson et al., 2004), and modulated PMv activity in phoneme categorization tasks (Alho et al., 2012; Callan et al., 2010; Chevillet et al., 2013). However, the neural activity patterns to a wide range of speech sounds in this study showed only a subset of vSMC to be active. Furthermore, while we observed differential responses to different speech sounds at the population level, these differences could not be explained by motor articulatory features, but instead reflected speech acoustics. These results are in direct contrast to predictions made by the motor theory of speech perception.

We found that, consistent with previous work, dorsal and ventral regions of the vSMC were active during speech perception. These distinct regions of the vSMC are known to control the laryngeal muscles, which in turn controls vocalization and pitch modulation (Brown et al., 2008). Though it remains to be seen how the two regions differ in laryngeal control, they are suggested to be a unique feature of human vocalization (Bouchard et al., 2013; Brown et al., 2008). In our study, these regions were tuned to high-order spectrotemporal properties, and showed a strong preference to voicing (100-200 Hz). Thus, the activity of these two regions reflects an acoustic sensory organization of speech sounds in motor cortex. These findings suggest the existence of a sensorimotor laryngeal motor region sensitive to speech perception and production.

Our results suggest sensory-motor laryngeal cortex in vSMC encodes auditory vocal information. These results have strong implications for speech production, as auditory feedback is potentially processed directly in the vSMC (Hickok et al., 2011). Speech production models currently propose a complex role for sensory feedback, where pathways exist for the activation of auditory cortex from vSMC activation (the forward prediction of production consequences), and the activation of vSMC from auditory and somatosensory input (the error correction signal) (Hickok et al., 2011). Given these proposed functional connections, activity in the vSMC from speech sounds may be a consequence of sounds activating the sensory feedback circuit (Hickok et al., 2011). However, our results also suggest that these models may need to be revised to incorporate a novel sensorimotor representation of speech in the dorsal and ventral regions of the vSMC that can account for PR models of speech perception.

3.4 Methods

3.4.1 Subjects

Five human subjects were implanted with high-density multi-electrode arrays as part of their clinical evaluation for epilepsy surgery. The array contained 256 electrodes with 4 mm pitch. Arrays were implanted on the lateral left hemispheres, but exact placement was determined entirely by clinical needs and varied between subjects. Using proprietary anatomic image fusion software from BrainLab (Munich, Germany), electrode positions were extracted by computed tomography (CT) scan, co-registered with the patient’s MRI and then superimposed on the subject’s 3D MRI surface reconstruction image.

All subjects were left hemisphere language dominant, as assessed by the Wada test. Subjects had self-reported normal hearing and normal neuropsychological language testing scores (including Boston Naming and verbal fluency tests). The study protocol was approved by the UC San Francisco Committee on Human Research, and all participants provided informed consent.

3.4.2 Task

Subjects completed three separate tasks that were designed to sample a range of phonetic features. First, subjects listened to eight consonant-vowel (CV) syllables between 5 to 18 times. Stimuli were presented randomly. Two subjects were instructed to simply passively listen to the sounds, which were presented with an interstimulus interval of 1-2 seconds. To remain alert, three subjects were asked to participate in a self-paced task to identify the syllable they heard by selecting from a multiple-choice question presenter on a computer with their ipsilateral hand. Each syllable was heard between 5 to 18 times.

Second, the subjects said the syllables aloud. Two subjects read the syllables aloud. Three subjects reported discomfort when reading, and were instructed to listen to an aurally presented stimulus and repeat aloud the syllable they heard. Only the neural activity from productions were used from this task. Each syllable was produced between 10-100 times.

Finally, subjects passively listened to natural speech samples from a phonemically transcribed public dataset, called TIMIT (Garofolo, 1993). We chose 500 unique sentences from 400 different male and female speakers. The sentences were divided into 4 sessions of approximately 124 sentences, with an interstimulus interval of 1-2 seconds. Each sentence was repeated at most two times. For analysis, we chose a subset of TIMIT phonemes that occurred more than 30 times. This resulted in an analysis of 33 phonemes.

Importantly, subjects were instructed to not move their articulators when listening. Studies have shown no detectable movement of articulators when listening to speech, suggesting that subvocalization does not occur during speech perception (McGuigan, 1979).

3.4.3 Data preprocessing

Electrocorticographic (ECoG) signals were recorded with a multichannel amplifier connected to a digital signal acquisition system (Tucker-Davis Technologies) sampling at 3,052 Hz. The produced speech was recorded with a microphone, digitized, and simultaneously acquired. The speech sound signals were presented monaurally from loudspeakers at a comfortable level, digitized, and also simultaneously acquired.

As in Chapter 2, line noise (60Hz and harmonics) was next removed from the signal with a notch filter. Each time series was visually and quantitatively inspected for additional excessive noise, and was subsequently removed from further analyses if its periodogram deviated more than two standard deviations away from the average periodogram of all other time series. The remaining time series were then common-average referenced and used for analyses. The analytic amplitude of each time series was extracted using eight bandpass filters (Gaussian filters, logarithmically increasing center frequencies (70-150Hz) with semi-logarithmically increasing bandwidths) with the Hilbert transform (Bouchard et al., 2013). The high-gamma power was calculated by averaging the analytic amplitude across these eight bands, and downsampling the signal to 100Hz. The signal was finally z-scored relative to the mean and standard deviation of baseline rest data for each channel.

3.4.4 Electrode selection

Supra-Sylvian cortical sites with robust evoked responses to both speech sounds and speech production were selected for this analysis. First, we implemented a t-test between neural responses to randomly selected time frames during the speech sound presentation and in silence ($p<0.01$). This resulted in 10, 22, 29, 27, and 27 electrodes for the five subjects ($n=115$). Next we implemented a t-test between neural responses to randomly selected time frames during speech production and in silence ($p<0.01$), resulting in 25, 74, 87, 92, and 84 electrodes ($n=362$). Finally, we took the intersect of these two electrode groups to arrive at our final motor electrodes set of 8, 16, 28, 22, and 24 electrodes active during listening and speaking ($n=98$). To analyze the responses of the auditory cortex, we restricted the infra-Sylvian cortical sites to those that were reliably evoked by speech sounds ($p<0.01$, t-test between silence and speech sounds neural responses).

3.4.5 Average neural response and peak high-gamma

For the speaking and listening to CV syllables tasks, the start of the syllable acoustics was used to align the responses of each electrode site. For the phoneme responses, the TIMIT phonetic transcriptions were used to align responses to the phoneme onset. Once responses were aligned to a stimulus, the average activity for each site to each stimulus was measured by taking the mean response over different trials of the same stimuli. The maximum of the mean responses to different stimuli were then used to measure the peak-high gamma distributions between different tasks and sites.

3.4.6 Syllable separability analysis

We measured the separability of the neural responses to different syllables by using the ratio of between-class to within-class variability (F-statistic). This statistic can be estimated as a function of time lag with respect to the onset of the syllable stimuli. This time analysis reveals the time points at which the neural representation of different syllables is most separable.

To assess if separability at a time point was significantly different from separability of neural responses during silence (null distribution), we used a bootstrap resampling method. We estimated the distribution of separability at a time point and the null distribution by sampling the F-statistic across subjects with replacement. The observed distribution was then assigned a bias-corrected and accelerated significance level (Efron et al., 1993). We finally corrected for multiple comparisons using the false discovery rate (FDR) approach (alpha < 0.005) (Benjamini et al., 1995).

3.4.7 Multidimensional scaling analysis

To examine the relational organization of the neural responses to syllables, we applied unsupervised multidimensional scaling (MDS) to the distance matrix of the mean neural responses to the electrodes of interest. We focused our analyses on the neural responses in the time window that showed most separability between responses. Syllables placed closer together in MDS space elicited similar neural response patterns, and those further apart from one another elicited more dissimilar patterns. To calculate the distance between a pair of mean neural responses, a mean neural response to one syllable was linearly correlated to another, and the resulting correlation coefficient was subtracted from 1.

3.4.8 Clustering analysis

We used unsupervised K-means clustering analysis to examine the grouping of the mean neural activity to syllables of the electrodes of interest. Focusing our analyses on the neural responses in the time window that had the most separable responses, we clustered the mean activity into 3 distinct clusters. This cluster number was chosen because there are 3 major place of articulations and manner of articulations in the syllable stimuli set that have been shown to play a major role in the neural organization of motor cortex during speech production and auditory cortex during speech perception.

After assigning the neural responses into three distinct groups, we measured the similarity of grouping to the grouping of syllables by place of articulation (Fig. 3.3C) and manner of articulation (Fig. 3.3F) using the adjusted Rand Index (RI_{adj}). The RI_{adj} is frequently used in statistics for cluster validation (Hubert et al., 1985; Rand, 1971). It measures the amount of agreement between two clustering schemes: one by a given clustering process (e.g. K-means), and the other by some external criteria, or gold-standard (e.g. place of articulation linguistic features). The RI_{adj} takes an intuitive approach to measuring cluster similarity by counting the number of pairs of objects classified in the same cluster under both clustering schemes, and controlling for chance (hence, "adjusted" RI). It has an expected value of 0 for independent clusterings, and a maximum value of 1 for identical clustering. It is defined as the following:

Let S be a set of *n* objects, $S = (o_1, o_2, \ldots, o_n)$. Partitioning the objects in two different ways such that $U = (U_1, \ldots, U_r)$ is a partition of *S* into *r* subsets, and $V = (V_1, \ldots, V_t)$ is a partition of *S* into *t* subsets, let:

a = number of pair of objects that are in the same set in *U* and in the same set in *V*,

b = number of pair of objects that are in the same set in *U* and in different sets in *V*,

c = number of pair of objects that are in different sets in *U* and in the same set in *V*,

d = number of pair of objects that are in different sets in *U* and in different sets in *V*.

Without adjusting for chance, the RI is simply:

$$RI = \frac{a+d}{a+b+c+d} = \frac{a+d}{\binom{n}{2}}.$$

Taking into account chance pairings, RI_{adj} becomes:

$$RI_{adj} = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c) + (c+d)(b+d)]}.$$

3.4.9 Phoneme selectivity index

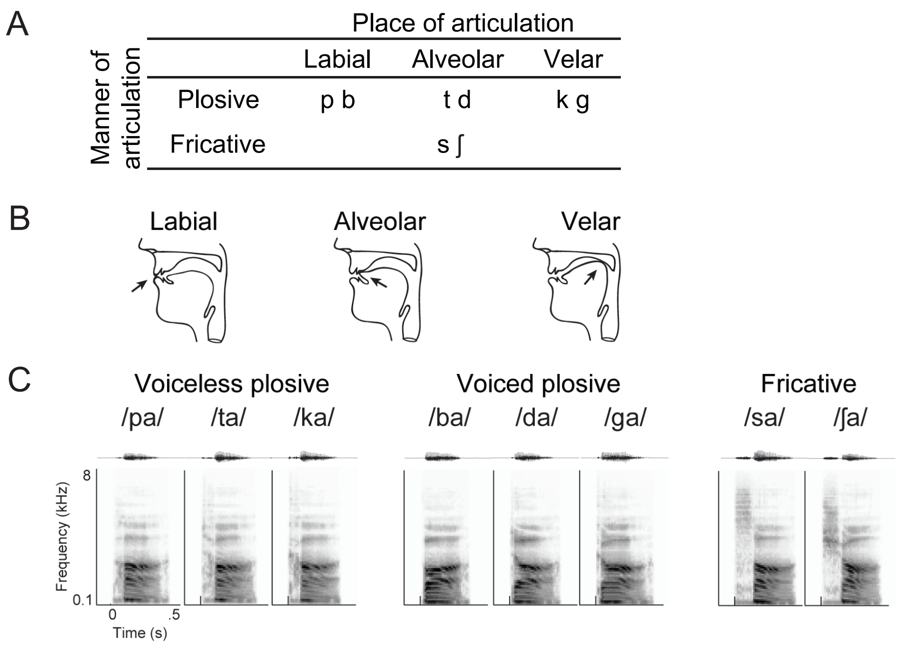
To characterize the phoneme selectivity of each electrode site, we implemented the PSI calculation described in Chapter 2 (Mesgarani et al., 2014). In short, for a single site, we summed the number of responses that were statistically different (Wilcox rank-sum test, $p<0.01$, corrected for multiple comparisons) from the response to a particular phoneme. This resulted in a PSI that ranges from 0 to 32, where a PSI = 32 is an extremely selective electrode and a PSI = 0 is not selective. A PSI describes an electrode's selectivity to one phoneme, and a vector of PSIs describes an electrode's selectivity profile to all phonemes.

