Dynamic and adaptive processing of speech in the human auditory cortex

Bahar Khalighinejad

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020
Abstract

Dynamic and adaptive processing of speech in the human auditory cortex

Bahar Khalighinejad

Communicating through speech is an important part of everyday life, and losing that ability can be devastating. Millions of patients around the globe have lost the ability to hear or speak due to auditory cortex deficits. Doctor’s ability to help these patients has been hindered by a lack of understanding of the speech processing mechanisms in the human auditory cortex. This dissertation focuses on enhancing our understanding of the mechanisms of speech encoding in human primary and secondary auditory cortices using two methods of electroencephalography (EEG) and electrocorticography (ECoG).

Phonemes are the smallest linguistic elements that can change a word’s meaning. I characterize EEG responses to continuous speech by obtaining the time-locked responses to phoneme instances (phoneme-related potential). I show that responses to different phoneme categories are organized by phonetic features, and each instance of a phoneme in continuous speech produces multiple distinguishable neural responses occurring as early as 50 ms and as late as 400 ms after the phoneme onset. Comparing the patterns of phoneme similarity in the neural responses and the acoustic signals confirms a repetitive appearance of acoustic distinctions of phonemes in the neural data. Analysis of the phonetic and speaker information in neural activations reveals that different time intervals jointly encode the acoustic similarity of both phonetic and speaker categories. These findings provide evidence for a dynamic neural transformation of low-level speech features as they propagate along the auditory pathway, and form an empirical framework to study the representational changes in learning, attention, and speech disorders.
Later in this dissertation, I use ECoG neural recordings to explore mechanisms of speech communication in real-world environments that require adaptation to changing acoustic conditions. I explore how the human auditory cortex adapts as a new noise source appears in or disappears from the acoustic scene. To investigate the mechanisms of adaptation, neural activity in the auditory cortex of six human subjects were measured as they listened to speech with abruptly changing background noises. Rapid and selective suppression of acoustic features of noise in the neural responses are observed. This suppression results in enhanced representation and perception of speech acoustic features. The degree of adaptation to different background noises varies across neural sites and is predictable from the tuning properties and speech specificity of the sites. Moreover, adaptation to background noise is unaffected by the attentional focus of the listener. The convergence of these neural and perceptual effects reveals the intrinsic dynamic mechanisms that enable a listener to filter out irrelevant sound sources in a changing acoustic scene.

Finally, in the last chapter, I introduce the Neural Acoustic Processing Library (NAPLib). NAPLib contains a suite of tools that characterize various properties of the neural representation of speech, which can be used for characterizing electrode tuning properties, and their response to phonemes. The library is applicable to both invasive and non-invasive recordings, including electroencephalography (EEG), electrocorticography (ECoG) and magnetoencephalography (MEG).

Together, this dissertation shows new evidence for dynamic and adaptive processing of speech sounds in the auditory pathway, and provides computational tools to study the dynamics of speech encoding in the human brain.
Table of Contents

List of Figures .......................................................................................................................... iii
List of Tables .............................................................................................................................. vi
Acknowledgments .................................................................................................................... vii

1 Introduction ............................................................................................................................ 1

2 Background ............................................................................................................................ 4
  2.1 The peripheral and central auditory pathways ................................................................. 4
  2.2 Auditory Cortex .................................................................................................................. 7
  2.3 EEG and ECoG recordings ................................................................................................. 8
  2.4 Acoustic and linguistic features of speech ....................................................................... 11

3 Phoneme related potentials .................................................................................................. 16
  3.1 Method ............................................................................................................................... 17
  3.2 Grand average phoneme related potential ....................................................................... 21
  3.3 Encoding of phonetic categories ..................................................................................... 23
  3.4 Time course of phonetic feature encoding ...................................................................... 25
  3.5 Recurrent appearance of acoustic similarity of phonemes ........................................... 29

4 Joint encoding of phonemes and speakers in PRPs ......................................................... 33
  4.1 Method ............................................................................................................................... 34
  4.2 variations in acoustic cues of different speakers is encoded in the PRP ....................... 34
  4.3 Speaker representation in neural responses of different phonemes ............................. 37
  4.4 Discussion ........................................................................................................................ 41
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Adaptation enhances phonetic distinctions</td>
<td>44</td>
</tr>
<tr>
<td>5.1</td>
<td>Method</td>
<td>45</td>
</tr>
<tr>
<td>5.2</td>
<td>Neural adaptation to changing background condition</td>
<td>49</td>
</tr>
<tr>
<td>5.3</td>
<td>Adaptation suppresses the representation of noise features</td>
<td>55</td>
</tr>
<tr>
<td>5.4</td>
<td>Adaptation enhances phonetic distinctions</td>
<td>62</td>
</tr>
<tr>
<td>6</td>
<td>Effect of brain regions and attention in adaption patterns</td>
<td>71</td>
</tr>
<tr>
<td>6.1</td>
<td>Method</td>
<td>72</td>
</tr>
<tr>
<td>6.2</td>
<td>Adaptation magnitude varies across neural sites</td>
<td>73</td>
</tr>
<tr>
<td>6.3</td>
<td>Adaptation is unaffected by the attentional focus</td>
<td>79</td>
</tr>
<tr>
<td>6.4</td>
<td>Spatial organization of the adaptation patterns</td>
<td>83</td>
</tr>
<tr>
<td>6.5</td>
<td>Discussion</td>
<td>89</td>
</tr>
<tr>
<td>7</td>
<td>NapLib: Neural acoustic processing library</td>
<td>93</td>
</tr>
<tr>
<td>7.1</td>
<td>Toolbox Description</td>
<td>94</td>
</tr>
<tr>
<td>7.2</td>
<td>Experiments</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>Conclusion</td>
<td>105</td>
</tr>
<tr>
<td>9</td>
<td>References</td>
<td>107</td>
</tr>
</tbody>
</table>
List of Figures

Figure 2.1 The structure of the human ear containing outer, middle and inner ear. ........................................ 5
Figure 2.2 The auditory pathway. .................................................................................................................. 7
Figure 2.3 Electrical fields generated by aligned pyramidal cells ................................................................. 9
Figure 2.4 Comparison of waveforms generated by ECoG and EEG activities ........................................... 10
Figure 2.5 An example of auditory spectrogram. ......................................................................................... 12
Figure 2.6 Articulatory organs ..................................................................................................................... 12
Figure 2.7 Comparison of articulation gesture and average phoneme spectrogram .................................... 15
Figure 3.1 PRP calculation method. .............................................................................................................. 21
Figure 3.2: The average EEG response to all phonemes. ........................................................................... 22
Figure 3.3 Hierarchical clustering of PRPs .................................................................................................. 24
Figure 3.4 PRP response to four manners of articulation .......................................................................... 25
Figure 3.5 F-statistic between four manners of articulation over time ..................................................... 26
Figure 3.6 Scalp topographies of the F-statistic for R1 to 4. ................................................................. 27
Figure 3.7 Confusion matrices of RLS classification. ............................................................................... 28
Figure 3.8 Effect size accumulated over subjects and stimulus duration ............................................... 28
Figure 3.9 Patterns of phoneme similarity in EEG. .................................................................................. 30
Figure 3.10 One-dimensional representation of the PRP distance matrices ........................................... 30
Figure 3.11 Recurrent appearance of patterns of phoneme similarity. .................................................... 32
Figure 4.1 Encoding of speaker characteristics in PRPs. ................................................................. 35
Figure 4.2 MDS diagram of phonemes and speakers ............................................................................. 36
Figure 4.3 Joint encoding of phonemes and speakers ............................................................................ 37
Figure 4.4 PRP of vowels for speaker 1 versus 2 ..................................................................................... 38
Figure 4.5 MDS diagram of vowels of the two speakers ......................................................................... 39
Figure 4.6 Comparison of timing and location for phonetic and speaker encoding. ............................ 40
Figure 5.1 Electrodes with significant response to speech. .................................................................... 49
Figure 5.2 Location of grid and depth electrodes on MRI of each individual subject. .......................................................... 50
Figure 5.3 The experimental design with changing background noise. ............................................................................. 51
Figure 5.4 Spectrotemporal properties of each background condition. ................................................................................. 52
Figure 5.5 Neural adaptation in transition to a new background condition. .......................................................... 56
Figure 5.6 Clean speech versus noisy speech task. ....................................................................................... 57
Figure 5.7 Average frequency profile of the conditions. ................................................................................. 58
Figure 5.8 R2 for individual trials. ........................................................................................................ 58
Figure 5.9 Speech masking for each condition. ............................................................................................. 59
Figure 5.10 Decrease in the signal-to-noise ratio (SNR) of the representation of speech for individual subjects. ... 60
Figure 5.11 The neural response to the noise and speech alone. ............................................................................. 61
Figure 5.12 Speech-specificity task and the speech-specific electrodes. .......................................................... 63
Figure 5.13 Average phoneme spectrograms. ............................................................................................. 64
Figure 5.14 Phoneme dissimilarity matrices in during adaptation versus after adaptation......................... 65
Figure 5.15 The discrimination between phonemes in DA versus AA. ............................................................. 66
Figure 5.16 Behavioral experiment designed to test the intelligibility of consonant-vowel pairs. ......................... 67
Figure 5.17 The psychoacoustic task user interface. ......................................................................................... 68
Figure 5.18 Comparison of correctly recognized phonemes occurring during adaptation (DA) and after adaptation (AA) intervals. ................................................................................................................................. 69
Figure 5.19 Comparison of correctly recognized phonemes for each condition separately. ......................... 70
Figure 6.1 Comparison of adaptation time and electrodes for different background conditions. ....................... 74
Figure 6.2 Diversity of adaptation patterns across electrodes. ............................................................................. 76
Figure 6.3 Neural adaptation patterns for individual subjects .................................................................................. 77
Figure 6.4 Average latency of electrodes from the top two clusters ........................................................................ 78
Figure 6.5 Visual distraction task ..................................................................................................................... 79
Figure 6.6 Decrease in comprehension accuracy during distracted condition. ....................................................... 80
Figure 6.7 Average high-gamma activity during distracted versus attended condition. ........................................ 81
Figure 6.8 Comparison of neural adaptation in attended and distracted conditions for individual subjects .......... 82

Figure 6.9 Spatial organization of adaptive responses to each background condition. ........................................ 83

Figure 6.10 Anatomical organization of electrodes with the largest adaptation to each background condition ...... 84

Figure 6.11 Spatial organization of the tuning parameters ................................................................................. 85

Figure 6.12 Calculating rate and frequency tuning properties from STRFs of the electrodes ............................... 86

Figure 6.13 R-squared values for predicting the adaptation indices ................................................................. 88

Figure 7.1 Pipeline of phoneme analysis can be implemented in real-time ......................................................... 99

Figure 7.2 Average response to phonemes vs STRF ......................................................................................... 101

Figure 7.3 Finding latency using Fratio maxima ............................................................................................... 102

Figure 7.4 Duration of speech required for generating phoneme mapping ....................................................... 103

Figure 7.5 Visualization of response location ................................................................................................. 104
List of Tables

Table 1 Phonetic features for English consonants. ............................................................................................................ 14

Table 2. Individual subject’s phoneme recognition score. .................................................................................................. 69
Acknowledgments

First and foremost, I would like to thank my Ph.D. advisor, Professor Nima Mesgarani. I have learned a lot from him during past years. His knowledge and unique approach in research always inspired me. He taught me that each problem has many different hidden angles, and to find the truth I should be hungry of learning and exploring. He pushed me to be open-minded and patient with new ideas, and it expanded the boundaries of my mind in a way that I could never imagine.

I would like to thank Professor Ashesh Mehta, and Professor Jose Herrero for many years of fruitful collaboration. I would like to thank Elsa Sanchez who has been truly caring and helpful during my years in Columbia University. I would like to thank Professor Rui Costa for his help when I needed it the most.