3.4.10 Spectrotemporal receptive fields

To characterize the receptive field of each side, we implemented the STRF calculation described in Chapter 2. In short, the spectrotemporal representation of speech sounds was estimated using a cochlear frequency model, which resulted in a two dimensional spectrotemporal representation of speech sounds simulating the pattern of activity on the auditory nerve (Wang et al., 1994).

We then estimated the spectrotemporal receptive fields (STRFs) of the sites from passive listening to TIMIT using normalized reverse correlation between spectrotemporal representation of the sentences and the evoked neural activity (STRFLab software package: http://strflab.berkeley.edu) (Theunissen et al., 2001). We calculated the final STRF and correlation between the predicted and actual neural response using a cross-validation approach. To do this, a STRF was derived using nine-tenths of the stimuli-response pairs, and a correlation number was measured by predicting the remaining one-tenth responses. This was repeated ten times with ten non-overlapping stimuli-response pair sets. The final STRF and correlation number were derived by averaging the ten STRFs and correlation numbers.

3.5 Supplementary materials



	Place of articulation		
Manner of articulation	Labial	Alveolar	Velar
Plosive	p b	t d	k g
Fricative		s ʃ	

Fig. S3.1. Consonant-vowel syllable tokens. (A) Consonants of all syllable tokens. Consonants were paired with the vowel /a/, and are organized by place and manner of articulation. (B) Vocal tract schematics for three occlusions made with the lips, tongue tip, and tongue body (arrow). (C) Acoustic waveform and spectrogram of the syllable tokens.

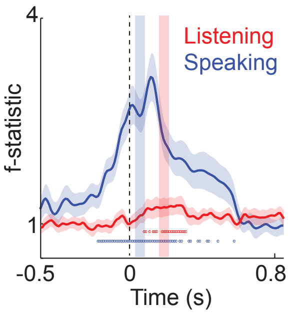


Fig. S3.2. F-statistic defined as the ratio of between-syllable to within-syllable variability derived from motor electrode activity when speaking and listening. Optimal time windows of 50ms (shaded) were determined to be centered at 50ms and 180ms for speaking and listening, respectively.

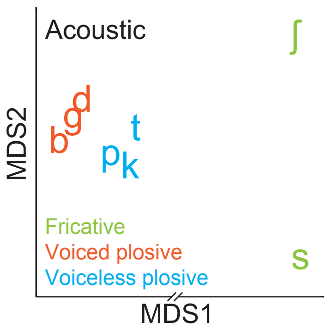


Fig. S3.3. Organization of acoustic spectrograms of tokens using MDS when listening. Tokens are colored by their main linguistically-defined acoustic feature.

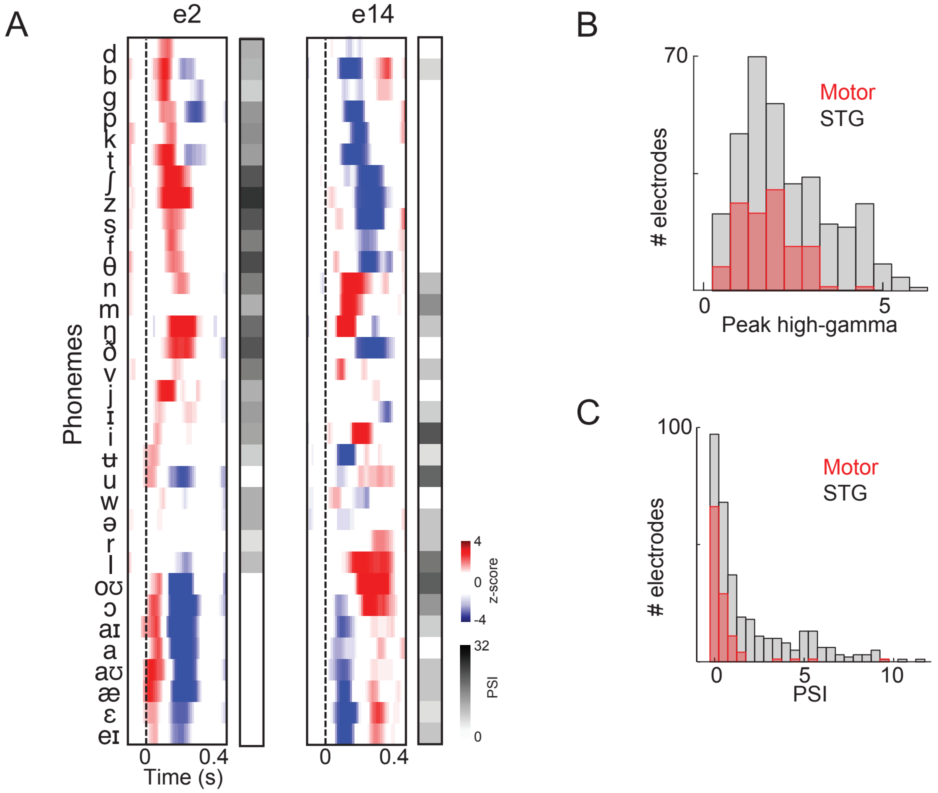


Fig. S3.4. (A) Average responses at sample electrodes to all English phonemes and their PSI vectors. (B) Peak high-gamma z-score distributions for motor and STG sites when listening. (C) Average PSI distributions for motor and STG sites.

CHAPTER 4

Real-time, time-frequency mapping of event-related cortical activation

As demonstrated in the previous chapters, direct human neural recordings enable us to understanding speech neuroscience, and have allowed us to make great academic strides in understanding this unique human function. Additionally, our understanding of speech processing using this recording method presents an exciting opportunity to affect clinical change.

Approximately 70,000 people are diagnosed with brain tumors and 150,000 with epilepsy each year in the United States. For thousands of these patients, resection of cortical tissue is the only viable option for effective treatment, and, depending on the cortical region of interest, awake functional mapping may be necessary for operative planning to avoid damaging eloquent cortex involved in speech processing. Approximately 30,000 of these cases are performed annually in the United States. Here, we explore the feasibility and advantages of ECoG recordings to functionally map human cortex, and revolutionize the current standard of care. Our system generates a robust functional motor and sensory cortical map in seconds, and demonstrates high concordance with results derived using independent stimulation mapping.

This work was done in collaboration with Edward F. Chang, and was published in The Journal of Neural Engineering (Cheung et al., 2012). The University of California, San Francisco has filed a PCT application on this work.

4.1 Introduction

Cortical mapping is a critical tool in safely carrying out neurosurgical procedures near "eloquent" brain regions. The traditional method for brain mapping is electrical cortical stimulation (ECS) (Haglund et al., 1994; Keles et al., 2004; Penfield et al., 1954; Penfield et al., 1959). Since only one site can be tested at a given time and must often be tested repetitively for confirmation, it is a highly inefficient method and can mis-represent important information about distributed cortical operations (Luders et al., 1991; Nii et al., 1996; Sinai et al., 2005). Cortical stimulation evokes unwanted seizures in up to 50% of cases (Ojemann et al., 1989; Sinai et al., 2005). Furthermore, ECS mapping can be highly operator dependent, and despite comprehensive ECS mapping patients can have neurologic impairments (Haglund et al., 1994; Keles et al., 2004; Krauss et al., 1996; Ojemann, 1979).

An alternative to ECS is measuring evoked cortical activity while patients are engaged in normal physiologic behavior. Electrocorticography (ECoG), which refers to direct recordings via electrodes placed on the cerebral cortex, can provide an accurate representation of spatiotemporal neural activity when sampling from a broad cortical surface (Canolty et al., 2007; Chang et al., 2010a; Crone et al., 1998a; Crone et al., 1998b; Edwards et al., 2010; Pfurtscheller et al., 2003). Recent studies have revealed that high gamma frequencies (70-150Hz) are an extremely discrete spatial and temporal marker of neural activity (Crone et al., 2001a; Crone et al., 2001b; Crone et al., 1998a; Crone et al., 1998b). Previous investigations have been largely constrained to offline analyses with a few exceptions of online application (Brunner et al., 2009; Lachaux et al., 2007; Miller et al., 2007; Roland et al., 2010; Schalk et al., 2008). In particular, several groups have demonstrated impressive passive localization of functional cortex (Brunner et al., 2009; Miller et al., 2007; Roland et al., 2010; Schalk et al., 2008). However, these recent demonstrations involve algorithms that effectively average signal across time, thus discarding temporal information that could be instrumental to understanding important dynamics, such as latency differences between activation in sensory, motor, and cognitive processes. If temporal information were to be retained, such dynamical differences could provide clinicians important insight into differing cortical functionality (e.g. motor versus somatosensory). This information can affect surgical planning decisions that drastically minimize morbidity (Haglund et al., 1994; Keles et al., 2004; Krauss et al., 1996; Ojemann, 1979; Ojemann et al., 1989). For example, patients can usually tolerate parasthesias from sensory cortex resection, but paralysis from inadvertent motor cortex injury is extremely debilitating. Furthermore, cognitive processes are often redundant or more distributed, and those impairments are more likely to recover from injury compared to deficits in basic sensorimotor processing (Duffau, 2012; Fox et al., 2012; Siegel et al., 2012).

Our goal was to develop a robust, reliable, and accurate system capable of performing real-time brain mapping based upon spectral changes evoked by brain activity during physiologic behavioral conditions. Intuitive visualization allows for rapid identification of spatial and temporal patterns during functional cortical activation with potential to supplement or replace ECS.

4.2 Results

To rapidly identify spatial and temporal patterns in ECoG signal during functional cortical activation, we implemented a real-time algorithm that processed ECoG signals while patients were engaged in sensorimotor tasks. Our system can be used to perform sensorimotor mapping using an event-triggered analysis in real-time within seconds.

4.2.1 Real-time mapping

Visualization: We first demonstrate that our method can be used to detect electrode-specific event-related spectral alterations. In the following presentation of data, neural activity is represented as z-score deviations from the baseline signal (depicted by a color scale). The event onset is denoted by a vertical dotted line at time 0 milliseconds. Each plot portrays all channels from the entire subdural array. In this section, we will discuss the three visual displays we believe to be most pertinent to this study.

Average plot: The average plot shows the running average event-related spectrogram, which is updated in real-time after each event trial. The average plots from subject A and E for the speaking and button press tasks are shown in Figures 4.1 and 4.2. Additional plots are included in the supplementary material (Fig. S4.1). It can be easily seen which electrodes are active immediately before and after the event onset. Thus, the average plot illustrates the localization of cortical functionality (e.g. sensory versus motor) based on the temporal patterns of activation while the patient is performing the task.