I would like to thank the past and present members of Neural Acoustic Processing Laboratory. They changed me to a more knowledgeable person. To Tasha, Minda, Laura, James, Sam, Vinay, Hassan, Menoua, Prachi, Moritz, Rashida, Ali, Yi, Zhuo, Cong, and everybody else. Thank you for many fun collaborations. Thank you for all the squash and tennis games, for all the ski trips and bouldering weekends, for all the super bowl nights and hangouts we had!

I also thank my friends and family for their support and friendship. I am very lucky to have grandparents, aunts, uncles, cousins and friends who always supported me. I am specifically thankful of Mitra, Shahram, Omid, and Grace for their supportive companionship during my PhD years.

Finally, my deepest gratitude to the best people in my life, Sadra, Minoo, Jalal and Ghazal. They understand me, they know me, and they always encourage me to do what I believe in. Their generous personalities, unconditional love and belief in me are what drives me in life. This thesis is dedicated to them!
1 Introduction

When listening to speech, humans have the ability to simultaneously extract information about both the content of the speech and the identity of the speaker. We are able to automatically accomplish these parallel processes by decoding multitude of cues encoded in the acoustic signal, including distinctive features of phonemic categories that carry meaning as well as identifiable features of the speaker such as pitch, prosody, and accent. Despite the extensive research to model and describe these processes, we still have no comprehensive and accurate framework of the transformation and representation of speech in the human brain. Recent invasive human neurophysiology studies have demonstrated the encoding of phonetic features in higher level auditory cortices. However, invasive recordings are limited to confined brain regions and are therefore impractical for studying the neural representation of acoustic features over time as speech sounds propagate through the auditory cortex.

Electroencephalography (EEG) has been used extensively in speech and language studies because it can measure the activity of the whole brain with high temporal resolution. EEG studies of speech perception are primarily based on event-related potentials (ERPs). For example, ERPs have been used to study the encoding of acoustic features in response to isolated consonant-vowel pairs, showing a discriminant encoding at multiple time points (e.g. P1-N1-P2 complex) and locations (i.e. fronto-central and temporal electrodes). In addition, ERPs have been used in studies of higher level speech units, such as word segmentation and multi-scale hierarchical speech processing.

Nonetheless, ERP approaches suffer from unnatural experimental constraints (for example, requiring isolated, non-overlapping events), which may result in only partial engagement of the
underlying cortical circuits \textsuperscript{21}. As a result, the generalization of these findings to more naturalistic settings remains unclear. Several recent studies have examined EEG responses to continuous speech by correlating the responses with the speech envelope \textsuperscript{22–25} and by regressing the neural responses against speech envelope \textsuperscript{26} or against the phonetic features and phonemes \textsuperscript{27}. To study the precise temporal properties of neural responses to acoustic features, we propose an ERP method where the events are the instances of phonemes in continuous speech.

Using PRPs, we characterized the properties of neural responses to phoneme instances in continuous speech. We show that each instance of a phoneme in continuous speech produces several observable neural responses at different times occurring as early as 50 ms and as late as 400 ms after the phoneme onset. Each temporal event explicitly encodes the acoustic similarity of phonemes, and linguistic and nonlinguistic information are best represented at different time intervals. We show a joint encoding of phonetic and speaker information, where the neural representation of speakers is dependent on phoneme category.

While we show the encoding of phonetic and speaker information, speech communication under real-world conditions rarely happens in a sound-proof room! In everyday life, we carry on conversations as cars and people pass by. We go from one noisy environment to another, and we are still able to filter out the background noise to understand the message that was mixed with the noise. Even though hearing in noise comes natural to us, the underlying neural process is quite complex. It requires a listener’s auditory system to continuously monitor the incoming sound, and tease apart the acoustic features of speech from the background noise \textsuperscript{28}. This process results in an internal representation of the speech signal that enables robust speech comprehension unaffected by the changes in the acoustic background \textsuperscript{29}. 
Studies of the representational properties of vocalization sounds have confirmed the existence of a noise-invariant representation in animal auditory cortex. Specifically, it has been shown that the auditory cortical responses in animals selectively encode the vocalization features over the noise features\textsuperscript{30–34}. A noise-invariant representation of speech in the human auditory cortex has also been shown\textsuperscript{35,36}, but the encoding properties of speech in noise in humans are less clear due to the limited spatiotemporal resolution of noninvasive neuroimaging methods. Previous studies of the neural representation of speech or vocalizations-in-noise have used constant background noises\textsuperscript{30–36}. As a consequence, their findings only show the aftereffects of adaptation and the properties of the neural representation once the noise has been removed. Therefore, it remains unclear how, when, and where adaptation unfolds from moment to moment as a new background noise suddenly appears in or disappears from the acoustic scene. For this reason, many important questions regarding the dynamic properties of adaptation to noisy speech in the human auditory cortex remain unanswered, such as I) how the invariant representation of vocalizations emerges over the time course of adaptation, II) how the neural representation and perception of phonetic features change over the time course of adaptation, and III) how cortical areas with different response properties adapt when transitioning to a new background condition. Answering these questions are crucial for creating a complete dynamic model of speech processing in the human auditory cortex.

In this thesis, we shed light on the dynamic mechanisms of speech-in-noise processing in the human auditory cortex. We show rapid suppression of noise features in the cortical representation of acoustic scene, resulting in enhanced neural representation and perception of phonetic features in noise.
2 Background

This chapter is divided into four sections and provides a review of the background and the basis of this dissertation. The first section provides an introduction to the peripheral and central auditory pathway in humans. The second section describes the auditory cortex and its role in speech processing. The third section presents the neural recording methods of EEG and ECoG, which are employed in all studies discussed in this dissertation. The fourth section discusses the auditory features of speech.

2.1 The peripheral and central auditory pathways

In its physical form, sound is the vibration of air that can be heard when it reaches the auditory pathway of humans or animals. The auditory pathway contains a series of stages, starting with the peripheral auditory system. The outer ear is the first stage of the peripheral auditory system, followed by the middle ear and the inner ear. The transformation of sound starts from the pinna. The pinna—the only visible part of the ear—constitutes of the folds of cartilage and skin. The pinna helps people locate sounds in the space and eliminates a small segment of frequency range known as the pinna notch.

The way the pinna transforms sound is dependent on the direction of sound and its frequency. The pinna affects the low-frequency sounds that come straight ahead less than the high-frequency sounds that come from above. The outer ear is connected to the middle ear through the ear canal. The ear canal has a resonant frequency of around 2–5 kHz and amplifies the sounds within this frequency range. Sound reaches the tympanic membrane or eardrum, which is at the end of the ear canal. As its name suggests, the eardrum is a thin, cone-shaped tissue that separates
the middle ear from the inner ear. The vibrations of the eardrum are transferred to the inner ear through three delicate bones known as the malleus, incus, and stapes.

The cochlear lays toward the end of peripheral auditory system. The cochlear acts like an analog filter bank and transforms the mechanical movement caused by vibration of sound to neuronal electrical activity. The three stages of the human ear are shown in Figure 2.1.

![Figure 2.1 The structure of the human ear containing outer, middle and inner ear.](Source: Noback 1967)

The inner ear is connected to the central auditory system via the vestibulocochlear nerve. The vestibulocochlear nerve transmits both sound and equilibrium. After the sound reaches the
central auditory system, the first stage is the cochlear nucleus (CN). The CN has two cranial nerve nuclei in the brainstem: the ventral cochlear nucleus (VCN) and the dorsal cochlear nucleus (DCN). The CN processes multiple types of information by passing the information through multiple types of principal cells. The VCN extracts information using the firing rate and population activity of the input nerves. The DCN performs non-linear spectral analysis and takes into account the location of the head, ears, and shoulders, and distinguishes the self-generated sounds from the unexpected input from the outside.37

Immediately exiting the CN, the sound enters the superior olivary complex, where the information from both ears converges and where analysis related to binaural hearing is performed. This stage calculates the interaural time difference and interaural level difference. The next stages are the inferior colliculus (IC) of the mid-brain and medial geniculate nuclease (MGN) of the thalamus.37 After the thalamus, sound arrives in the auditory cortex, which is the focus of this dissertation. Figure 2.2 shows the pathway of the auditory system from the ear to auditory cortex.
2.2 Auditory Cortex

Sound perception in natural conditions requires extraction of multiple acoustic attributes to form distinct unites such as phonemes, words, and speaker identities.\textsuperscript{4,39-42} Higher processing primarily occurs in the auditory cortex. The auditory cortex is located bilaterally on the upper sides of the brain’s temporal lobes. The three main regions of the human auditory cortex are Heschl’s gyrus (HG), planum temporale (PT), and superior temporal gyrus (STG). HG is an important part of the primary auditory cortex and STG is a key region in the secondary auditory cortex. The classical view of speech processing in humans portrays that simple acoustic spectrum features such
as frequency are processed in the primary auditory cortex. Next, the phonemes are extracted in the posterior superior temporal gyrus and sulcus, and words are extracted in the lateral and ventral temporal cortex.\textsuperscript{7,40,43} This view is partially adopted from the macaque auditory cortex, which can be divided into three main areas: core, belt, and parabelt.\textsuperscript{44} The primary region includes core processes and the primary feature is frequency.\textsuperscript{45}

It is difficult to obtain direct measurements of brain activity in humans; therefore, most of the current information on functional organization of the human auditory cortex is based on fMRI studies. However, fMRI is not a direct measure of neuronal activity and does not have adequate temporal resolution to study the sound features that occur at the scale of hundreds of milliseconds, such as latency and phonemes. To achieve an adequate temporal and spatial resolution we used two methods of EEG and ECoG.

### 2.3 EEG and ECoG recordings

EEG was discovered by Hans Berger on 1929 and is one of the oldest methods of neural recordings. The recorded electroencephalogram is a spatiotemporally smoothed version of the local field potentials (LFP) and is integrated over an area of 10 cm\(^2\) or more. Approximately 10\(^5\) neurons exist under a mm\(^2\) of a cortical surface; therefore, only a minor relationship exists between EEG and firing patterns of individual underlying neurons. The distorting and attenuating effects of tissues between the electrical activity source and the electrodes causes a low spatial resolution in EEG.\textsuperscript{46} The EEG potentials are largely generated by a population of cortical cells oriented perpendicularly to the cortical surface. These cells have branch-like dendrites and are referred to as pyramidal cells. The synchronous activity of pyramidal cells produces the potentials observed in EEG (see Figure 2.3).
To overcome the low spatial resolution of EEG, we also used intracranial, subdural electrodes in epilepsy patients. Due to the frequency of medial temporal seizures, the surface electrocorticography (ECoG) electrodes and depth intracranial (iEEG) electrodes are often placed over or into the superior temporal lobe. This provides a rare possibility for researchers to record potentials from the surface of the human brain while the patient is performing a task or listening to stimuli. The depth and grid electrodes have a much higher signal to noise ratio in comparison to EEG. To better compare the shape of patterns in ECoG versus EEG electrodes, in Figure 2.4 (adopted from Buzsáki 2012, figure 1 46), we show the simultaneous recordings from three depth electrodes (two selected sites each) in the left amygdala and hippocampus (measuring the local field potential (LFP)); a subdural grid electrode array placed over the lateral left temporal cortex.
(measuring the ECoG); two four-contact strips placed under the inferior temporal surface (measuring the ECoG); an eight-contact strip placed over the left orbitofrontal surface (measuring the ECoG); and scalp electroencephalography (EEG) over both hemispheres (selected sites are the Fz and O2) in a patient with drug-resistant epilepsy.

Figure 2.4 Comparison of waveforms generated by ECoG and EEG activities

(source: Buzsáki 2012, figure 1 46).

Together, the ECoG and EEG recordings help us better understanding the neural mechanisms underlying phoneme and speaker processing in clean and noisy hearing conditions.
2.4 Acoustic and linguistic features of speech

Auditory Spectrogram

A spectrogram is the frequency-time representation of sound that can be generated from the waveform using a filter bank and non-linear functions. Auditory spectrogram is a type of spectrogram where the filter banks and nonlinearities are calculated based on the properties of the peripheral auditory system. In this thesis, the auditory spectrogram is estimated using a computational model inspired by psychoacoustical and neurophysiological findings in early and central stages of the auditory system \(^47\). The first stage of the model consists of a bank of constant 128 asymmetric filters equally spaced on a logarithmic axis. The filter bank output is subject to nonlinear compression, followed by a first order derivative along spectral axis (modeling lateral inhibitory network), and finally an envelope estimation operation. This results in a two dimensional representation simulating the pattern of activity on the auditory nerve \(^47\). The Matlab code to calculate the auditory spectrogram is available at:

https://isr.umd.edu/Labs/NSL/Software.htm.

Figure 2.5 shows the auditory spectrogram for the sentence “How to identify a bird”. As it is shown with highlighted color in the auditory spectrogram, speech has many high-frequency segments in addition to low-frequency ones. The high frequencies are enhanced in auditory spectrogram because filter banks in auditory spectrogram are designed in a way that they have a smaller band and are more discriminative at lower frequencies and have wider band and are less discriminative at higher frequencies.
Figure 2.5 An example of auditory spectrogram.

**Phonemes**

Phonemes are the smallest linguistic unit that can change a word’s meaning. For example, the difference between two words of /dad/ and /bad/ is the difference between phoneme /d/ and /b/. It is fascinating that more than ten articulatory organs work together to generate the distinct sound of each phoneme. Figure 2.6 (adopted from http://www.indiana.edu/~hlw/PhonUnits/vowels.html) shows a number of articulatory organs, later in this section we show how these organs generate distinct phonetic features.

Figure 2.6 Articulatory organs

The 44 English phonemes fall into two main categories of vowels and consonants. Vowels such as /æ/, /ei/, /i/ are generated when the mouth is fairly open and airflow can pass without obstruction or friction with articulatory organs. Consonants such as /b/, /f/, /l/ are generated when the mouth is fairly closed, and when friction and obstruction of air occurs as airflow passes the articulatory organs.

Manner of articulation is an important feature of consonants and it is dependent on the interaction between speech organs such as the tongue, lips, and palate, and the passage of airflow from these organs. Plosives such as /b/, /d/, and /g/ are generated by making a short and complete obstruction in the airflow. Nasals such as /m/, and /n/ are generated by passing the airflow from the nose. Fricatives such as /z/, and /s/ are generated by continuous friction between airflow and articulatory organs, and finally approximants or semivowels such as /w/ and /r/ are the closest to vowels, as their name suggests. Approximates have very little obstruction.

Along with manner of articulation, place of articulation has an important effect on how phonemes sound. Place of articulation refers to the location of obstruction generated by articulatory organs. Eight places of articulation exist in the English language. Starting from the front of the mouth, the places of articulation for different consonant phonemes are bilabial, labiodental, dental, alveolar, post-alveolar, palatal, velar and glottal as shown in Table 1.

Many consonant phonemes have two versions where manner of articulation and place of articulation are the same but the voicing is different. Unvoiced sounds are produced with vocal cords apart whereas voiced sounds are produced with vocal cords touching and vibrating. The
examples of voiced phonemes are /b/,/v/,/z/,/d/, and their unvoiced versions are /p/,/f/,/s/,/t/ respectively.

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Post-alveolar</th>
<th>Palatal</th>
<th>Velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>/b/,/p/</td>
<td></td>
<td></td>
<td>/t/,/d/</td>
<td></td>
<td>/k/,/g/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>/f/,/v/</td>
<td>/th/,/dh/</td>
<td>/s/,/z/</td>
<td></td>
<td>/ng/</td>
<td></td>
<td></td>
<td>/h/</td>
</tr>
<tr>
<td>Affricate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>/ch/,/sh/</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>/m/</td>
<td></td>
<td></td>
<td></td>
<td>/n/</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td>/w/</td>
<td></td>
<td>/r/,/l/</td>
<td></td>
<td>/y/</td>
<td>/w/</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 Phonetic features for English consonants.

To better understand how articulatory features determine acoustic properties of phonemes, the average auditory spectrogram generated by averaging over 300 phone instances from our stimuli corpus were compared to the articulation gestures that are performed to generate the phonemes. The articulation gestures were adopted from

http://www.indiana.edu/~hlw/PhonUnits/consonants1.html.

Figure 2.7 illustrates the two plosives of /d/ and /g/ where a complete closure of airflow occurs. This closure results in a sudden silence (the blue part) in the spectrogram. The effect of place of articulation can also be observed; for example, /d/ is an alveolar plosive and shows higher
frequencies during the release in comparison to /g/, which is a glottal phoneme. It can also be seen that /m/, which is a nasal, has lower frequency power in comparison to /d/ and /g/.

Figure 2.7 Comparison of articulation gesture and average phoneme spectrogram

In next chapters, we will examine the representation of phonemes and articulatory features in the neural data. We show how phonetic features are encoded in EEG and ECoG data and what happens to their representation when the listener enters a noisy environment, and how brain responds to acoustic features of phonemes when they are embedded in a noisy background.
3 Phoneme related potentials

Portions excerpted from:

3.1 Method

Participants

Participants were 22 native speakers of American English with self-reported normal hearing. 20 were right-handed. 12 males, 10 females.

Stimuli and Procedure

EEG data were collected in a sound-proof, electrically shielded booth. Participants listened to short stories with alternating sentences spoken by a male and a female speaker; we alternated sentences to normalize time-varying effects such as DC drift on speaker-dependent EEG responses. The stimuli were presented monophonically at a comfortable and constant volume from a loudspeaker in front of the subject. Five experimental blocks (12 minutes each) were presented to the subject with short breaks between each block. Subjects were asked to attend to the speech material. To assess attention, subjects were asked three questions about the content of the story after each block. All subjects were attentive and could correctly answer more than 60% of the questions. Participants were asked to refrain from movement and to maintain visual fixation on the center of a crosshair placed in front of them. All subjects provided written informed consent. The Institutional Review Board (IRB) of Columbia University at Morningside Campus approved all procedures.

Recording

EEG recordings were performed using a g.HIamp bio-signal amplifier (Guger Technologies, Austria) with 62 active electrodes mounted on an elastic cap (10-20 enhanced montage). EEG data was recorded at a sampling rate of 2 kHz. A separate frontal electrode (AFz) was used as ground and the average of two earlobe electrodes were used as reference. The choice of earlobe as reference in studies of auditory evoked potentials is motivated by the highly correlated activity
across electrodes which makes common reference averaging unsuitable. EEG data were filtered online using a 0.01Hz fourth-order high-pass Butterworth filter to remove direct current (DC) drift. Channel impedances were kept below 20 kΩ throughout the recording.

**Preprocessing**

EEG data were filtered using a zero-lag, FIR bandpass filter with cut-off frequencies of 2 and 15 Hz. The frequency range was determined by measuring the average power of the PRP at different frequencies which shows that PRP peaks at 8 Hz (the syllabic rate of speech). For each subject, we normalized the neural response of each EEG channel to ensure zero mean and unit variance.

**Phoneme-related potential**

To obtain a time-locked neural response to each phone, the stimuli were first segmented into time-aligned sequences of phonemes using the Penn Phonetics Lab Forced Aligner Toolkit. The EEG data were then segmented and aligned according to phoneme onset. Response segments where the magnitude exceeded +/-10 units were rejected to reduce the effect of biological artifacts such as eye blinking. On average of 8% of data was removed for each subject. Neural responses within the first 500 ms after the onset of each utterance were not included in the analysis to minimize the effect of onset responses.

Phoneme-related potentials and average auditory spectrograms of phonemes were calculated by averaging the time-aligned data over each phoneme category. Defining $s(f, t)$ as the acoustic spectrogram at frequency $f$ and time $t$, and $r(e, t)$ as the EEG response of electrode $e$ at time $t$, the average spectrograms and PRP for phoneme $k$ which occurs $N_k$ times and starts at time points $T_{k1}, T_{k2}, \ldots, T_{kn}$ are:

$$
\tilde{S}(k, f, \tau) = \frac{1}{N_k} \sum_{n=1}^{N_k} s(f, T_{kn} + \tau), \quad PRP(k, e, \tau) = \frac{1}{N_k} \sum_{n=1}^{N_k} r(e, T_{kn} + \tau)
$$

(1)
Where $\bar{S}(k, f, \tau)$ is the average auditory spectrogram of phoneme $k$, at frequency $f$ and time $\tau$, and $PRP(k, e, \tau)$ is the average response of phoneme category $k$, at electrode $e$ and time $\tau$ relative to the onset of the phoneme \textsuperscript{52}. As shown in equation 1, $PRP$ is a function of time relative to the onset of phonemes.

To group the PRPs based on their similarity, we performed unsupervised hierarchical clustering based on the unweighted pair group method with arithmetic mean algorithm (UPGMA, Euclidean distance) \textsuperscript{53}. To study the separability of different manners of articulation in neural and acoustic space, we used F-statistic at each time point to measure the ratio of the distance between and within different manner of articulation groups.

**Neural representation of acoustic phonetic categories**

Pairwise phoneme distances were estimated using a Euclidean metric \textsuperscript{54} to measure the distance of each phoneme relative to all other phonemes. This analysis results in a two-dimensional symmetric matrix reflecting a pattern of phoneme similarity which can be directly compared to the distance patterns estimated at different time points.

We compared neural versus acoustic organization of phonemes by finding the covariance \textsuperscript{55} value between distance matrices in the acoustic and neural signals. The covariance was calculated from only the lower triangular part of the distance matrices to prevent bias caused by the symmetric shape of the matrix. Calculating the covariance values at all time lags in acoustic and neural spaces results in a two dimensional neural-acoustic similarity measure at all time lags.

In addition to the neural-acoustic covariance matrix, we calculated a neural-neural similarity matrix by comparing the pairwise phoneme distances at different time lags in PRPs.

To visualize the relational organization of PRPs at different time lags, we applied one-dimensional unsupervised multidimensional scaling (MDS) using Kruskal’s normalized criterion.
to minimize stress for one dimension. The MDS was set to zero when no electrode showed a significant response (multiple comparison corrected via FDR, q<0.001).
3.2 Grand average phoneme related potential

We recorded EEG data from 22 native speakers of American English. Participants listened to simple stories comprised of alternating sentences uttered by two speakers (one male, one female). To investigate whether phonemes in continuous speech elicit distinct and detectable responses in the EEG data, we used phonetic transcription of speech data to segment and align the neural responses to all phoneme instances (Figure 3.1). We refer to the resulting time-locked evoked responses to phonemes as phoneme-related potentials (PRPs).

![Speech spectrogram, EEG waveform, Phonetic transcription]

**Figure 3.1 PRP calculation method.**

By averaging over all phonemes, we found a robust PRP response at most electrodes. The response of a representative electrode (central electrode Cz) is shown in Figure 3.2. We applied two-tailed paired t-test (Corrected for False Discovery Rate [FDR] , q<0.01) to compare the PRP response with baseline activity. We observed three statistically significant time intervals of 50 to 90 ms (Response 1 or R1, positive deflection), 100 to 160 ms (R2, negative deflection) and
190 to 210 ms (R3, positive deflection). The distribution of the PRP across electrodes shows a broadly distributed response strongest in fronto-central electrodes, a finding that is consistent with the topographical map of the standard auditory-evoked potential (AEP) on fronto-central electrodes\textsuperscript{26,58}, even though the individual phonemes in continuous speech are not isolated events. Figure 3.2 (top-left) shows the average acoustic spectrogram for all phonemes. Top-right shows the average EEG response to all phonemes averaged over all electrodes. Individual subjects are shown in gray; grand average PRP is shown in red. Time points where PRP shows a significant response are shaded in yellow (central electrode Cz, t-test, multiple comparison corrected via FDR, \( q<0.01 \)). The scalp topographies (bottom of the figure) shows the scalp map of three significant time points of 70 ms, 130 ms, and 200 ms.