In Figure 4.1, the average plot is shown for the evoked spectrograms when the subject repeated the syllable /la/ (Task 1). Electrodes 15, 29, and 30 in the ventral precentral gyrus (primary motor cortex) show strong activation that are clearly event-locked, and begins slightly before event onset (-118 ms), suggesting the efferent neural motor commands for articulation. Electrode 29 demonstrated tongue movement with ECS. It can be seen that electrode 54 in the posterior superior temporal gyrus also reveals an increase in activity, however this only occurs after event onset (72 ms), and therefore is related to auditory afferent sensory processing (feedback of hearing oneself speak). Anomic speech arrest was found to occur at electrode 54 during ECS mapping while patient was tested with confrontational picture naming. Key channels from the plot can be more carefully examined on the right. Previously, the delineation of motor from sensory sites was done in offline analyses (Chang et al., 2010b; Crone et al., 1998a; Edwards et al., 2009).

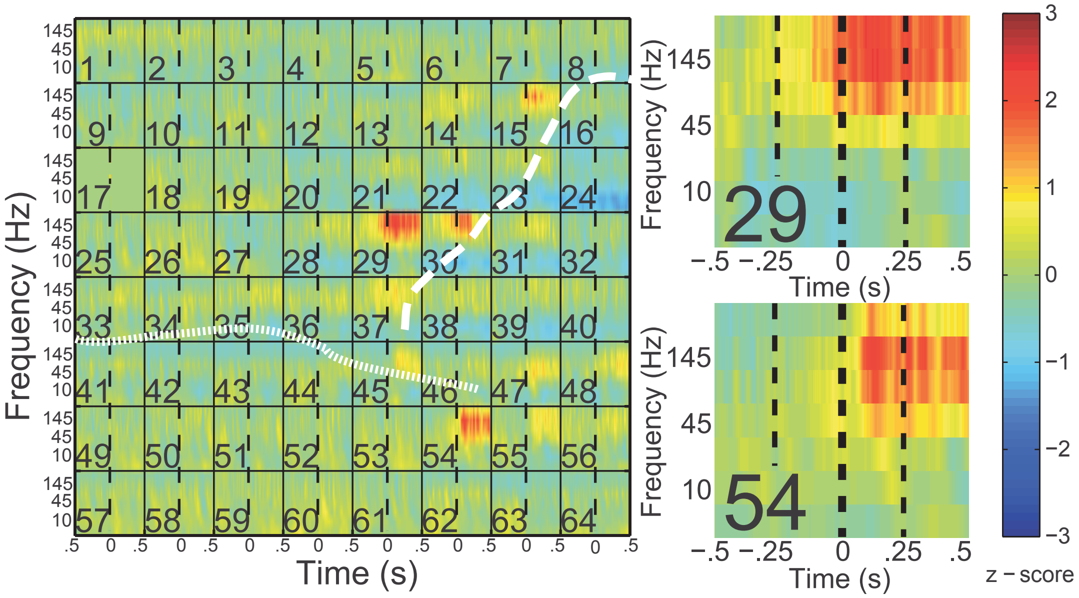


Figure 4.1. Average event-related spectrogram. Plots generated in real time are shown depicting average event-related spectrograms for Subject A. Each electrode is represented as a square on the 8x8 grid, and has its own separate time-frequency axes. The horizontal axis represents the progression of time, with the dotted line representing the onset of the event. Averaged data from half a second before the event, to half a second after the event is displayed. The 7 frequency bands are shown on the vertical axis (4-7, 8-12, 13-30, 31-59, 61-110, 111-179, 181-260 Hz). The central sulcus and Sylvian fissure are outlined with dashed and dotted lines, respectively. The average plot was generated after the subject produced the syllable /la/ for approximately 1 minute. Spectrograms for specific channels are plotted on the right

In Figure 4.2, the utility of real-time spectral mapping of sensory and motor organization for the hand is shown. Subject E was implanted with a 256 channel array over the right peri-Rolandic cortex (Fig. 4.2A). The average plot was generated in real-time while subject E was pressing a button with his left thumb (Fig. 4.2B). Electrodes along the central sulcus can be rapidly visually identified to have an event-related increase in the high gamma power (61-260 Hz), and an event-related decrease in alpha power (4-12 Hz), consistent with previous findings (Crone et al., 1998a; Crone et al., 1998b). The anterior ventral-most electrodes showed high gamma activation onset nearly 200 milliseconds prior to the button press, consistent with the precentral gyrus cortical motor processing specifically for the thumb which is the most ventral representation in the homunculus. In contrast, the posterior electrodes showed somatosensory activation after the button press. Using median nerve stimulation, the typical N20-P20 phase reversal in the averaged evoked potentials is observed along the central sulcus (Fig. 4.2C), and confirmed the localization mapping from real time spectral mapping. Note that median nerve stimulation is not exactly identical to thumb movement alone in the button press, which may explain some of the minor differences in localization between the plots.

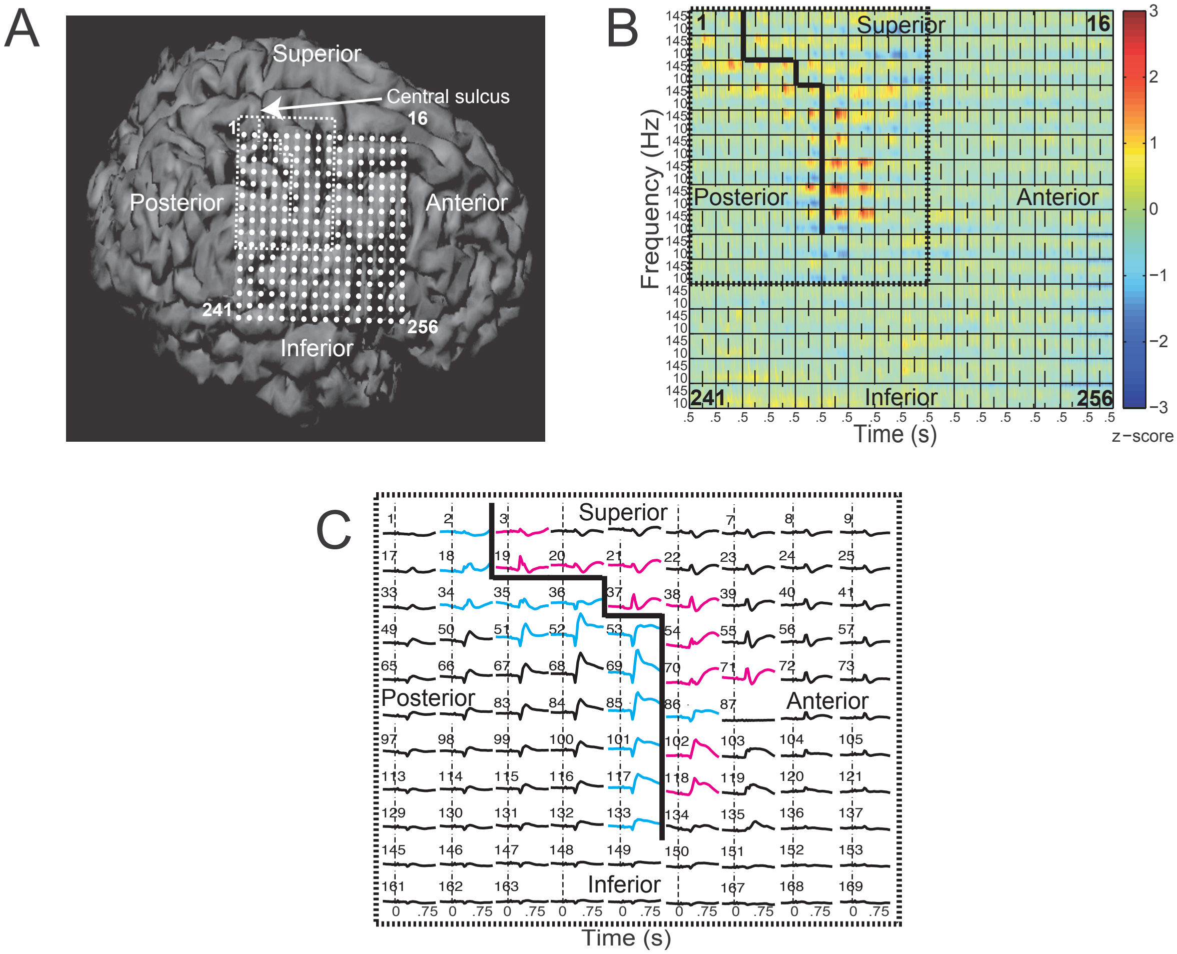


Figure 4.2. ECoG grid superimposed on patient brain, average event-related spectrogram, and median nerve stimulation results. (A) A 256-channel subdural ECoG electrode array implanted over Subject E's right lateral hemisphere is shown. Electrodes are superimposed on the 3D MR surface reconstruction image. The boundary of the N20-P20 phase reversal is outlined. (B) Plots generated in real-time are shown depicting average event-related spectrograms during the button press task (Task 3). Each electrode is represented as a square on the 16x16 grid, and has its own separate time-frequency axes. The horizontal axis represents the progression of time, with the dotted line representing the onset of the event. Averaged data from half a second before the event, to half a second after the event is displayed. The 7 frequency bands are shown on the vertical axis (4-7, 8-12, 13-30, 31-59, 61-110, 111-179, 181-260 Hz). The boundary of the N20-P20 phase reversal is outlined. (C) Somatosensory evoked potentials gathered during median nerve stimulation are shown. Onset of stimulation is indicated with a dotted line at time 0 seconds. A phase reversal of the N20-P20 peak between a pair of electrodes indicate that the pair straddles the central sulcus. Blue potentials denote those that displayed a N20 peak during median nerve stimulation. Red

potentials indicate a P20 peak was seen during stimulation. The boundary of the N20-P20 phase reversal is outlined.

Spatial layout of electrodes over the MRI surface reconstruction at different time epochs: In order to provide an intuitive visualization which readily relates activity to anatomical regions, the magnitude of the running average event-related high gamma (61-260 Hz) activations at four different time intervals (-500 to -250 ms, -250 to 0 ms, 0 to 250 ms, 250 to 500ms) was projected on to the subject's 3D MRI surface reconstruction image, and updated after every trial. Plots from the same speech production task (/la/) are shown in Figure 4.3. It can be easily seen which electrodes are active immediately before and after the event onset, and where they are located on the subject's individual brain. The four separate time epochs allow one to distinguish the pre-event motor from post-event sensory cortical activations.