![Average spectrogram and Average phoneme-related potential](image)

**Figure 3.2:** The average EEG response to all phonemes.
3.3 Encoding of phonetic categories

To study whether different phonemic categories elicit distinct neural responses, we averaged the PRP responses over all instances of each phoneme and across all subjects, excluding phoneme categories that contained less than 0.01 percent of all phones. Visual inspection of PRPs elicited by each phoneme suggests that they vary in their magnitude and latency, with a varied degree of similarity relative to each other. For example, PRPs for vowels show similar patterns of activation, which differ from that of consonants. Figure 3.3 shows the average response for the fronto-central electrode (FCz), averaged over all subjects.

To determine whether PRPs can be characterized by phonetic feature hierarchy \(^2,5^9\), we used an unsupervised clustering method based on the Euclidean distance between PRPs of different phoneme categories. Hierarchical clustering was performed on neural responses over an interval of 0 to 400 ms after phone onset. This window was chosen to ensure the inclusion of significant components of the average PRP. The Hierarchical clustering shown in Figure 3.3 reveals different tiers of grouping corresponding to different phonetic features: The first tier distinguishes obstruent from sonorant phonemes \(^1\). Within the obstruent tier, a second tier further differentiates categories based on manner of articulation, where plosives (blue) formed a separate group from the fricative (red) phoneme group. Place of articulation features appear in the lower tiers of the hierarchy, separating high vowels from low vowels (light green for high vowels versus dark green for low vowels). Overall, the clustering analysis of PRPs shows that manner of articulation is the dominant feature expressed in the responses, followed by place of articulation particularly for vowels. This finding is consistent with neural representation of speech on the lateral surface of superior temporal gyrus \(^4,5\), the acoustic correlates of manner and place of articulation features \(^2\), and psychoacoustic studies showing more confusions among phonemes with the same manner of articulation \(^60,6^1\).
Figure 3.3 Hierarchical clustering of PRPs

Left: Grand average PRPs for 30 American English phonemes. Right: Hierarchical clustering of PRPs using all electrodes.
3.4 **Time course of phonetic feature encoding**

To study the temporal characteristics of PRPs, we grouped the PRPs according to the top clusters identified in previous section, which also corresponds to the manner of articulation categories of plosives, fricatives, nasals, and vowels. Each of these phonemic categories have distinctive spectro-temporal properties. For example, plosives have a sudden and spectrally broad onset, fricatives have an energy peak in higher frequencies, and vowels have relatively centered activity at low to medium frequencies. As the vowels become more “Front”ed, the single peak broadens and splits. In comparison to vowels, nasals are spectrally suppressed. The time course of manner-specific PRPs shows discrimination between different manners of articulation as early as 10 ms after phoneme onset to as delayed as 400 ms after phoneme onset. As shown in the next section, this early response (R1, 10 to 50 ms) is mainly due to the structure of speech stimulus which influences the preceding phonemes.

![Figure 3.4 PRP response to four manners of articulation.](image)

Figure 3.4 PRP response to four manners of articulation.
We used the F-statistic to measure the ratio of variance within and between different manners to systematically study the temporal separability of PRPs for different manners of articulation. F-statistic analysis reveals significant separability between manners of articulation (multiple comparison corrected via FDR, q<0.05) with F-statistic peaks observed at four distinct time points (components) centered around 50ms, 120ms, 230ms, and 400ms. As shown in Figure 3.5, repeating the same analysis using the acoustic spectrograms instead of EEG data does not show the late response components, validating their neural origin as opposed to possible contextual stimulus effects. Distinct temporal components were also observed in the average PRP with R1 at 70 ms, R2 at 130 ms, and R3 at 200 ms.

Comparing the F-statistic and average PRP reveals the unique characteristics of each temporal component. For example, although the first component of PRP (R1) elicits the response with the largest magnitude, it is comparatively less dependent on phoneme category compared to R2 and R3, as evidenced by a smaller F-statistic. The peak of the F-statistic indicates that the most phonetically-selective PRP response appears at 120 ms (R2). Additionally, the PRP component occurring at 400 ms (R4) in the F-statistic was not apparent in the average PRP because the opposite signs of deflection at this time point for different manners cancel out.

![Figure 3.5 F-statistic between four manners of articulation over time.](image)
Calculating the F-statistic between manners of articulation for individual electrodes (Figure 3.6) show different scalp distribution for early and late PRP components with a varying degree of asymmetry. For example, two temporal electrodes of T7 and T8 show significant discriminability at R2 and R3 but not at R4. It has been shown that cortical sources of ERP responses recorded at T7 an T8, known as T-complex \(^{63}\), are independent from fronto-central activities \(^{64}\). This suggests that various cortical regions may contribute differently to the response components of R1 to R4 in phoneme-related potentials.

![Scalp topographies of the F-statistic for R1 to 4.](image)

To examine both the separation and overlap of different manners of articulation, we trained a regularized least square (RLS) classifier \(^{65}\) to predict the manner of articulation for individual instances of PRPs (10% of data used for cross validation). The classification accuracy is observed to be significantly above chance for all categories. To compare the confusion patterns of manners of articulation in neural and acoustic spaces, we also tested the classification of manners using spectrograms of phones. Figure 3.7 shows that the confusion patterns in neural and acoustic spaces
are highly correlated ($r=0.59$, $p=0.016$, t-test), suggesting that the acoustic overlap between various phones is also encoded in the neural responses.

Figure 3.7 Confusion matrices of RLS classification.

Finally, to determine the variability of PRPs across subjects, we estimated F-statistics for manners of articulation accumulated over subjects and recording time (Figure 3.8). This analysis is particularly informative because it specifies the minimum number of subjects needed to obtain a statistically significant PRP response for a given experimental duration.

Figure 3.8 Effect size accumulated over subjects and stimulus duration.
3.5 Recurrent appearance of acoustic similarity of phonemes

The previous section illustrates how phonetic feature categories shape PRPs and their distinct temporal components. However, it does not explicitly examine the relationships between the EEG responses and the acoustic properties of speech sounds. Because speech is a time-varying signal with substantial contextual and duration variability, it is therefore crucial to compare the neural and acoustic patterns over time to control for the temporal variability of phonemes. We therefore used pairwise phoneme similarities calculated at each time point relative to the onset of phonemes, and compared the similarity patterns in neural and acoustic data at each time. As a result, this direct comparison can separate the intrinsic dynamic properties of neural encoding from the temporal dependencies that exist in natural speech. Moreover, this analysis focuses on the encoding of similarities and distinctions between categories rather than the encoding of individual items, and has been widely used in the studies of visual system to examine representational geometries and to compare models and stages of object processing 66,67.

We start by calculating the Euclidean distance between the PRPs for each phoneme pair and at every time lag, yielding a time-varying pairwise phoneme distance matrix. We use $m - D$ as a measure of similarity, where $D$ is the distance matrix, and $m$ is the mean value of elements of matrix $D$. Figure 3.9 shows the phoneme neural similarity matrices calculated at time points R1, R2, R3 and R4, where red values indicate more similar phoneme pairs.
To illustrate the temporal progression of relative distances between the PRPs, we used multi-dimensional scaling (MDS) analysis and projected the PRP distance matrices at each time lag to a single dimension, where the MDS values are derived from the responsive electrodes at each time point. The MDS result shows the phoneme separability is largest at R2. Figure 3.10 also shows the difference in timing of responses to different manners of articulation is most apparent at PRP component R3 compared to R1, R2, and R4 components.

Figure 3.9 Patterns of phoneme similarity in EEG.

Figure 3.10 One-dimensional representation of the PRP distance matrices.
To compare phoneme similarity patterns in acoustic and neural data over time, we calculated the acoustic similarity matrix using the acoustic spectrogram of phones \[^{69}\], and found the covariance between the corresponding similarity matrices. The covariance values (neural-acoustic matrix) demonstrate distinct time intervals when the organization of phonemes in PRP mirrors the acoustic organization of phonemes (significant was assessed using bootstrapping, n=20, multiple comparison corrected via FDR, q<0.0001). In particular, the acoustic distance matrix at time interval 10 to 60 ms is significantly similar to the neural data at three time intervals, approximately centered at 120ms (R2); 230ms (R3), and 400ms (R4) after the phoneme onset. R1 (40ms) in neural data, on the other hand, is similar to acoustic patterns at about -30 ms, showing that the observed distinctions between phonemes at R1 are mainly caused by the acoustic structure of the preceding phonemes (Figure 3.11A). We also calculated the covariance between PRP distance matrices at different time lags (neural-neural matrix, bootstrapping, n=20, multiple comparison corrected via FDR, q<0.0001). Figure 3.11B shows that the PRP similarity matrix at R3 is significantly similar to the similarity matrices at R2 and R4. The main diagonal of neural-neural covariance matrix demonstrates the start and ending of the significant PRP responses, as well as showing the strength of phoneme distinction at each duration. In summary, Figure 3.11 shows that the organization of neural responses at time intervals R2, R3 and R4 mirrors the acoustic similarities of different phonemes, and provides compelling evidence for a repetitive appearance of acoustic phonetic distinctions in the neural data.
Figure 3.11 Recurrent appearance of patterns of phoneme similarity.
4 Joint encoding of phonemes and speakers in PRPs

Portions excerpted from:

4.1 Method

Calculating speaker-dependent pairwise phoneme distances

We calculated the pairwise Euclidean distance of PRPs for each speaker resulting in a pairwise phoneme distance matrix with four quadrants, where diagonal quadrants represent within-speaker distances and off diagonal quadrants represent the between-speaker distances. We measured a speaker index by subtracting between-group distances from within-group distances, both in the PRP and spectrogram data. We calculated the correlation between speaker-dependent patterns in neural and acoustic spaces for each time point which yielded a speaker-dependent neural-acoustic correlation matrix.

The speaker-dependent encoding (SE) of phoneme category $i$ is defined as follow, where the distance matrices can be estimated from either the neural or acoustic representations:

$$SE(i) = \frac{1}{2} \sum_{j=1}^{N} \left( d_{WS1}(i, j) + d_{WS2}(i, j) - d_{BS1}(i, j) - d_{BS2}(i, j) \right)$$

where $d_{WS1}(i, j)$ and $d_{WS2}(i, j)$ are the distances between phonemes $i$ and $j$ of each speaker (within speaker distances), and $d_{BS1}(i, j)$ and $d_{BS2}(i, j)$ are the distances between phoneme $i$ and $j$ of different speakers (between speaker distances).

4.2 variations in acoustic cues of different speakers is encoded in the PRP

The previous chapter showed that the encoding of phonetic distinctions in the PRPs can be directly related to the acoustic characteristics of phonetic categories. However, in addition to deciphering the semantic message encoded in the speech signal, a listener also attends to acoustic cues that specify speaker identity. To study whether the variations in acoustic cues of different speakers is encoded in the PRP, we modified the pairwise phoneme similarity analysis by
estimating the pairwise distances between phonemes of each speaker, and between phonemes of different speakers. To measure speaker dependency of EEG responses, we subtracted the sum of the pairwise phoneme distances for each speaker and across speakers, an approach that highlights the PRP components that show a differential response between the two speakers. The correlations between the speaker distance matrices in acoustic and the speaker distance matrices in PRPs are shown in Figure 4.1 (top) where the most significant resemblance between speaker-dependent matrices occurs at R3 (200ms, r = 0.46, p<0.01). This observation differs from the timing of the largest phonetic distinctions in the PRP observed at R2, showing a significant time differences in the encoding of different acoustic features. The scalp location of speaker feature encoding is shown in Figure 4.1 (bottom).
We used a multidimensional scaling analysis to visualize the relative distance of the PRPs estimated separately for each speaker. As shown in Figure 4.2, speaker-dependent characteristics (indicated by white and black fonts) are secondary to phonetic features (indicated by colored bubbles), meaning that the phonetic feature distinctions in the PRP are greater than speaker dependent differences. We quantified this preferential encoding using silhouette index $^70$, and found a silhouette index significantly greater for the PRP clusters corresponding to manner of articulation compared to the PRP clusters that represent speaker differences (Silhouette index = 0.18 vs. 0.001).

Figure 4.2 MDS diagram of phonemes and speakers.
4.3 Speaker representation in neural responses of different phonemes

Next, we wanted to examine the encoding of the variable degree of acoustic similarity between different phonemes of the two speakers. This varied acoustic similarity is caused by the interactions between the physical properties of the speakers’ vocal tracts and the articulatory gestures made for each phoneme. To test the dependence of speaker representation in neural responses on different phonemes, we defined an index (Speaker Encoding, SE) that measures the response similarity between the phonemes of the two speakers. Therefore, this index would be zero if the responses to the same phonemes of two speakers were identical. We compared speaker dependent phoneme distances in acoustic and neural signals (Figure 4.3, r = 0.46, p<0.014), where the high correlation value implies a joint encoding of speaker-phoneme pairs. Our analysis shows that the separation between the two speakers is higher in the group of vowels.

Figure 4.3 Joint encoding of phonemes and speakers.
To more explicitly study speaker representation of vowels, we found the average PRPs for vowels for each of the two speakers. The average vowel PRPs of the two speakers show a significant separation at approximately 200 ms after the phoneme onset (corresponding to R3, Figure 4.4).

![Figure 4.4 PRP of vowels for speaker 1 versus 2.](image)

To visualize vowel separation at this time interval, in Figure 4.5, we used three-dimensional MDS diagram where the separation between the two speakers is readily observable. We quantified the separation of speakers within the group of vowels using the silhouette index\(^7\) (Figure 4.5), which revealed greater separation within the group of vowels in comparison to the separation of speakers in all PRPs.
Finally, Figure 4.6 summarizes the scalp location and timing for the three main analyses in our study; 1) the average PRP of all phonemes (green), 2) response components corresponding to acoustic phoneme similarity patterns (blue), and 3) response components correlated with speaker differences (red). The largest average PRP component appears at R1, maximum phonetic distinctions are encoded at R2, and speaker dependency was best represented at R3.
Figure 4.6 Comparison of timing and location for phonetic and speaker encoding.
4.4 Discussion

We observed that EEG responses to continuous speech reliably encode phonetic and speaker distinctions at multiple time intervals relative to the onset of the phonemes. The responses are primarily organized by phonetic feature, while subtler speaker variations appear within manner of articulation groups, consistent with previous studies showing a larger role for phonetic over speaker characteristics in shaping the acoustic properties of phones \cite{71,72}.

Our finding of repetitive appearance of phonetic distinction in the neural response is consistent with auditory evoked potential (AEP) studies of isolated consonant-vowel pairs \cite{12-15,73}. However, relating the PRP components (R1 to R4) to specific AEP events such as P1-N1-P2 or N2-P3-N4 complex requires further investigation. A challenge of this comparison is caused by the differences in the shape of PRP and AEP responses, including the sign of the deflection. For example, the sign of PRP deflection for different manner groups is not always positive-negative-positive, as is the case in AEP. In particular, R2 deflection is positive for plosive phonemes and negative for the vowels. Possible reasons for the observed differences between AEP and PRP is the dominance of onset response in AEP which was excluded in the calculation of PRPs, in addition to contextual effects which may influence the average responses to a particular phoneme. In addition, continuous speech is likely to engage higher level, speech-specific regions that may not be activated when a person hears isolated consonant-vowel tokens \cite{21,74}.

While our observation of scalp distributions at each time point suggest a different underlying pattern of neural activity for each component, the neural sources contributing to R1 to R4 remains unclear. In the rich literature of AEP, many studies have shown that AEP can be subdivided into three categories where the responses with latency less than 10ms are associated with brain-stem, response latencies between 10 to 50 ms are associated with thalamic regions and response latencies
beyond 50 ms are mostly generated by cortical regions\textsuperscript{75,76}. Within cortical responses, comparison of high-gamma and AEP\textsuperscript{77,78} as well as attention and development studies\textsuperscript{79–82} has shown that different cortical regions are responsible for generating P1, N1, P2, N2 and N4. Based on these findings, it is possible that the diverse timing of the observed components of PRP could be the combined effect of the activity of several cortical regions. The pairing of source connectivity analysis along with complementary neuroimaging techniques should allow for more detailed characterizations of neural processes in future studies\textsuperscript{83}. Additionally, systematically manipulating the stimulus, task and behavior may yield better characterization of the sensory and perceptual processes contributing to the representation of the acoustic features we observed at different time intervals\textsuperscript{17–19}.

One major difference between our study and previous work is the direct comparison between the organization of neural responses and acoustic properties of speech sounds. Therefore, the neural encoding of acoustic features can be investigated at each time point which may represent the underlying stages of cortical processing. In contrast with regression-based approaches\textsuperscript{27} which average neural responses over the duration of phonemes, our approach maintains the precise temporal features of the neural response. Our results lays the groundwork for several research directions, where explicit changes in the representational properties of speech can be examined in speech development\textsuperscript{84}, phonotactic probabilities in speech\textsuperscript{85}, contexts where a listener learns new acoustic distinctions\textsuperscript{86,87}, second language acquisition\textsuperscript{88,89}, and changes in the representational properties of speech through varying task demands\textsuperscript{90}. Given that the N1 and P1 sequences in AEP are not fully matured in children and teenagers, it remains to be seen how this can change the PRP components we report in this thesis\textsuperscript{14,80,91}. The ability to directly examine the representational properties of the spoken language stimulus in neural responses is a powerful tool for distinguishing
between the many factors involved in sensory processing \(^92,93\). For example, speech and communication disorders can be caused by a loss of linguistic knowledge, or as a result of a degraded representation of relevant acoustic cues, such as in disorders of the peripheral and central auditory pathways. The source of the problem is unclear for speech disorders, such as aphasia \(^94–\)
\(^96\). Since phoneme-related potentials can track the representational properties of speech as it is processed throughout the auditory pathway, these potentials could be instrumental in comparing healthy and disordered brains and identifying possible problem sources.
5 Adaptation enhances phonetic distinctions

Portions excerpted from:

5.1 Method

Intracranial recordings

Eight adults (five females) with pharmacoresistant focal epilepsy were included in this study. Subjects 1 to 6 were presented with the complete noisy speech task. Subjects 7 and 8 were presented with the visual distraction task. All subjects underwent chronic intracranial encephalography (iEEG) monitoring at North Shore University Hospital to identify epileptogenic foci in the brain for later removal. Six subjects were implanted with stereo-electroencephalographic (sEEG) depth arrays, one with grids and strip arrays, and one subject with both (PMT, Chanhassen, MN, USA). Electrodes showing any sign of abnormal epileptiform discharges, as identified in epileptologists’ clinical reports, were excluded from the analysis. All included iEEG time series were manually inspected for signal quality and were free of interictal spikes. All research protocols were approved and monitored by the institutional review board at the Feinstein Institute for Medical Research, and informed written consent to participate in research studies was obtained from each subject before implantation of electrodes.

Intracranial EEG (iEEG) signals were acquired continuously at 3 kHz per channel (16-bit precision, range ± 8 mV, DC) with a data acquisition module (Tucker-Davis Technologies, Alachua, FL, USA). Either subdural or skull electrodes were used as references, as dictated by recording quality at the bedside after online visualization of the spectrogram of the signal. Speech signals were recorded simultaneously with the iEEG for subsequent offline analysis. The amplitude of the high-gamma response (75–150 Hz) was extracted using the Hilbert transform and was resampled to 100 Hz. The high-gamma responses were normalized based on the responses recorded during a 2-minute silent interval before each recording.

Stimulus
Speech material was short stories recorded by four voice actors (two male and two female voice actors; duration: 20 minutes, 11025 Hz sampling rate). The three noises were taken from the NOISEX-92 corpus. Different three- or six-second segments of the noise were chosen randomly for each transition and were added to the speech at a 6 dB signal-to-noise ratio (noisy speech task). The SNR of 6 dB was chosen to ensure the intelligibility of foreground speech. In three of the subjects, we ran an additional task after the adaptation task, where they listened to the same speech utterances without the additive noises (clean speech task).

All stimuli were presented using a single Bose SoundLink Mini 2 speaker situated directly in front of the subject. To reduce the inevitable acoustic noise encountered in uncontrolled hospital environments, all electrical devices in patients’ room were unplugged except the recording devices and the door and windows were closed during the experiment to prevent interruption. We also recorded the clean speech task without the noise in three of the subjects for direct comparison of neural responses in the same hospital environment. Speech volume was adjusted to a comfortable listening volume.

**Speech-specificity task**

To quantify the speech specificity of each neural site, four of the subjects (subjects 1, 2, 4, and 6) performed the speech-nonspeech task. Subjects listened to 30 minutes of audio containing 69 commonly heard sounds. The sounds consisted of coughing, crying, screaming, different types of music, animal vocalization, laughing, syllables, sneezing, breathing, singing, shooting, drum playing, subway noises, and speech by different speakers. To determine the speech-specificity index, we first normalized the response of each site using the mean and variance of the neural data during the silent interval. We then averaged the normalized responses over the presentation of each
sound. Finally, we performed an unpaired t-test between the averaged responses of all speech and all nonspeech sounds to obtain a t-value for each site denoting the selectivity to speech over nonspeech sounds.

**Behavioral task**

12 subjects (seven males, five females) with self-reported normal hearing participated in this experiment. The task consisted of six consonant-vowel pairs (CVs, /pa,ta,ka,ba,da,ga/) spoken by two male and two female speakers (a total of 24 tokens). The tokens were embedded in changing background noise identical to the main speech in the noise experiment. Half of the CVs were uttered immediately after the transition to a new background noise (during adaptation, DA), and the other half of the CVs were uttered 1.5 s after transition (after adaptation, AA). Noises were added to CVs at SNR of -4 dB. The task was presented to the subjects using Matlab. The participants responded via a multiple-choice graphical user interface (GUI) in Matlab that included the six CVs in addition to an unsure option. Subjects were required to report the CV continually and were all able to keep up with the rapid speed of CV presentation. All subjects provided written informed consent. The Institutional Review Board (IRB) of Columbia University approved all procedures.

**Stimulus reconstruction**

We used a linear model to map the neural responses (R) to the auditory stimulus (S). We trained the model on clean speech that was played to the subject after the noisy speech experiment. We used time lags from -250 to 0 ms of the neural data as the input to the ridge regression (R). The model (g) is calculated by minimizing the MSE between reconstructed and original spectrograms,
which results in the cross-correlation of the stimulus and the ECoG data normalized by the autocorrelation of the ECoG data.

**Spectrotemporal receptive fields**

STRFs were computed by normalized reverse correlation algorithm\textsuperscript{101} using STRFLab\textsuperscript{101}. Regularization and sparseness parameters were found via cross-validation. The best-frequency and response latency parameters were estimated by finding the center of the excitatory region of STRF along frequency and time dimensions. The best-rate parameter was estimated from the 2-dimensional wavelet decomposition of the STRF\textsuperscript{47,102}. The wavelet decomposition extracts the power of the filtered STRFs at different temporal modulations (rates)\textsuperscript{47,102}. The modulation model of STRFs has four dimensions: scale, rate, time, and frequency. To estimate rate, we first averaged the model over three dimensions of time, frequency, and scale to calculate a rate vector. Next, we found the weighted average of the rate vector, where weights are the rate values.