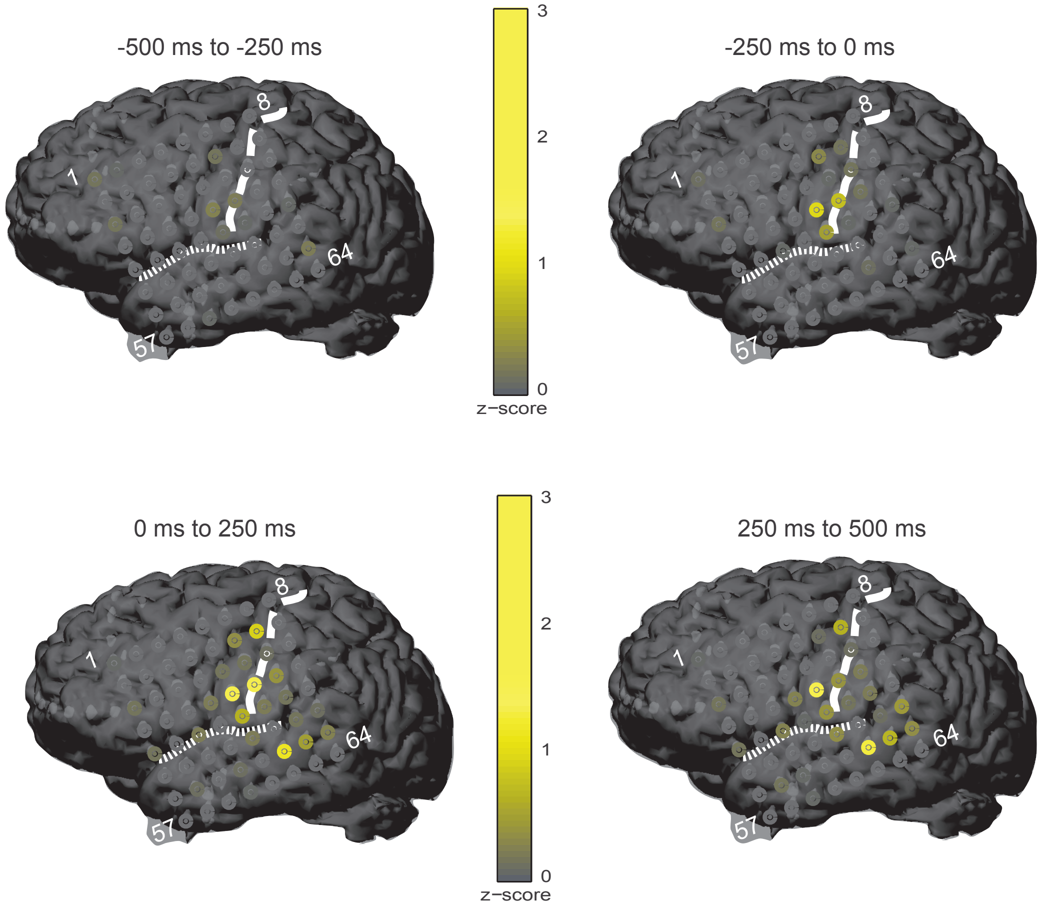


Figure 4.3. Average activity at different time epochs on MRI surface reconstruction. Plots generated in real-time of average event-related high gamma (61-260 Hz) activations on Subject A's 3D MRI surface reconstruction image at four different time intervals (-500 to -250 ms, -250 to 0 ms, 0 to 250 ms, 250 to 500ms) during production of the syllable /la/ are shown. The central sulcus and Sylvian fissure are outlined with dashed and dotted lines, respectively.

Single-trial raster plot: The raster plot displays cumulative single trials spectral changes for a given frequency range (111-179 Hz used for this example) measured in z-score deviations from the baseline. The stacked single trial plot is generated simultaneously with the same speech production task (/la/), and is shown in Figure 4.4. For each new event the plot is updated in real-time with the single event trial added as a horizontal row.

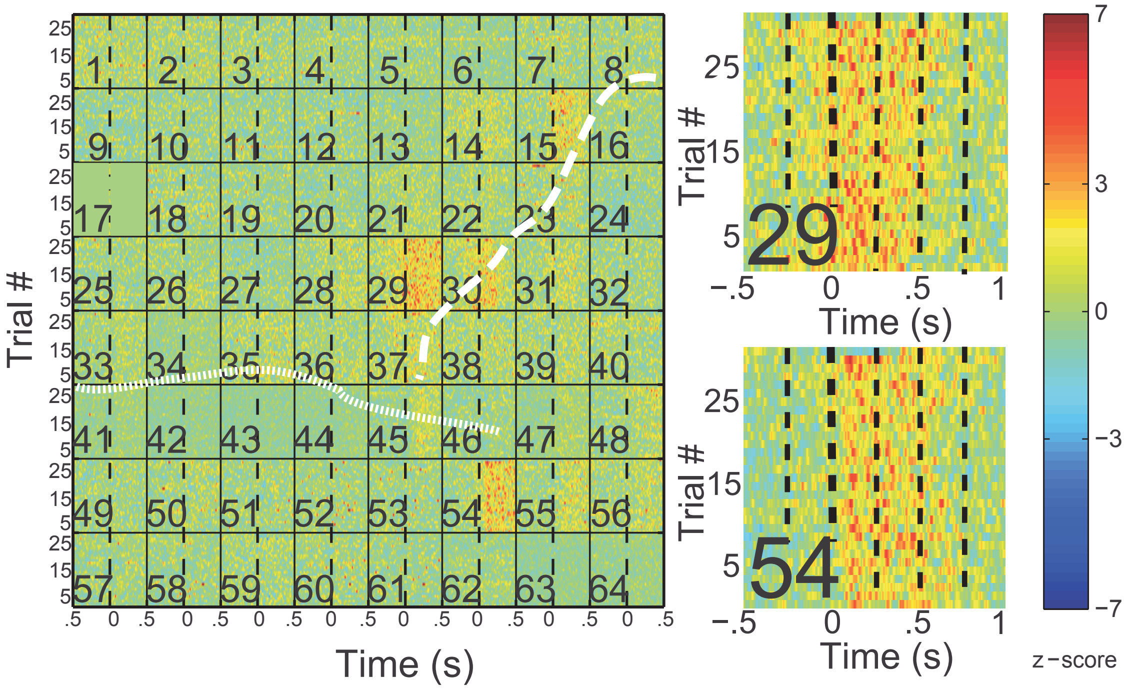


Figure 4.4. Single-trial raster plots. Plots generated in real-time are shown depicting single-trial raster plots for Subject A's 8x8 electrode grid (figure 1a). Each electrode has its own separate time axes. The horizontal axis represents the progression of time, with the dotted line representing the onset of the event. Data from half a second before the event, to half a second after the event are displayed. On the vertical axis are each event's activations in one frequency band (111-179 Hz). Z-score deviations from baseline are in colorscale. This raster plot was generated after the subject produced the syllable /la/ for approximately 1 minute. The central sulcus and Sylvian fissure are outlined with dashed and dotted lines, respectively. Rasters for specific channels are plotted on the right. The time axis has been extended to show a full second after the event.

The patterns observed across events demonstrate the robustness of the physiologic approach in single trials, and provides a visualization that is not susceptible to outlier values (in contrast to the average plots). Key channels from the raster plot can be more carefully examined on the right. Note the time axis has been extended to show a full second after the event. Again, these plots demonstrate that activity occurs earlier in the motor electrodes compared to the sensory auditory channels.

Analysis of the single plot and continual plot is included in the supplementary material (Fig. S4.2).

4.2.2 Robustness of real-time mapping

To examine the robustness of our algorithm, the variance of running average spectrograms was analyzed offline for subjects with ECS mapping results. All N trials within an experiment were randomized and binned to contain n number of trials in each bin (n ranging from 2 through N/2). An average spectrogram was then created from each bin of trials and a mean and variance between these spectrograms were calculated. It is important to note that no one trial is used in more than one average spectrogram, thus no bias is created. This was repeated 100 times and the average variance was calculated as a function of number of trials (Fig. 4.5A). A sharp exponential decay exists across all subjects. The variance between average spectrograms becomes minimal after incorporating approximately 5 trials in the average, and virtually negligible after 10 events. This suggests that a very stable representation of cortical activity is achieved with very few repetitions.

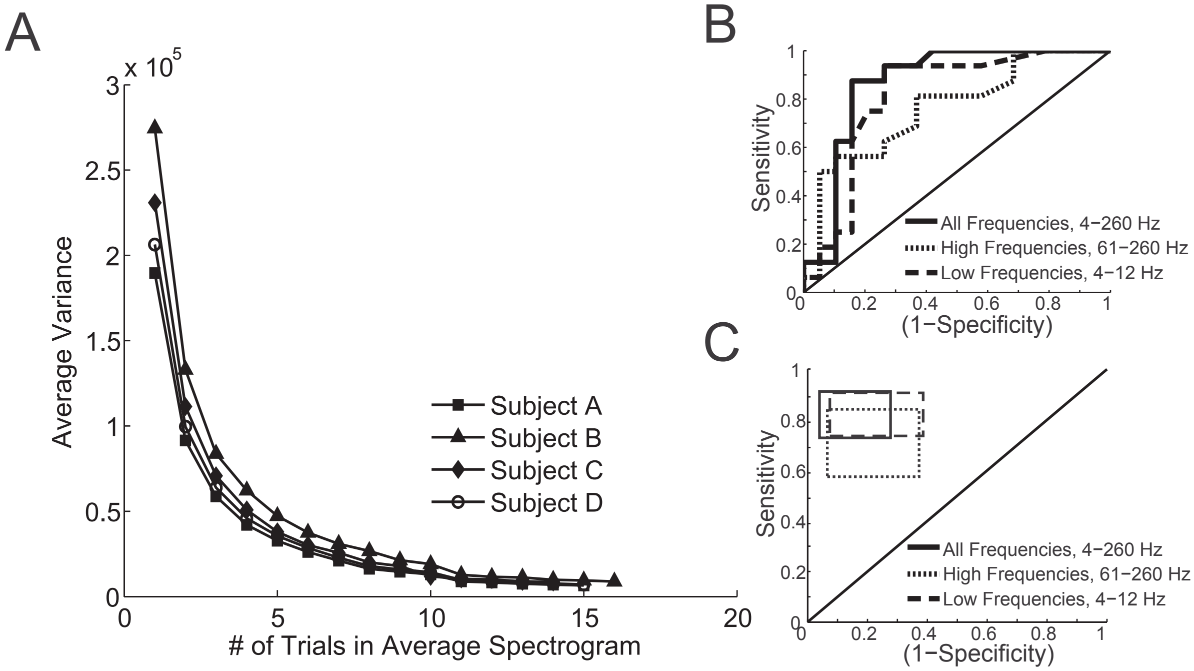


Figure 4.5. Average variance and sensitivities and specificities for subjects. (A) The average variance of average plots is shown as a function of the number of trials incorporated into the average. (B) Subject C's ROC curve. Electrode channels were categorized as positive or negative using average squared z-scores and compared against ECS maps. Z-scores from the average spectrogram after the 10th trial were used. Three spectral ranges were examined (4-260 Hz, 61-260 Hz, 4-12 Hz). The ROC curve plots the rate of true positives (sensitivity) against the false positive rate (1-specificity). The diagonal line (line of no-discrimination) divides the ROC space between good and poor classification results. Values in the upmost left corner indicate perfect classification. (C) Using all four subjects, sensitivities and specificities ranges (95% confidence intervals) were found using subject-specific optimal thresholds and plotted in the ROC space.

4.2.2 Sensitivity and specificity

The sensitivity and specificity of our mapping algorithm were analyzed offline for subjects A-D at overlapping ECS sites. Sites below the Sylvian fissure were discarded since our tasks and ECS results targeted primarily the motor and somatosensory cortex that are usually mapped with ECS. Three sets of data were used: the average spectrogram z-scores after the first 10 trials for (i) all frequencies (4-260 Hz) (ii) the low frequencies (4-12 Hz), and (iii) the high frequencies (61-260 Hz). Sites were deemed positive (+) if the average squared z-score exceeded some threshold ranging from 0 to 2 with 0.01 increments, and deemed negative (-) otherwise. ECS sites were categorized by neurologists as (+) if stimulation produced unusual or involuntary sensations or movements during clinical motor mapping, and (-) otherwise. Though adjacent electrodes were stimulated together, each site was counted separately so that the two mapping techniques could be compared.

A receiver operating characteristic (ROC) curve and the area-under-the-ROC-curve (AUC) were calculated using ECS results as the correct classification. Subject-specific ROCs are shown in Figure 4.5B and in the supplementary material (Fig. S4.3). Averaging across patients, AUC confidence intervals were found for the three sets of data. Using all frequencies, the AUC was 0.85 ± 0.083 (95% confidence interval). High frequency data yielded an AUC range of 0.73 ± 0.11 (95% confidence interval). Using low frequencies, the AUC range was 0.79 ± 0.10 (95% confidence interval). To verify that analyses were not biased from using only one baseline per day of experimentation, data was reanalyzed offline by renormalizing the data from each experimental session with it’s own baseline. These results were not significantly different, and suggest that only one baseline collection is necessary per day of experimentation

The optimal threshold – defined as the threshold that yielded the sensitivity and specificity coordinate pair closest to 100% classification accuracy on the ROC curve – varied across subjects. With high frequency data, and using the subject-specific optimal threshold, the sensitivity was 70.8 ± 13.4% (95% confidence interval) relative to ECS (+) sites. The specificity was 78.1 ± 15.3% (95% confidence interval) relative to ECS (-) sites. For the low frequencies, the sensitivity was 82.1 ± 8.5% (95% confidence interval) and the specificity was 77.0 ± 15.6%

(95% confidence interval). Using all frequencies, the sensitivity was 82.0 ± 9.2% (95% confidence interval) and the specificity was 84.2 ± 11.9% (95% confidence interval). These intervals are depicted in Figure 4.5C. Again, data using baselines from the same experimental session were not significantly different.

4.3 Discussion

This paper describes a novel technique for real-time signal analysis that can safely localize physiologic cortical function within seconds. Novel instantaneous visualization of average and single-trial activity with clear spatial and temporal resolution allow for quick visual assessment of cortical function.

Of late, there have been several demonstrations of real-time brain mapping (Brunner et al., 2009; Lachaux et al., 2007; Miller et al., 2007; Roland et al., 2010; Schalk et al., 2008). In particular, Schalk et al., sought to identify sensorimotor cortex using a modified Competitive Expectation Maximization statistical method called SIGFRIED. Their powerful method was able to provide a visual interface for the localization of isolated electrodes covering pertinent cortical regions. Our results extend these efforts to provide critical temporal resolution that is needed to differentiate temporal phases of activity, including those between sensory and motor cortex – an essential attribute to any clinically-performed functional mapping. As Figures 4.1-4.4 demonstrate, our visualizations are capable of highlighting temporal and spectral differences between electrodes that can give clinical investigators insight to such functional differences without the need of a priori thresholds.

Comparing our algorithm's results to those obtained by ECS, we conclude that we can reasonably classify positive and negative electrodes in the low and high frequency domains, and can classify remarkably well when all frequencies are used. In addition, we see an exponential decay in variance between running average spectrograms. This suggests that very few repetitions (5-10 event trials) are needed to obtain a stable functional map. Indeed, the single trial raster plots demonstrate that much of the evoked neural activity can be observed on a per trial basis. This helps to interpret potential confounding in the average plots generated from outlier data. With a more extensive battery of tests, we believe this method is capable of mapping critical eloquent cortex (e.g. expressive languages, sensory, motor) within minutes, with minimal effort required from the patient and no morbidity.

A few plausible explanations exist to explain the discrepancies between our algorithm's and ECS maps. First, it is difficult to compare methodologies due to the inherent differences in approaches. For example, due to the all-or-none nature of ECS responses, it is necessary for our comparison to also analyze ECoG data using a simple positive-or-negative threshold (Sinai et al., 2005). However, inter-subject variation in ECoG signal and behavior make it difficult to define an absolute threshold. In addition, clinical stimulation procedures require pairs of electrodes to be stimulated together. Bipolar stimulation is thought to be more spatially specific as the current is localized between the two electrodes, whereas unipolar stimulation is thought to be less specific as current can spread unpredictably to the corticospinal tract. In this study, we count electrodes of a (+) stimulation pair as two separate (+) sites, which may over-count the number of true positive sites. Furthermore, ECS evokes non-physiologic stereotyped movements. This is fundamentally different from our method, which is designed to capture cortical activity that occurs during normal behavioral movement.

Another explanation may be that ECS itself can be erroneous. ECS-induced after-discharges often spread to cortex outside the current field, thereby overestimating the extent of functional cortex (Blume et al., 2004; Lesser et al., 1984; Luders et al., 1991; Motamedi et al., 2002; Nii et al., 1996). Thresholds for ECS mapping can be extremely variable between individuals, and in some cases, no motor sites can be detected despite exhaustive testing. Thus, it is possible that ECS is less specific than previously assumed, and perhaps does not replace the ultimate gold-standard, which is patient neurological outcome.

Limitations in our own methods may also account for discrepancies. One such limitation is that our method relies on collecting a baseline rest period, and re-referencing to the common average. Outliers (e.g. seizure activity, movements, poor signal to noise ratio) could affect results by skewing baseline statistics or the common average used to calculate relative change during event-related activity. For example, contamination of the baseline could lead to more variable signal, making it more difficult to detect significant event-related spectral alterations using a z-score, which is sensitive to the variance of the baseline. Thus, contaminated signal could potentially lead to a higher number of false negatives and lower sensitivity rates. Since using different baselines did not significantly change our results, it seems unlikely that our results were adversely affected by noisy baselines. However, this limitation highlights the importance of monitoring for outlier activity during experimentation. We currently visually reject electrodes with poor signal quality. Though susceptible to human error, this is common practice among clinicians and researchers. Future studies can evaluate the automation of this process using the known biophysical properties of ECoG recordings to reject artifact-laden signals (Miller et al., 2009). In addition, we plan to explore methods of isolating motor elements from cortical maps that do not require baseline statistics, such as subtraction of sensory activations.

Additionally, the current study performed only four tasks that were designed to highlight sensorimotor cortical regions during speech production and hand movement. However, motor stimulation mapping routines are tailored to specific individuals, which lead to a variable number of stimulation sites and differing responses between patients. For example, subject A had a limited number of sites tested during stimulation, which could account for the high sensitivity and specificity rates (Fig. S4.3A). Stimulation for some patients also elicited wrist movement, and throat movement. Our tasks did not require strong wrist or throat movements, and thus our algorithm could not have been expected to detect these (+) sites.

The determination of a diagnostic "threshold" does have significant implications for the long-term clinical application of our approach. While it offers a novel method for the determination of local cortical processing, our algorithm needs validation across many patients across institutions to carefully assess a threshold that maintains high sensitivity and specificity.

For these reasons, we have strongly advocated direct representation of the data analysis (e.g. illustrating complete spectrograms for each electrode, both single trial and average, in addition to the high gamma plots alone) since a simple threshold derived "yes/no" response may be misleading. For now, this method will require users to have some familiarity with signal processing involved to make safe decisions about its clinical application.

The ethical considerations for adoption of this mapping tool are similar to those other tools used for preoperative or pre-resection mapping, such as functional magnetic resonance imaging or magnetic source imaging. That is, there needs to be clear benefit to the patient from this new technology. To achieve this, future studies must prospectively validate that it is more effective or equivalent than the current standard of ECS mapping. Our paper is first step in describing the technology. At this point, we use these findings only to supplement stimulation mapping.

The speed, accuracy, and safety of this algorithm make it a promising candidate for future applications in both clinical and basic research, but there are several ways in which our mapping method can be further developed. To improve upon the clinical utility of this system, subsequent developments will examine different user interfaces that might be useful to clinicians, such as anatomically derived coordinate systems and easy-to-use graphical user interfaces.

4.4 Methods

4.4.1 Subjects

We recorded ECoG in 5 refractory epilepsy patients (see Table 4.1) undergoing intracranial monitoring for the localization of an epileptogenic focus. Subjects underwent a craniotomy for chronic (1-2 weeks) implantation of a subdural platinum–iridium electrode array over either the left or right hemisphere. The study protocol, approved by the UC San Francisco Committee on Human Research, presented minimal risk to participating subjects and did not interfere with the clinical recordings. Placement of the array was determined entirely by clinical needs and varied between subjects. All participants provided informed consent.

Table 4.1. Patient characteristics

Subject	Age	Gender	Hemispheric coverage	Age of seizure onset	Resection location	Engel seizure outcome classification	Hemispheric dominance for language
A	23	Male	Left	3	Temporal lobe	I	Left
B	36	Female	Left	18	Inferior parietal lobe	II	Left
C	48	Male	Left	2	Anterior temporal lobe	I	Left
D	45	Male	Left	24	Posterior temporal cortex	I	Left
E	30	Male	Right	7	Inferior parietal cortex	III	Left

For the purpose of this paper, the majority of the subject-specific figures were generated from a representative example (Subject A). Using proprietary anatomic image fusion software from BrainLab (Munich, Germany), electrode positions were extracted by computed tomography (CT) scan, co-registered with the patient's MRI, and then superimposed on the subject's 3D MRI surface reconstruction image (Fig. 4.6A). Registrations were verified with intraoperative photographs, and by another third party open-source imaging software, Osirix.

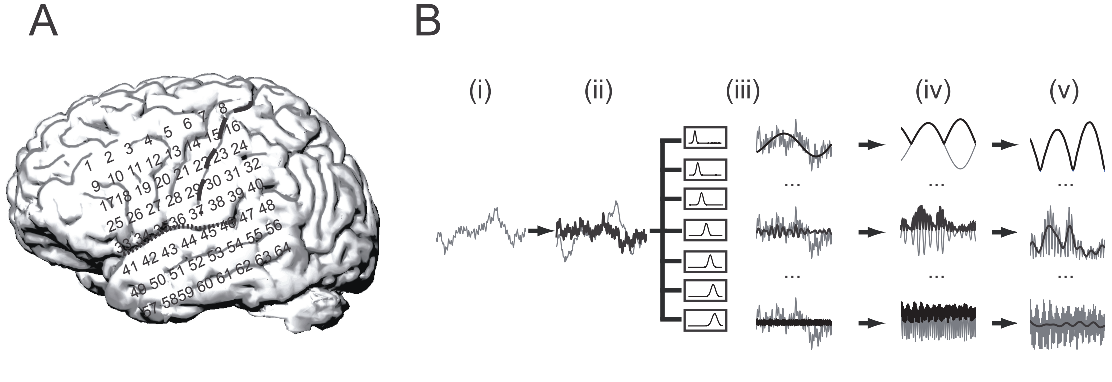


Figure 4.6. ECoG grid superimposed on patient brain and real-time signal processing algorithm. (A) A 64-channel subdural ECoG electrode array implanted over Subject A's left lateral hemisphere is shown. Electrode numbers are superimposed on the 3D MR surface reconstruction image. The central sulcus and Sylvian fissure are outlined with dashed and dotted lines, respectively. (B) The real-time signal processing algorithm is depicted. (i) An example of raw ECoG signal from a single channel is shown. (ii) Line noise is eliminated and the signal is re-referenced to the common average signal. (iii) The signal is separated by a filter bank (4-7 Hz, 8-12 Hz, 13-30 Hz, 31-59 Hz, 61-110 Hz, 111-179 Hz, 181-260 Hz). (iv) The absolute value of the time series is taken. (v) The approximate amplitude envelope of the signal is extracted by low-pass filtering (cutoff frequencies: 5 Hz, 5 Hz, 20 Hz, 20 Hz, 40 Hz, 40 Hz, 40 Hz). The resulting signal after the stepwise process has been performed is shown in black. The preceding signal before the processing step is shown in grey.

4.4.2 Experimental set up

For the majority of the subjects (A-D), the subdural ECoG grids implanted were standard 64-channel platinum-iridium electrodes with 10 mm center-to-center spacing arranged in an 8x8 configuration (Ad-Tech, Racine, WI, USA). Each electrode had an exposed diameter of 2 mm. Subject E received a high-density 256-channel grid with 4 mm center-to-center spacing over the right hemisphere. Each electrode had exposed diameter of 1.25 mm.

ECoG signals were split between the clinical monitoring system and a customized research data acquisition and processing system. The ground and reference signal were split from a scalp electrode, usually on the patient's forehead. A 30 second baseline rest period dataset was collected. During this baseline period, the room was quieted and subjects were instructed to simply rest with eyes open without moving. Generally only one baseline collection was needed per day of experimentation, however, multiple were often collected for post-hoc comparisons, or if there was a long interval between data collection periods.