**Phoneme responses**

We segmented the speech material into time-aligned sequences of phonemes using the Penn Phonetics Lab Forced Aligner Toolkit\textsuperscript{51}, and the phoneme alignments were then manually corrected using Praat software\textsuperscript{103}. The spectrograms were aligned to the onset of phonemes with a time window of 200 ms. To minimize the preprocessing effects, we did not normalize the natural variation in phoneme length. The phoneme pairwise distances were calculated based on the Euclidean distance between each pair of phonemes.
5.2 Neural adaptation to changing background condition

We recorded electrocorticography data from six human subjects implanted with high-density subdural grid (ECoG) and depth (stereotactic EEG) electrodes as a part of their clinical evaluation for epilepsy surgery. One subject had both grid and depth electrodes, four subjects had bilateral depth electrodes, and one subject had only grid electrodes. The location of the electrodes is shown in Figure 5.1.

![Figure 5.1 Electrodes with significant response to speech. (t-val>10, t-test speech versus silence)](image)

The location of electrodes for individual subjects of one to six is shown in Figure 5.2. Subjects 1, 3, 4 and 5 have bilateral depth electrodes. Subject 2 has only grid on the left hemisphere. Subject 6 has both grid and depth electrodes only on the left hemisphere.
Figure 5.2 Location of grid and depth electrodes on MRI of each individual subject.
Subjects listened to 20 minutes of continuous speech by four different speakers (two male speakers and two female speakers). The background condition changed randomly every three or six seconds between clean (no background noise), jet, city, and bar noises were added to the speech at a 6 dB signal-to-noise ratio as shown in (Figure 5.3).

These three types of background noise were chosen because they represent a diversity of spectral and temporal acoustic characteristics as evident by their average of frequency, non-stationariness and rate. The properties of each noise is shown in Figure 5.4, where part “A” shows the weighted average (center of gravity) of frequencies for four conditions of jet, city, bar and clean computed from the acoustic spectrograms. It shows that

\[ \text{Average Frequency of Jet} > \text{City} > \text{Bar} > \text{Clean} \]

In part “B” and “C” of Figure 5.4, we defined the non-stationariness by finding the standard deviation of each frequency band over time and then averaging across frequencies. The higher standard deviation indicates higher nonstationarity of the condition. The difference between stationarity of the conditions is significant and varies in the following order:

\[ \text{stationariness of Jet} > \text{Bar} > \text{City} > \text{Clean speech} \]
In part “D” and “E”, we show the Frequency-rate model of each of the background conditions and clean speech. We show that Jet has the highest rate across conditions, and clean has the lowest rate.

In part “F”, We computed the degree of speech masking for each condition by calculating the overlap between the frequency profile of each noise and clean speech:

\[
\text{masking}_{\text{bar}} > \text{masking}_{\text{city}} > \text{masking}_{\text{jet}}
\]

Figure 5.4 Spectrotemporal properties of each background condition.
As it is evident from the acoustic properties shown in Figure 5.4, the jet noise has high frequency and high temporal modulation power, the city noise has uniformly distributed power over frequencies, and the bar noise has mostly low-frequency power. In total, there were 294 transitions between background conditions, distributed evenly among the four conditions. The background noise segments were not identical and were randomly taken from a minute-long recording. To ensure that the subjects were engaged in the task, we paused the audio at random intervals and asked the subjects to report the last sentence of the story before the pause. All subjects were attentive and could correctly repeat the speech utterances. All subjects were fluent speakers of American English and were left-hemisphere language dominant (as determined with Wada test).

We extracted the envelope of the high-gamma band (75–150 Hz), which has been shown to reflect the average firing of nearby neurons. For all analyses, the electrodes were selected based on a significant response to speech compared with silence (t-test, false discovery rate [FDR] corrected, p<0.01). This criterion resulted in 167 electrodes in perisylvian regions, including Heschl’s gyrus (57 electrodes), the transverse temporal sulcus (12 electrodes), the planum temporale (26 electrodes), and the superior temporal gyrus (STG, 39 electrodes), from both brain hemispheres (97 left, 70 right).

To study how the neural responses to speech are affected when the background condition changes, we aligned the responses to the time of the background change and averaged over all transitions to the same background condition. The average response in Figure 5.5A shows a short-term transient peak, which occurs immediately after the background changes (average duration = 670 ms, t-test, FDR corrected, t-test, p<0.01). This short-term response appears in all four conditions, even in the transition to the clean condition (e.g., from speech with jet noise to clean...
speech). Figure 5.5A also illustrates that the selectivity and magnitude of this adaptive response to different background conditions varies across neural sites.
5.3 Adaptation suppresses the representation of noise features

To study what features of the acoustic scene are encoded in the responses over the time course of adaptation, we used the method of stimulus reconstruction\textsuperscript{106,107}. Reconstruction methods typically rely on learning the linear mapping that relates evoked neural population responses to a time-frequency (spectrogram) representation of sound. This method enables a direct comparison between original and reconstructed spectrograms, making it possible to analyze what spectrotemporal features are encoded at the neural population level.

We first trained the reconstruction model on responses to clean speech without the added background noise for each subject separately and used it to reconstruct the spectrograms from the neural responses to speech with added background noises. The reconstructed spectrograms were then averaged across all subjects. We aligned the original and reconstructed spectrograms to the time of the background changes and calculated averages over all trials that shared the same new condition. Comparison of the average original (Figure 5.5B) and reconstructed (Figure 5.5C) spectrograms shows that immediately after a transition, the neural responses encode the acoustic features of the background noise, which can be seen from the similarity of the reconstructed and original spectrograms after a transition (e.g., the high-frequency energy in jet noise or the low-frequency energy in bar noise after transition). The acoustic features of noises in the reconstructed spectrograms, however, fade away quickly when the adaptation is over, resulting in a noise-invariant representation of speech sounds.
To better illustrate this noise-invariant representation, in Figure 5.6A, we show the temporal shape of reconstructed spectrograms by averaging over their frequency dimension. For comparison, we also presented the same speech materials to the subjects but without any added background noises (clean speech task). The similarity of average reconstructions from the responses in noisy speech task and clean speech task after the adaptation interval is shown in Figure 5.6. Additionally, we also directly compared the neural responses in noisy and clean speech task and observed the same initial transient divergence of responses after transitioning to a new noise which then converged to the neural response to clean speech task after the adaptation interval (Figure 5.6B).
To illustrate the appearance of the spectral features of noise more explicitly, we averaged the reconstructed and the original spectrograms over two time intervals, during adaptation (DA, 0–0.39 s after transition) and after adaptation (AA, 2–2.39 s after transition), and we normalized each to its maximum value. We defined the adaptation interval for the reconstructed speech by comparing the envelope of the reconstructed and clean spectrograms (average duration = 390 ms, \( t \)-test, \( p<0.01 \)). For comparison, Figure 5.7A shows the average frequency power from the original spectrograms. Figure 5.7B shows that the average reconstructed frequency profile during adaptation resembles the frequency profiles of the noises (\( R^2 = 0.64 \) using 5–fold crossvalidation for each condition, \( t \)-test, \( p<10^{-6} \)). However, the average reconstructed frequency profile after adaptation in all three noise conditions converges to the frequency profile of clean speech

\[ R^2 = 0.91 \text{ using 5–fold crossvalidation for each condition, } t \text{– tes, } p<10^{-6}. \]
Figure 5.7 Average frequency profile of the conditions.

Figure 2c also shows this shift for individual trials. We quantified the time course of this effect by measuring the coefficient of determination ($R^2$) between reconstructed spectrograms with both original noisy and original clean spectrograms over time. Each dot is the average of five trials. Trials during the adaptation interval (DA, 0-0.4 s after transition) are shown with orange, and trials after the adaptation interval (AA, 2.0-2.4 s after transition) are shown with blue.

Figure 5.8 $R^2$ for individual trials.

In addition, the degree of overlap between the reconstructed spectral profile during adaptation (DA) and the spectral profile of clean speech varies across noises, as quantified by $R^2$ between
reconstructed and clean speech spectrograms in Figure 5.9A. The overlap was highest for the bar noise and lowest for the jet noise, meaning that during the adaptation phase, the bar noise masks the acoustic features of clean speech more than the jet noise does. This difference is a direct result of acoustic similarity between bar noise and clean speech. The \( R^2 \) differences over time are shown in Figure 5.9B, where they show an average time of switching between the similarity of reconstructed spectrograms from noisy to clean at 420 ms (std = 70 ms). This finding shows a brief and significant decrease in the signal-to-noise ratio (SNR) of the representation of speech in the auditory cortex as the neural responses are undergoing adaptation, but the SNR is subsequently increased after the adaptation is over.

![Figure 5.9 Speech masking for each condition.](image)

To confirm the effect of decrease in the signal-to-noise ratio (SNR) of the representation of speech we calculated the time course of the coefficient of determination \( (R^2) \) of reconstructed spectrograms with respect to (wrt) original noisy (orange) and with respect to original clean (black) spectrograms (20% cross-validations, \( n = 15 \)) for individual subjects. As it is show in Figure 5.10 this effect is consistent across subjects.
Moreover, we confirmed that the decreased response to background noise is not due to the lack of responsiveness of electrodes to the noise stimulus relative to speech\(^{33}\). We played the noises and speech separately in four of the subjects (117 electrodes). The goal of this experiment was to test whether the adaptation patterns are caused by the non-responsiveness of electrodes to noise stimulus relative to speech (Figure 5.11). We observed that unlike the noisy speech task, the neural responses to the noise stimuli continued over time, showing sustained response to noise-only stimuli when it was presented to the subject without adding the foreground speech (t-test, \(p<0.001\)). This means that the suppression of the background noise is not an inherent tuning property of the neural response to the noises and instead is contingent upon the presence of foreground speech\(^{108,109}\). Figure 5.11B shows the comparison of the average neural response when transitioning from silence to noise-only stimulus (purple) with the average neural response when
transitioning from clean speech to speech in noise (green). Neural responses are normalized based on pre-transition intervals. Right: comparison of the baseline change after adaptation interval (2 to 3 sec after transition) relative to the pre-transition interval (-0.5 sec to 0 sec) in the two conditions.

**Figure 5.11** The neural response to the noise and speech alone.
5.4 Adaptation enhances phonetic distinctions

The reconstruction analysis showed the encoding of spectrotemporal features of the stimulus in the population neural responses. Speech, however, is a specialized signal constructed by concatenating distinctive units called phonemes, such as the /b/ sound in the word /bad/\(^1\). In addition, the human auditory cortex has regions specialized for speech processing that respond substantially more to speech than to other sounds\(^{110,111}\). Using a separate speech-nonspeech task, we also found many electrodes that responded significantly more to speech than to nonspeech sounds.

Figure 5.12 shows that out of 117 electrodes, 54 (45%) responded significantly (FDR corrected, p<0.05) more to speech over nonspeech sounds (speech-specific electrodes). The remaining 55% of electrodes responded equally to both speech and nonspeech sounds (speech-nonspecific electrodes). The majority of speech-specific electrodes were located on the left superior temporal gyrus, the same area that showed a stronger transient response when the background noise stopped.

We therefore extended the spectrotemporal acoustic feature analysis to explicitly examine the encoding of distinctive features of phonemes during and after adaptation intervals. To examine how the cortical representation of phonetic features is affected when the background condition changes, we segmented the original and reconstructed spectrograms into individual phonemes and averaged the spectrograms of phonemes that occurred in the time intervals of during (DA) and after adaptation (AA). Figure 5.13 shows the original and reconstructed spectrograms of three example phonemes. The distinctive spectrotemporal features of these phonemes in reconstructed spectrograms are distorted during adaptation but are significantly enhanced afterward. For example, the phoneme /b/ is characterized by an onset gap followed by low-frequency spectral power. Both the gap and the low-frequency feature are masked during adaptation but are
subsequently restored after adaptation. Another example is the vowel /ih/, which is characterized by its first two formant frequencies. The frequency peaks of /ih/ vowels that occurred after the adaptation interval are enhanced compared to those during the adaptation interval.