Experimental sessions were divided into blocks, each lasting1 to 2 minutes. There were four simple tasks designed to illustrate the algorithm: 1) Speaking, 2) Listening, 3) Hand button press, 4) Tactile hand somatosensation. Each block consisted of 15 - 40 repeated trials of the same task.

The speaking task involved self-paced production of /ba/ and /la/ syllables. A listening task involved simple passive listening to the experimenter's repeated production of speech syllables. The third task required the subject to press a handheld button press device with the thumb of the contralateral hand. Subjects were instructed to press the button when a cue was presented. In the fourth task, to isolate the somatosensory response, the experimenter briefly applied the same button press device against the subject's finger while the subject was resting.

4.4.3 Real-time mapping

To rapidly identify spatial and temporal patterns in ECoG signal during functional cortical activation, the following algorithm was implemented.

1) Signal Processing: First, ECoG signal was acquired in real-time using a portable customized multichannel neurophysiology workstation (Tucker-Davis Technologies, TDT, Alachua, FL, USA). All channel signals were amplified independently. The data was sampled at 500 Hz and recorded in a circular buffer for real-time analysis.

2) Real-time signal processing: Real-time signal processing was carried out on the portable workstation. To ensure that signals analyzed contained no ambient electrical line noise, 60 Hz and its harmonics were removed using a second order Butterworth notch filter with a 5 Hz stopband. The common average reference was then removed from each channel by subtracting the average of the raw signal across all electrodes. Electrode channels were omitted if they, upon visual inspection, had poor signal quality due to electrode drift, poor electrode contact, or excessive high frequency noise. Following the re-referencing, the signal was band pass filtered at seven different frequency bands (4-7 Hz, 8-12 Hz, 13-30 Hz, 31-59 Hz, 61-110 Hz, 111-179 Hz, 181-260 Hz) using a second order Butterworth filter. The resulting bandpass signals can be viewed as carrier signals whose amplitudes are modulated by slower periodic signals.

The slower periodic signal can be extracted by approximating the envelope of the bandpass signal. This was done by low pass filtering the absolute value of the resulting signal. Since the cutoff frequency of a modulating signal is always at most half the bandwidth of the bandpass signal, frequency cutoffs were set at 5 Hz, 20 Hz, and 40 Hz for signals band passed filtered between 4-12 Hz, 13-59 Hz, and 61-260 Hz, respectively. A schematic diagram of the preprocessing stream can be seen in Figure 4.6B.

3) Event Detection: A basic event detection method was implemented in order to determine the timing of cortical events related to the execution of a particular task. Both speaking and listening events were recorded using a microphone. For the hand movement and somatosensory tasks, a simple button press device was used. Both analog signals were recorded synchronously with the multichannel ECoG data. The signal was converted to a digital signal and downsampled to 500 Hz and smoothed using an exponential smoothing average with a factor of 0.005. Threshold voltages were predetermined for the outputs of the microphone at 80 dB sound pressure level (SPL) and the button press device. Event onset was defined to be the time when analog voltage exceeded the pre-defined threshold voltage. To ensure that inaccuracy in event detection did not skew any results, any experimental session where false events (e.g. coughing) were registered were discarded.

4.4.4 Visualization

To allow for rapid visual identification of spatial and temporal patterns, MATLAB (Mathworks, Natick, MA, USA) was used for visualization. The envelope of the ECoG signal and event signal were used as inputs to our customized program.

To examine the event-related change in ECoG activity, the neural signal was normalized with respect to the baseline rest signal. The log power of the signal was first calculated by taking the logarithm of the incoming envelope signal. To avoid taking the logarithm of 0, which is undefined, the median envelope signals during the baseline period were added as constants prior to taking the logarithm. The log power was then z-score normalized using the mean and standard deviation of the baseline log power signal.

During the task, five plots of the entire electrode grid were continuously refreshed. This means the software was set to read and analyze newly acquired data immediately after the previous cycle was completed. Each cycle took on average 0.7 seconds to complete, with each plot taking approximately 0.10 - 0.15 seconds to calculate and display. Listed below are names and descriptions of the plots.

1) Average plot: Displays the running average event-related spectrogram. It is calculated by summing all event-related spectrograms and dividing by the number of event trials. Since it is a running average, the spectrogram is recalculated and plotted with each new event.
2) Average plot on MR: Displays running average event-related activations on the subject's 3D MRI surface reconstruction image.
3) Single-trial raster plot: Displays single trial activations in a stacked raster plot. Power from only one frequency band is shown
4) Single plot: Displays the event-related spectrogram of the most recent single event trial.
5) Continual plot: Displays the mean z-score of the last 20 ms of data from a specified frequency.

All event windows were taken from 500 ms prior to the onset of the event, and 500 ms after the onset. The z-score measurements were presented using a normalized color scale.

To remove common artifact data, we implemented an algorithm for removing trials from occasionally corrupted noisy data caused by faulty electrodes in the average plot and single-trial raster. A window of data for a particular frequency band was deemed an artifact if more than 50% of the event-window was below 1.5 baseline standard deviations or if more than 80% was above 1.5 baseline standard deviations. These data were displayed with a solid band of green on the plot itself.

4.4.5 Electrical cortical stimulation

ECS mapping was performed on subjects with the 64-channel grid (subjects A-D) according to clinical motor mapping routine in 2-3 hour sessions over 1-2 days. These procedures utilized constant current bipolar electrical stimulation between pairs of adjacent electrodes using a Grass S-88 cortical stimulator (Grass-Telefactor, West Warwick, RI, USA). Trains of 1-5 seconds, 50 Hz, 0.3 ms, alternating polarity square-wave pulses were sent through the stimulator starting with an intensity of 1 mA. The intensity was increased at 1 mA increments, up to a maximum of 15 mA. Intensities were adjusted so that after-discharges were not produced. Patients reported any unusual or involuntary sensations or movements. Disruption of motor function was detected by observing patients during voluntary movements. It should be noted that not all electrode sites were stimulated. Sites suitable for stimulation were determined entirely by clinical neurologists and were dependent on the patient and probable resection and/or eloquent cortical areas.

4.4.6 Median nerve stimulation

Due to the high-density electrode configuration of subject E, standard ECS mapping was not carried out. However, standard localization of the central sulcus using the phase reversal of somatosensory evoked potential (N20-P20) was performed (Dinner et al., 1987). The contralateral median nerve was stimulated at the wrist with 5.1 pulses/second and a current intensity between 5 mA and 10 mA until twitches of the thumb were observed visually. The anode was placed just proximal to the palmar crease, and the cathode was placed between the tendons of the palmaris longus muscle, 3 cm proximal to the anode. ECoG signals were recorded simultaneously. The raw times series was averaged over 187 trials.

4.5 Supplementary materials

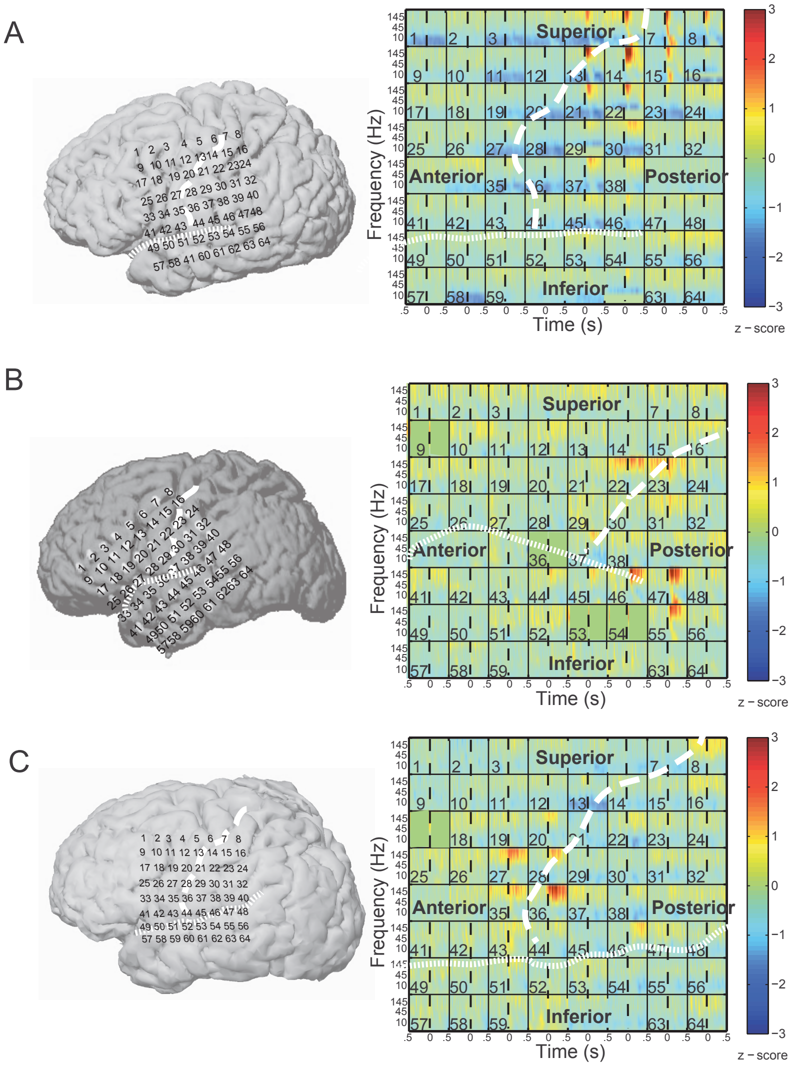


Figure S4.1. Average event-related spectrograms. Plots are shown depicting average event-related spectrograms for Subjects B, C, and D. Plotting conventions are the same as figure 2 in the main text. (A) The average plot for Subject B was generated after the experimenter pressed the button press device against the subject's finger while the subject was resting for approximately 1 minute (Task 4). (B) The average plot for Subject C was generated after the subject listened to the syllable /ba/ for approximately 1 minute (Task 2). (C) The average plot for Subject D was generated after the subject produced the syllable /ba/ for approximately 1 minute (Task 1).