Quantifying the similarity of the reconstructed phoneme spectrograms during and after adaptation with the clean phoneme spectrograms shows a similar effect (Figure 5.14A). Furthermore, using the high-gamma activity, we examined the relative distances between the neural representations of phonemes during and after adaptation. We generated a phoneme dissimilarity matrix, which summarizes the pairwise correlation between all phoneme pairs. We found that the relative phoneme distance in the neural responses collapses during adaptation but is significantly increased after the adaptation interval (Figure 5.14B).
The discriminability of different reconstructed phonetic features is also reduced during adaptation to a new background condition but is increased thereafter. In Figure 5.15, we quantified the discriminability of all phonemes using the ratio of between-group to within-group variability (F-statistics) for five phonetic features of manners of articulation (plosives, fricatives, vowels, nasals, and approximants). In Figure 5.15B, we performed the same analysis as in neural responses to speech without the changing background conditions, and we did not observe a suppression in pairwise distances of phonemes.

Figure 5.14 Phoneme dissimilarity matrices in during adaptation versus after adaptation.
Motivated by this observation, we designed a psychoacoustic task to study the perception of phonetic features during and after adaptation intervals (Figure 5.16). The task consisted of six consonant-vowel pairs (CVs, /pa,ta,ka,ba,da,ga/), chosen to cover a wide range of frequency profiles (low-frequency labials /pa,ba/, mid-frequency velar /ka,ga/, and high-frequency alveolar
The CVs were embedded in the same changing background noise as the speech in the noise task. Half of the CVs were uttered during adaptation to a new background noise (DA, colored in orange), and the other half of the CVs were uttered after the adaptation interval was over (AA, colored in blue).

**Figure 5.16** Behavioral experiment designed to test the intelligibility of consonant-vowel pairs.

The task consisted of 200 phonemes randomly assigned to the during adaptation (DA) and after adaptation (AA) intervals. Subjects had to click on the consonant-vowel pair (CV) that they perceived as soon as the CV ends. The time interval between each CV was 1.5 s. The task that was given to the subject is shown in Figure 5.17.
We observed that the recognition score of the CVs during adaptation was significantly lower than that of the CVs after adaptation in all 12 subjects as shown in Table 2.

<table>
<thead>
<tr>
<th>Subject</th>
<th>During adaptation</th>
<th>After adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6000</td>
<td>0.6900</td>
</tr>
<tr>
<td>2</td>
<td>0.5700</td>
<td>0.6500</td>
</tr>
<tr>
<td>3</td>
<td>0.5000</td>
<td>0.5600</td>
</tr>
<tr>
<td>4</td>
<td>0.5600</td>
<td>0.6100</td>
</tr>
<tr>
<td>5</td>
<td>0.6500</td>
<td>0.7100</td>
</tr>
<tr>
<td>6</td>
<td>0.4800</td>
<td>0.6200</td>
</tr>
<tr>
<td>7</td>
<td>0.3100</td>
<td>0.3900</td>
</tr>
<tr>
<td>8</td>
<td>0.5600</td>
<td>0.7100</td>
</tr>
<tr>
<td>9</td>
<td>0.1458</td>
<td>0.2188</td>
</tr>
<tr>
<td>10</td>
<td>0.1649</td>
<td>0.1856</td>
</tr>
</tbody>
</table>
We combined all subjects and observed that there is a significant increase after adaptation as shown in Figure 5.18 (Yate’s corrected chi-square test, p<0.01). The match between neural and perceptual degradation of phonetic features during adaptation suggests an important role for neural adaptation in enhancing the discriminative features of speech that may ultimately contribute to the robust perception of speech in noise in humans.  

Table 2. Individual subject’s phoneme recognition score.

<table>
<thead>
<tr>
<th></th>
<th>Recognition accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>0.5200</td>
</tr>
<tr>
<td>12</td>
<td>0.5500</td>
</tr>
</tbody>
</table>

Figure 5.18 Comparison of correctly recognized phonemes occurring during adaptation (DA) and after adaptation (AA) intervals.

Figure 5.19 shows the behavioral effect of adaptation in each background condition separately. The improvement in the recognition accuracy is highest for bar and lowest for the jet noise. This difference correlates well with the masking of speech features in the neural responses during adaptation to each noise. Interestingly, we also found an increase in the recognition rate after adaptation to the clean condition, meaning that phoneme recognition accuracy is also decreased immediately after the noise stops, similar to the findings of forward masking studies.
Figure 5.19 Comparison of correctly recognized phonemes for each condition separately.
6 Effect of brain regions and attention in adaption patterns

Portions excerpted from:

6.1 Method

Brain maps

Electrode positions were mapped to brain anatomy using registration of the postimplant computed tomography (CT) to the preimplant MRI via the postop MRI\textsuperscript{114}. After coregistration, electrodes were identified on the postimplantation CT scan using BioImage Suite\textsuperscript{115}. Following coregistration, subdural grid and depth electrodes were snapped to the closest point on the reconstructed brain surface of the preimplantation MRI. We used the FreeSurfer automated cortical parcellation\textsuperscript{116} to identify the anatomical regions in which each electrode contact was located within approximately 3 mm resolution (the maximum parcellation error of a given electrode to a parcellated area was < 5 voxels/mm). We used Destrieux’s parcellation, which provides higher specificity in the ventral and lateral aspects of the medial lobe\textsuperscript{117}. Automated parcellation results for each electrode were closely inspected by a neurosurgeon using the patient’s coregistered postimplant MRI.

Calculating adaptation indices

To characterize the adaptation index (AI), we measured the t-value of a paired t-test between the neural response of each neural site in time intervals of 0 to 0.7 s (during adaptation, DA) and 2 to 2.7 s (after adaptation, AA) after the transition to each background condition (time 0). AIs were normalized by subtracting the minimum over all conditions, followed by a division by their sum.

Visual attention task

To control for the effect of attention on the adaptation patterns, we designed a visual experiment which we tested on two subjects (subject 7 and 8). We used 10 minutes of the adaptation task and
presented it in two conditions to the subject: I) when the subject was engaged in the visual task (auditory distracted) and II) without the visual task (auditory attended). The subjects were presented with the distracted condition first where they were asked to perform a visual search task and ignore the sound (the speech in noise task) that was presented simultaneously. The visual search task was a two-choice test. The subjects had a maximum of seven seconds to answer each question and had to count either the number of colors (Question 1) or the number of shapes (Question 2) and respond whether the answer shown on the screen was right or wrong. In the visual condition, subject 1 could answer 132 questions in 10 minutes with 74% accuracy, and subject 2 could answer 164 questions in 10 minutes with 87% accuracy. In the Attended experimental condition, the subject attended to the sound (the adaptation task) without the visual secondary task. To control for the possible confounding effect of visual stimulus in the distracted experimental condition, we asked the subject to fixate on the visual search task while different questions were shown, but the subject was not required to answer any of those questions. To measure the efficacy of the visual task in engaging the attention of the subject, at the end of each experimental block we asked the subjects contextual questions about the speech stories. The subject had three options: 1) Right, 2) Wrong and 3) Unsure. The total number of questions was 72.

6.2 Adaptation magnitude varies across neural sites

Our analysis in the previous sections focused on the encoding of the acoustic features of speech and noise by the population of neural sites. To examine how individual electrodes respond when the background condition changes, we first compared the magnitudes of the responses during (DA) and after adaptation (AA) by pooling electrodes across all subjects. We found variable numbers of electrodes with significant response changes during transitions to different background
conditions (104 for jet, 120 for city, 122 for bar, and 78 for clean conditions, t-test, FDR corrected, p<0.05). Figure 6.1A shows the Mean squared error (MSE) between the high-gamma responses to clean and noisy speech over time. The horizontal bars on top of the plot highlight the time points where at least 25% of electrodes showed a significant difference between the two conditions (corrected for false discovery rate, q<0.01). Figure 6.1B shows the distortion effect on single electrodes for each of the conditions. The x-axis shows the z-scored high-gamma response after adaptation, the y-axis shows the z-scored high-gamma response during adaptation intervals, and electrodes with significant distortion are colored (paired t-test, p<0.01).

Figure 6.1 Comparison of adaptation time and electrodes for different background conditions.
We also found 16 electrodes that showed no significant transient response to any of the background conditions, even though these electrodes were similarly responsive to speech (t-test between responses to speech vs. silence, FDR corrected, p<0.01).

To explain the variability of adaptive response patterns across electrodes, we first defined an adaptation index (AI) as the t-value of a paired t-test between the magnitude of the responses during and after adaptation intervals. The AI for each electrode is calculated for each background condition and is normalized by subtracting the minimum over all conditions. We performed unsupervised hierarchical clustering (minimum variance algorithm, Euclidean distance) on AIs to group electrodes based on the similarity of their adaptive responses across the four background conditions (Figure 6.2A). Comparison of the top two clusters of electrodes in Figure 6.2B shows that the primary difference between adaptation patterns is the presence or absence of an adaptive response in transition to the clean condition (e.g., transition from speech in jet noise to clean speech). The secondary factor that further separates electrodes is their selective adaptation to different background noises. This is evident in the average responses of electrodes in each cluster, shown in Figure 6.2C. For example, the first three clusters all show minimal adaptation to the clean condition but have significant adaptation to jet, city, and bar noise, respectively. Clusters 4 and 5, on the other hand, show significant adaptation to the clean condition.
Figure 6.2 Diversity of adaptation patterns across electrodes.

In addition, we checked the adaptation patterns across individual subjects. In Figure 6.3, the adaptation patterns are plotted for each subject in transition to four background conditions. The plots for each condition show the average response of all electrodes with significant adaptation to that condition. The percentage of electrodes that showed the highest adaptive response to each background condition is shown on the right. As an example, in subject 1, 11% of adaptive electrodes are most adaptive to jet, 8% are most adaptive to city, 24% are most adaptive to clean, and 58% are most adaptive to bar noise.
Furthermore, we show that the latency of the response is significantly higher for clusters of electrodes with high adaptation to clean conditions and nonadaptive clusters. The two tiers shown in Figure 6.4 refers to the top clusters that are shown in Figure 6.2. Given that the response latency approximates the number of synapses between the auditory periphery and the neural site, this
suggests that nonadaptive sites and sites with larger adaptation responses to clean condition are in
the higher processing stages of the auditory pathway.

Figure 6.4 Average latency of electrodes from the top two clusters
6.3 Adaptation is unaffected by the attentional focus

To examine the effect of subject’s attention on adaptive responses to background noise changes, we engaged the subjects in a demanding visual task as they heard the speech in noise sounds. As it is shown in Figure 6.5 subjects had to count the number of shapes and colors in the picture and respond whether the equation that was shown to them is correct or not.

![Visual distraction task](image)

**Figure 6.5 Visual distraction task.**

We then repeated the speech in noise task without the visual task and asked the subjects to attend to speech instead. Figure 6.6 shows a significant difference between the speech comprehension accuracy with and without the secondary visual task and confirms the efficacy of the secondary task in distracting the subjects from the auditory task. Figure 6.6 shows the subjects’ behavioral response accuracy in response to the contextual questions from speech stories in two conditions of Distracted (Black) and Attended (orange) (mean ± s.e.m., N_{distracted} = 48, N_{attended} = 96, *** Yate’s corrected chi-square test, p<0.01).
Despite the large difference between the attentiveness of the subjects to the auditory task in these two experimental conditions (Yate’s corrected chi-square test, $p<0.001$), we did not observe a significant difference between the neural responses to speech and the adaptation patterns with and without the secondary visual task. As shown in Figure 6.7, the average high-gamma activities are the same for distracted versus attended condition.
Figure 6.7 Average high-gamma activity during distracted versus attended condition.