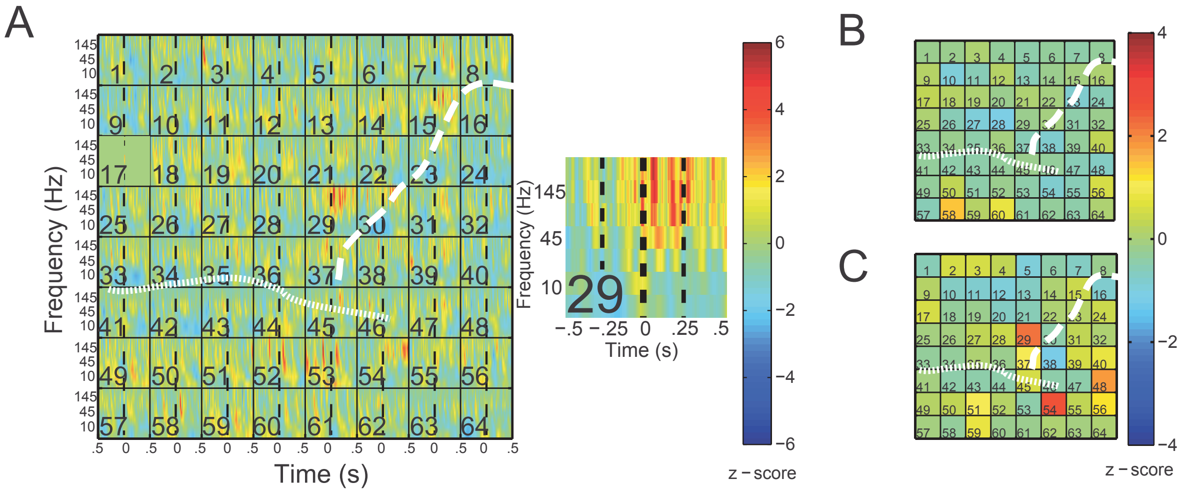


Figure S4.2. Single-trial event-related spectrograms and continual plots. (A) Plots are shown depicting event-related spectrograms for one event during Subject A's /la/ task. On the left, each electrode is represented as a square on the 8 x 8 grid with the same time-frequency axis used in the average plot (Fig. 4.1). On the right is a spectrogram for channel 29. High event-locked z-scores can be seen, especially in channel 29. In offline analyses the maximum z-score in channel 29's single-event spectrogram was found to be significant using a Bonferroni-corrected alpha level of 0.05 ($p < 1.9 \times 10^{-4}$, 181-260 Hz, t = 44 ms). While this significant activation in channel 29 corresponds well with observations made with the average and single-trial raster plots, a broad range of activations throughout the grid (yellow-red coloring) makes it difficult to visually identify other crucial channels with the single-event spectrograms. (B, C) Plots are shown depicting average z-scores of 20 ms of data during Subject A's /la/ task in the 111-179 Hz spectral range. Each electrode is represented as a square on the 8 x 8 grid. Shown are average z-scores captured during a period of rest (B) and during the production of syllable /la/ (C). During rest nearly all electrodes to have insignificant p-values (p>0.87, Bonferroni-corrected). During speech production, channels 29 and 54 show a high change from baseline. These channels correspond to those identified in the average plots (Fig. 4.1). However, channels 15 and 30, seen activated in the average plot, do not show similar changes. The central sulcus and Sylvian fissure are outlined with dashed and dotted lines, respectively.

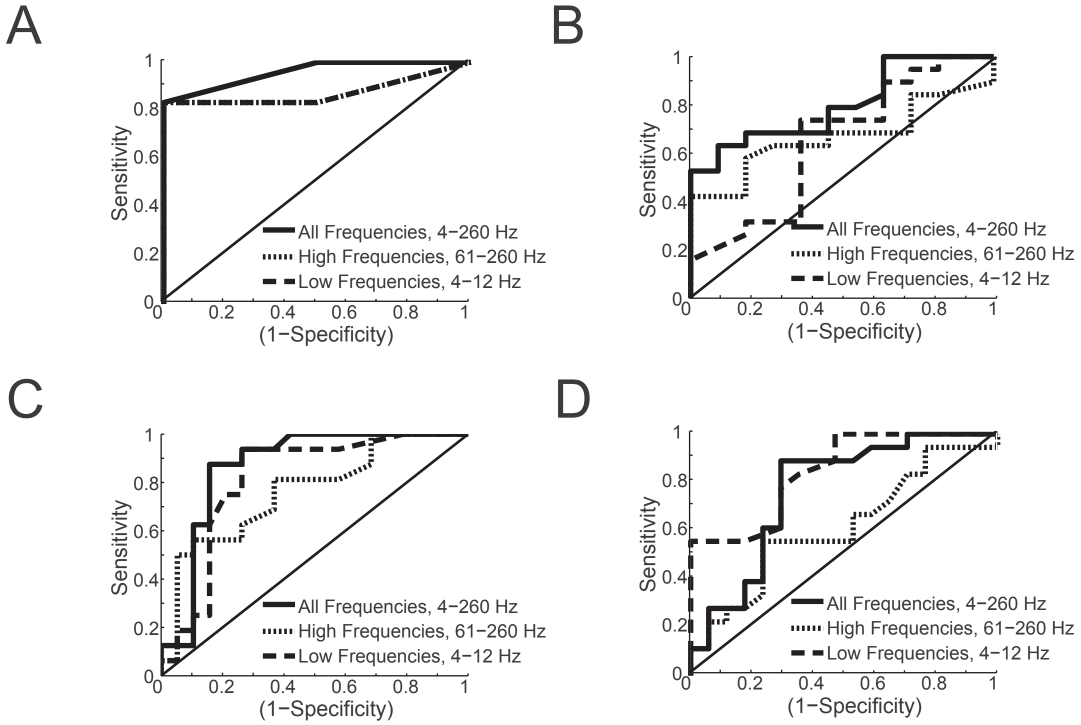


Figure S4.3. Sensitivities and specificities for subjects. ROC plots for Subjects A, B, C, and D. Plotting conventions are the same as Figure 4.5B. Figure 4.5B has been repeated here for ease of comparison.

CHAPTER 5

Conclusion

This dissertation sought to elucidate how the human cortex processes speech sounds, and presents a physiologic framework for validating linguistic theories of speech perception. Additionally, it outlines an innovative clinical brain mapping system that aids in the preservation of eloquent cortex, and was built upon the principles of human electrophysiology. Together, this work details the human speech processing system and the benefits of physiologic design for the creation of clinical tools. Below, we summarize the key findings of this dissertation, and discuss open questions in our understanding of the human speech system and future directions for biomedical engineering.

5.1 Summary of contributions

In Chapter 2, direct cortical recordings of human patients chronically implanted with subdural grids revealed the STG representation of the entire American English phonetic inventory. Spatially local selectivity to acoustic-phonetic features was found to give rise to our internal representation of a phoneme, and could account for common human perceptual inaccuracies. These findings are consistent with auditory perceptual models that posit a feature hierarchy organized along acoustic cues (Stevens, 2002).

In Chapter 3, the role of the motor system in perception was explored in the context of speech. Neural activity was localized to the ventral-most and dorsal-most regions in motor cortex during listening, whereas it was distributed throughout motor cortex during speaking. We found that the motor cortex contained focal sites with defined spectrotemporal tuning properties to acoustic features, such as voicing. These results are in direct contrast to predictions made by the motor theory of speech perception. Furthermore, they provide evidence for an acoustic sensory representation of speech sounds in motor cortex, and suggest the existence of a human motor cortex site specialized for sensorimotor speech integration.

In Chapter 4, a clinical brain mapping system was designed to safely carry out cortical resections near eloquent brain regions. The traditional method for brain mapping is electrical cortical stimulation, which causes dangerous seizures and uses hours of valuable medical resources. With insights gathered from our previous studies, we designed a physiologic algorithm to map eloquent brain regions safely, accurately, and in a fraction of the original time. This system promises to revolutionize the current standard of care in neurosurgical operations.

5.2 Open questions and future directions

This work provides unprecedented insight into how the human cortex represents speech sounds, but there are still many open questions to be answered and new directions to be taken. First, our results demonstrate the human STG is selective to distinct acoustic-phonetic features of speech. Such a representation of speech suggests a universal basis for languages, and provides a fundamental description of speech in the brain. However, to achieve a more complete understanding of speech, future studies should explore the encoding of high-order speech features such as prosody and semantics.

Furthermore, our findings suggest the existence of a sensorimotor laryngeal motor region sensitive to both the production and perception of speech sounds. This region has been suggested to be a unique feature of vocalization in humans, as it is not found in nonhuman primates (Brown et al., 2008). It remains to be seen how precisely this region participates in laryngeal control. A detailed analysis of neural responses in this region in a speech production setting may shed light on this evolutionary novelty, and answer questions about how human-specific vocalizations are controlled.

Insights gained about physiologic neural activity gathered from subdural electrode arrays led to the development of a functional mapping system to aid preoperative surgical planning. This system has the potential to serve as a replacement for the current standard of care, electrical cortical stimulation. But, novel clinical systems have major barriers to overcome to reach market. To be market approved, products such as our mapping system must undergo formal clinical trials to prove efficacy and equivalency to current standards. These clinical trials can be long and costly, which is undesirable in a risk-adverse funding environment. In order to lower the barrier to entry, further research studies should be pursued to define key design parameters to create a more robust, turn-key clinical mapping system.

References

1. Alho, J., Sato, M., et al. (2012). Enhanced early-latency electromagnetic activity in the left premotor cortex is associated with successful phonetic categorization. *Neuroimage, 60*(4), 1937-1946. doi: http://dx.doi.org/10.1016/j.neuroimage.2012.02.011

2. Basso, A., Casati, G., & Vignolo, L. (1977). Phonemic identification defect in aphasia. *Cortex, 13*(1), 85-95.

3. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

4. Binder, J. R., Frost, J. A., et al. (2000). Human Temporal Lobe Activation by Speech and Nonspeech Sounds. *Cerebral Cortex, 10*(5), 512-528. doi: 10.1093/cercor/10.5.512

5. Blume, W., Jones, D., & Pathak, P. (2004). Properties of after-discharges from cortical electrical stimulation in focal epilepsies. *Clinical Neurophysiology, 115*(4), 982-989.

6. Blumstein, S. E., Baker, E., & Goodglass, H. (1977a). Phonological factors in auditory comprehension in aphasia. *Neuropsychologia, 15*(1), 19-30.

7. Blumstein, S. E., Cooper, W. E., et al. (1977b). The perception and production of voice-onset time in aphasia. *Neuropsychologia, 15*(3), 371-383.

8. Boatman, D., Lesser, R. P., & Gordon, B. (1995). Auditory Speech Processing in the Left Temporal Lobe: An Electrical Interference Study. *Brain and Language, 51*(2), 269-290. doi: http://dx.doi.org/10.1006/brln.1995.1061

9. Bouchard, K. E., Mesgarani, N., et al. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature, 495*(7441), 327-332.

10. Broca, P. (1861). Perte de la parole, ramollissement chronique et destruction partielle du lobe anterieur gauche du cerveau. *Bull Soc Anthropol, 2*, 235-238.

11. Brown, S., Ngan, E., & Liotti, M. (2008). A larynx area in the human motor cortex. *Cerebral Cortex, 18*(4), 837-845.

12. Brunner, P., Ritaccio, A. L., et al. (2009). A practical procedure for real-time functional mapping of eloquent cortex using electrocorticographic signals in humans. *Epilepsy & Behavior, 15*(3), 278-286.

13. Callan, D., Callan, A., et al. (2010). Premotor cortex mediates perceptual performance. *Neuroimage, 51*(2), 844-858.

14. Canolty, R. T., Soltani, M., et al. (2007). Spatiotemporal dynamics of word processing in the human brain. *Frontiers in neuroscience, 1*(1), 185.

15. Chan, A. M., Dykstra, A. R., et al. (2013). Speech-Specific Tuning of Neurons in Human Superior Temporal Gyrus. *Cerebral Cortex*, bht127.

16. Chang, E. F., Edwards, E., et al. (2010a). Cortical Spatio-temporal Dynamics Underlying Phonological Target Detection in Humans. *Journal of Cognitive Neuroscience, 23*(6), 1437-1446. doi: 10.1162/jocn.2010.21466

17. Chang, E. F., Rieger, J. W., et al. (2010b). Categorical speech representation in human superior temporal gyrus. *Nat Neurosci, 13*(11), 1428-1432.

18. Chechik, G., & Nelken, I. (2012). Auditory abstraction from spectro-temporal features to coding auditory entities. *Proceedings of the National Academy of Sciences, 109*(46), 18968-18973.

19. Cheung, C., & Chang, E. F. (2012). Real-time, time-frequency mapping of event-related cortical activation. *Journal of neural engineering, 9*(4), 046018.

20. Chevillet, M. A., Jiang, X., et al. (2013). Automatic Phoneme Category Selectivity in the Dorsal Auditory Stream. *The Journal of Neuroscience, 33*(12), 5208-5215.

21. Chomsky, N., & Halle, M. (l968). *The sound pattern of English*. New York.

22. Clements, G. N. (1985). The geometry of phonological features. *Phonology yearbook*, 225-252.

23. Cogan, G. B., Thesen, T., et al. (2014). Sensory-motor transformations for speech occur bilaterally. *Nature*.

24. Creutzfeldt, O., Ojemann, G., & Lettich, E. (1989). Neuronal activity in the human lateral temporal lobe. *Experimental brain research, 77*(3), 451-475.

25. Crone, N., Boatman, D., et al. (2001a). Induced electrocorticographic gamma activity during auditory perception. Brazier Award-winning article, 2001. *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology, 112*(4), 565.

26. Crone, N., Hao, L., et al. (2001b). Electrocorticographic gamma activity during word production in spoken and sign language. *Neurology, 57*(11), 2045.

27. Crone, N., Miglioretti, D., et al. (1998a). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. *Brain, 121*(12), 2301.

28. Crone, N., Miglioretti, D., et al. (1998b). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. I. Alpha and beta event-related desynchronization. *Brain, 121*(12), 2271.

29. Damasio, H., & Damasio, A. R. (1980). The anatomical basis of conduction aphasia. *Brain, 103*(2), 337-350.

30. Delattre, P. C., Berman, A., & Cooper, F. S. (1962). Formant transitions and loci as acoustic correlates of place of articulation in American fricatives. *Studia linguistica, 16*(1,Äê2), 104-122.

31. Delattre, P. C., Liberman, A. M., & Cooper, F. S. (2005). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America, 27*(4), 769-773.

32. di Pellegrino, G., Fadiga, L., et al. (1992). Understanding motor events: a neurophysiological study. *Experimental brain research, 91*(1), 176-180.

33. Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology, 1*(2), 121-144.

34. Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annu. Rev. Psychol., 55*, 149-179.

35. Dinner, D. S., Luders, H., et al. (1987). Cortical generators of somatosensory evoked potentials to median nerve stimulation. *Neurology, 37*(7), 1141.

36. Duffau, H. (2012). The "frontal syndrome" revisited: Lessons from electrostimulation mapping studies. *Cortex, 48*(1), 120-131.

37. Edwards, E., Nagarajan, S. S., et al. (2010). Spatiotemporal imaging of cortical activation during verb generation and picture naming. *Neuroimage, 50*(1), 291-301.

38. Edwards, E., Soltani, M., et al. (2009). Comparison of time-frequency responses and the event-related potential to auditory speech stimuli in human cortex. *Journal of neurophysiology, 102*(1), 377-386.

39. Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap* (Vol. 57): CRC press.

40. Engineer, C. T., Perez, C. A., et al. (2008). Cortical activity patterns predict speech discrimination ability. *Nature Neuroscience, 11*(5), 603-608.

41. Fadiga, L., Craighero, L., et al. (2002). Short communication: Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience, 15*, 399-402.

42. Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*: MIT press.

43. Formisano, E., De Martino, F., et al. (2008). " Who" Is Saying" What"? Brain-Based Decoding of Human Voice and Speech. *Science, 322*(5903), 970-973.

44. Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*(1), 3-28.

45. Fox, P. T., & Friston, K. J. (2012). Distributed processing; distributed functions? *Neuroimage, 61*(2), 407-426.

46. Gallese, V., Fadiga, L., et al. (1996). Action recognition in the premotor cortex. *Brain, 119*(2), 593.

47. Garofolo, J. S. (1993). *TIMIT: acoustic-phonetic continuous speech corpus*: Linguistic Data Consortium.

48. Haglund, M., Berger, M., et al. (1994). Cortical localization of temporal lobe language sites in patients with gliomas. *Neurosurgery, 34*(4), 567.

49. Hatsopoulos, N. G., & Suminski, A. J. (2011). Sensing with the motor cortex. *Neuron, 72*(3), 477-487.

50. Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron, 69*(3), 407-422.

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

52. Holt, L. L., & Lotto, A. J. (2008). Speech perception within an auditory cognitive science framework. *Current Directions in Psychological Science, 17*(1), 42-46.

53. Hubel, D. H. (1995). *Eye, brain, and vision*: Scientific American Library/Scientific American Books.

54. Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification, 2*(1), 193-218.

55. Iverson, P., & Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception & Psychophysics, 62*(4), 874-886.

56. Kandel, E., Schwartz, J., et al. (2013). *Principles of Neural Science, Fifth Edition*: McGraw-Hill Education.

57. Keles, G., Lundin, D., et al. (2004). Intraoperative subcortical stimulation mapping for hemispheric perirolandic gliomas located within or adjacent to the descending motor pathways: evaluation of morbidity and assessment of functional outcome in 294 patients. *Journal of neurosurgery, 100*(3), 369-375.

58. Krauss, G. L., Fisher, R., et al. (1996). Cognitive Effects of Resecting Basal Temporal Language Areas. *Epilepsia, 37*(5), 476-483. doi: 10.1111/j.1528-1157.1996.tb00594.x

59. Lachaux, J., Jerbi, K., et al. (2007). A blueprint for real-time functional mapping via human intracranial recordings. *PLoS One, 2*(10), 1094.

60. Ladefoged, P., & Johnson, K. (2010). *A Course in Phonetics*: Cengage Learning.

61. Lesser, R., Luders, H., et al. (1984). Cortical afterdischarge and functional response thresholds: results of extraoperative testing. *Epilepsia, 25*(5), 615-621.

62. Liberman, A. M., Cooper, F. S., et al. (1967). Perception of the speech code. *Psychological review, 74*(6), 431.

63. Liberman, A. M., Delattre, P. C., et al. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied, 68*(8), 1.

64. Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1-36.

65. Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science, 243*(4890), 489.

66. Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and speech, 10*(1), 1-28.

67. Luders, H., Lesser, R., et al. (1991). Basal temporal language area. *Brain: a journal of neurology, 114*, 743.

68. Massaro, D., & Chen, T. (2008). The motor theory of speech perception revisited. *Psychonomic Bulletin & Review, 15*(2), 453-457. doi: 10.3758/pbr.15.2.453

69. Matyas, F., Sreenivasan, V., et al. (2010). Motor Control by Sensory Cortex. *Science, 330*(6008), 1240-1243. doi: 10.1126/science.1195797

70. McGuigan, F. J. (1979). *Psychophysiological measurement of covert behavior: A guide for the laboratory*: L. Erlbaum Associates.

71. Meister, I. G., Wilson, S. M., et al. (2007). The essential role of premotor cortex in speech perception. *Current Biology, 17*(19), 1692-1696.

72. Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature, 485*(7397), 233-236.

73. Mesgarani, N., Cheung, C., et al. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science, 343*(6174), 1006-1010. doi: 10.1126/science.1245994

74. Mesgarani, N., David, S. V., et al. (2008). Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America, 123*(2), 899-909.

75. Miceli, G., Gainotti, G., et al. (1980). Some aspects of phonological impairment in aphasia. *Brain and Language, 11*(1), 159-169.

76. Miller, G. A., & Nicely, P. E. (1955). An analysis of perception confusions among some english consonants. *JASA, 27*, 338-352.

77. Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America, 85*(5), 2114-2134.

78. Miller, K., denNijs, M., et al. (2007). Real-time functional brain mapping using electrocorticography. *Neuroimage, 37*(2), 504-507.

79. Miller, K. J., Sorensen, L. B., et al. (2009). Power-Law Scaling in the Brain Surface Electric Potential. *PLoS Comput Biol, 5*(12), e1000609.

80. Motamedi, G., Lesser, R., et al. (2002). Optimizing parameters for terminating cortical afterdischarges with pulse stimulation. *Epilepsia, 43*(8), 836-846.

81. Nelken, I. (2008). Processing of complex sounds in the auditory system. *Current opinion in neurobiology, 18*(4), 413-417.

82. Nii, Y., Uematsu, S., et al. (1996). Does the central sulcus divide motor and sensory functions. *Brain, 46*, 360-367.

83. Obleser, J., Leaver, A. M., et al. (2010). Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. *Frontiers in psychology, 1*.

84. Ojemann, G. (1979). Individual variability in cortical localization of language. *Journal of neurosurgery, 50*, 164-169.

85. Ojemann, G., Ojemann, J., et al. (1989). Cortical language localization in left, dominant hemisphere. An electrical stimulation mapping investigation in 117 patients. *Journal of neurosurgery, 71*(3), 316-326.

86. Penfield, W., & Jasper, H. (1954). *Epilepsy and the functional anatomy of the human brain*. Oxford, England: Little, Brown & Co.

87. Penfield, W., & Roberts, L. (1959). Speech and brain mechanisms.

88. Peterson, G. E., & Barney, H. L. (2005). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24*(2), 175-184.

89. Pfurtscheller, G., & Cooper, R. (1975). Frequency dependence of the transmission of the EEG from cortex to scalp. *Electroencephalography and clinical neurophysiology, 38*(1), 93-96.

90. Pfurtscheller, G., Graimann, B., et al. (2003). Spatiotemporal patterns of beta desynchronization and gamma synchronization in corticographic data during self-paced movement. *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology, 114*(7), 1226.

91. Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: implications for voicing perception in stops. *The Journal of the Acoustical Society of America, 61*(5), 1352-1361.

92. Pulvermuller, F., & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience, 11*(5), 351-360.

93. Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association, 66*(336), 846-850.

94. Ray, S., & Maunsell, J. H. R. (2011). Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. *PLoS Biol, 9*(4), e1000610.

95. Roland, J., Brunner, P., et al. (2010). Passive real-time identification of speech and motor cortex during an awake craniotomy. *Epilepsy & Behavior, 18*(1-2), 123-128.

96. Schalk, G., Leuthardt, E., et al. (2008). Real-time detection of event-related brain activity. *Neuroimage, 43*(2), 245-249.

97. Siegel, M., Donner, T. H., & Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nat Rev Neurosci, 13*(2), 121-134.

98. Sinai, A., Bowers, C., et al. (2005). Electrocorticographic high gamma activity versus electrical cortical stimulation mapping of naming. *Brain, 128*(7), 1556-1570.

99. Steinschneider, M., Fishman, Y. I., & Arezzo, J. C. (2008). Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. *Cerebral Cortex, 18*(3), 610-625.

100. Steinschneider, M., Nourski, K. V., et al. (2011). Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cerebral Cortex, 21*(10), 2332-2347.

101. Steinschneider, M., Volkov, I. O., et al. (2005). Intracortical responses in human and monkey primary auditory cortex support a temporal processing mechanism for encoding of the voice onset time phonetic parameter. *Cerebral Cortex, 15*(2), 170-186.

102. Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America, 111*(4), 1872-1891.

103. Stevens, K. N., & Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *The Journal of the Acoustical Society of America, 55*(3), 653-659. doi: doi:http://dx.doi.org/10.1121/1.1914578

104. Sundberg, J., Nord, L., & Carlson, R. (1991). *Music, language, speech and brain*: Macmillan Basingstoke.

105. Sussman, H. M. (1986). A neuronal model of vowel normalization and representation. *Brain and Language, 28*(1), 12-23.

106. Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America, 79*(4), 1086-1100.

107. Theunissen, F. E., David, S. V., et al. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems, 12*(3), 289-316. doi: doi:10.1080/net.12.3.289.316

108. Tremblay, S. p., Shiller, D. M., & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature, 423*(6942), 866-869.

109. Wang, K., & Shamma, S. (1994). Self-normalization and noise-robustness in early auditory representations. *Speech and Audio Processing, IEEE Transactions on, 2*(3), 421-435.

110. Wernicke, C. (1874). Der aphasische Symptomencomplex.

111. Wilson, S. M., & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. *Neuroimage, 33*(1), 316-325.

112. Wilson, S. M., Saygin, A. P., et al. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*(7), 701-702.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Connie Cheung

5/16/2014

______________________________ ______________________

Connie Cheung Date