The average high-gamma activity for each individual subject also shows the same similarity between neural activities in distracted versus attended conditions. In Figure 6.8, the responsive electrodes to speech were selected for subjects 7 and 8. The average of all responsive electrodes in transition to a new noisy condition is shown for both attended (red) and distracted (blue) tasks. In addition to the average of electrodes, we also did not find any significant difference between the adaptation patterns of attended and distracted at the level of single electrodes.
Figure 6.8 Comparison of neural adaptation in attended and distracted conditions for individual subjects.

In conclusion, the similarity of responses in the two attention conditions suggests that adaptation to changing background noise may primarily be a bottom-up phenomena $^{30,31}$. 
6.4 Spatial organization of the adaptation patterns

We examined the spatial organization of the adaptive responses to different background conditions. Figure 6.9 shows the AIs of individual electrodes on an average Freesurfer brain.

![Image showing spatial organization of adaptive responses to each background condition.](image)

Figure 6.9 Spatial organization of adaptive responses to each background condition.

In Figure 6.10 we magnified the perisylvian region and computed the histogram of adaptation indices by averaging the neighboring electrodes. Each pixel in Figure 6.10 is a 2 mm x 2 mm square, and the color of the square at each location is chosen based on the maximum AI at that location across the four background conditions. It shows that adaptation to jet noise is
strongest in the medial (deep) electrodes on both hemispheres, while adaptation to bar noise is stronger in the lateral (superficial) electrodes.

The spatial organization of adaptive responses is largely due to the spatial organization of tuning properties. Figure 6.11(left) shows the correlation between the tuning parameters and medial-lateral distance of electrodes (ML distance). Figure 6.11(right) shows the correlation between adaptation indices and ML distance of electrodes (error bar from depth electrodes, n=128).
Furthermore, an intriguing observation from Figure 6.10 is that electrodes with the largest adaptation to the clean condition are mostly located in the STG and in the left brain hemisphere. The spatial organization of the two tiers from the unsupervised clustering is also consistent with the spatial organization of the adaptation to the noises and to the clean because tier two mostly consists of electrodes that show the strongest adaptation in transition to the clean condition. Moreover, the stronger adaptation to the clean condition in higher-level cortical areas, such as the STG, is highly correlated with the spatial organization of speech specificity of electrodes ($r = 0.51$, $p < 10^{-9}$).

To study why neural sites adapt differently to the background conditions, we examined the relationship between adaptation patterns and both the spectrotemporal tuning and the speech specificity of electrodes. We characterized the tuning properties of an electrode by calculating its spectrotemporal receptive field (STRF). We measured two parameters from each STRF to describe the electrodes’ preferred frequency (best frequency) and preferred temporal modulation (best rate). The best-frequency parameter differentiates tuning to high versus low acoustic frequencies and is defined as the spectral center of the excitatory region of the STRF. The best-

Figure 6.11 Spatial organization of the tuning parameters.

Furthermore, an intriguing observation from Figure 6.10 is that electrodes with the largest adaptation to the clean condition are mostly located in the STG and in the left brain hemisphere. The spatial organization of the two tiers from the unsupervised clustering is also consistent with the spatial organization of the adaptation to the noises and to the clean because tier two mostly consists of electrodes that show the strongest adaptation in transition to the clean condition. Moreover, the stronger adaptation to the clean condition in higher-level cortical areas, such as the STG, is highly correlated with the spatial organization of speech specificity of electrodes ($r = 0.51$, $p < 10^{-9}$).

To study why neural sites adapt differently to the background conditions, we examined the relationship between adaptation patterns and both the spectrotemporal tuning and the speech specificity of electrodes. We characterized the tuning properties of an electrode by calculating its spectrotemporal receptive field (STRF). We measured two parameters from each STRF to describe the electrodes’ preferred frequency (best frequency) and preferred temporal modulation (best rate). The best-frequency parameter differentiates tuning to high versus low acoustic frequencies and is defined as the spectral center of the excitatory region of the STRF. The best-

85
rate parameter is measured from the modulation transfer function\textsuperscript{47} and differentiates tuning to slow and fast acoustic features. Figure 6.12A shows the computations that were performed to find the rate-scale model of STRFS. STRFs are passed through a 2D wavelet transform to find the temporal modulation and spectral modulation of the STRF. Figure 6.12B shows the example STRFs sorted based on their best rate values. STRFs with higher and lower rates are more sensitive to fast and slow acoustic changes, respectively.

Figure 6.12 Calculating rate and frequency tuning properties from STRFs of the electrodes.

In addition, we also measured the degree of speech specificity of the electrodes, defined as the t-value of a paired t-test between the responses of each electrode to speech and nonspeech sounds.
To study the contribution of each tuning dimension in predicting how an electrode responds in transition to a particular background condition, we used linear regression to predict AIs from the tuning parameters. Figure 6.13 shows the predictive power for each tuning parameter and the overall correlation between the actual and predicted AIs of each background condition. The AIs of all conditions except city (the least stationary noise) are highly predictable from electrodes’ response properties ($R^2_{\text{jet}} = 0.41, p < 10^{-12}; R^2_{\text{city}} = 0.02, p < 0.01; R^2_{\text{bar}} = 0.4, p < 10^{-12}; R^2_{\text{clean}} = 0.42, p < 10^{-12}$). It also shows that electrodes with tuning to higher frequencies also show higher adaptation to the high-frequency jet noise (positive main effect, 0.27). On the other hand, lower-frequency neural sites show higher adaptation to low-frequency bar noise (negative main effect, -0.46, t-test, p<0.001).Temporal modulation tuning of electrodes is positively correlated with the AI of jet noise (positive main effect, 0.38, t-test, p<0.001), which is also the condition with fastest temporal modulation. Temporal modulation (rate) is negatively correlated with the AI of the clean condition (negative main effect, -0.47, t-test, p<0.001), meaning that electrodes with a longer temporal integration window had the highest adaptive response in transition to the clean condition. The speech specificity of electrodes was positively correlated with the AI of the clean condition (positive main effect, 0.48, t-test, p<0.001), indicating that the electrodes that show the highest adaptation in transition from noisy to clean speech are the ones that also respond more selectively to speech over nonspeech sounds. Together, these results show that the adaptation patterns across electrodes are largely predictable from the response properties of those electrodes, such that electrodes that are tuned to the acoustic properties of a background condition also show the strongest adaptation to that condition.
Figure 6.13 R-squared values for predicting the adaptation indices
6.5 Discussion

In chapter 5 and 6, we examined the reduction in background noise in the human auditory cortex using invasive electrophysiology combined with behavioral experiments. We found that when a new background noise appears in the acoustic scene during speech perception, the auditory neural responses momentarily respond to noise features, but rapidly adapt to suppress the neural encoding of noise, resulting in enhanced neural encoding and perception of phonetic features of speech. We found a diversity of adaptation pattern across electrodes and cortical areas, which was largely predictable from the response properties of electrodes. Moreover, adaptation was present even when the attention of the subjects was focused on a secondary visual task.

Previous studies have shown that the auditory cortex in animals and humans encodes a noise-invariant representation of vocalization sounds\textsuperscript{30–36,119}. Our study takes this further by examining the dynamic mechanisms of this effect and how they change the representation of the acoustic scene as adaptation unfolds. Our finding of reduced neuronal responses to noise is consistent with studies that propose adaptation as an effective coding strategy that results in an enhanced representation of informative features when the statistical properties of the stimulus change\textsuperscript{120–122}. Although the adaptive encoding of a particular stimulus dimension has been shown in several subcortical\textsuperscript{123–126} and cortical areas\textsuperscript{30,31}, our study goes further by identifying the specific acoustic features of speech and background noise that are encoded by the neural responses over the time course of adaptation.

We found that the magnitude of adaptation to different background noises varied across neural sites, yet it was predictable from the spectrotemporal tuning properties of the sites. This observation was made possible by the sharp spectral contrast between the three background noises used in our study. This means that the neural sites whose spectral tuning match the spectral profile
of a particular noise also have a stronger adaptive response to that noise. We also found a population of neural sites that did not show any adaptation to the noises in our study, which could be due to the sparse sampling of the spectrotemporal space caused by the limited number of noises we used. In addition to the spectral overlap, previous studies have shown that separating an auditory object from a background noise that has a temporal structure requires integration over time\textsuperscript{127,128}. Experiments that systematically vary the temporal statistics of the background noise\textsuperscript{21} are needed to fully characterize the dependence of adaptation on the statistical regularity and the history of the stimulus\textsuperscript{129}.

We found that adaptation in the transition from noisy to clean speech occurred only in higher cortical areas, such as in the left-hemisphere STG. While previous studies have already established the specialization of the STG for speech processing\textsuperscript{110,111}, our finding uncovers a dynamic property of this area in response to speech. The magnitude of the adaptive response in transition to the clean condition was highly predictable from the speech specificity of electrodes, which is a nonlinear tuning attribute. It is worth mentioning that these sites were also highly responsive to foreign languages that were incomprehensible to the subjects. Therefore, the speech specificity of neural sites in our study is likely caused by tuning to speech specific spectrotemporal features and not by higher order linguistic structures\textsuperscript{40}. The transient response to the clean condition observed in the speech-specific electrodes may indicate adaptation of these sites to the unmasked features of speech, which reappear when the noise stops and indicate the recovery of speech-selective responses from their noise-adapted state\textsuperscript{130}. This result is also consistent with studies of the neural mechanism of forward masking, which has been reported in the auditory periphery\textsuperscript{131} and the auditory cortex\textsuperscript{130}, where the neural response to a clean target sound changes depending on the sound that preceded the target.
Using a behavioral paradigm, we show that the recognition of phonemes is degraded during the adaptation interval to a new background condition. Moreover, we found that the decrease in the phonetic feature recognition was greater when transitioning to a background noise that overlaps spectrally with speech, such as in the case of bar noise. This reduced phoneme recognition accuracy was consistent with the observed degradation of the phoneme representation in the neural data. This finding confirms the role that adaptation plays in enhancing the signal contrast with the background \(^{132}\), which results in an improved identification of its distinct features that are relevant for perception. Interestingly, we also observed a reduced behavioral accuracy in the perception of the phonemes when transitioning from a noisy background to the clean condition. This behavioral observation is consistent with the psychophysical studies of forward masking, where the detection of a target sound can be impaired by the preceding sound\(^{113}\), particularly when the acoustic properties of the noise and target overlap\(^{133}\).

We found that the strength of adaptation to background noises was stronger when listening to speech in noise compared to listening to noise alone. This means that the presence of speech was necessary for the observed suppression of noise features in the neural responses. The representation of speech in the human auditory cortex is also modulated by top-down signals, including the semantic context \(^{41,134-136}\) and attention \(^{137-139}\). It was therefore plausible that a momentary lapse in the subjects’ attention at the point of background switch can cause the transient neural responses we report here. Controlling for this possibility, we found that adaptation results are equally present even when the attention of the subject was directed towards a demanding secondary visual task. Although the behavioral performance of the subject during the auditory task significantly decreased with the added visual task, there was no detectable difference in adaptation patterns in the two experimental conditions. Moreover, While we used speech stories in the native
language of the subjects, our behavioral experiment showed a decrease in phoneme recognition accuracy even when nonsense speech (CVs) was used, suggesting that the enhanced effect of adaptation exists independent of linguistic context\textsuperscript{40,41}. As a result, the adaptation results we observed are likely due to bottom-up nonlinear mechanisms such as synaptic depression\textsuperscript{31,140} and divisive gain normalization\textsuperscript{30,141}. These mechanisms can separate an acoustic stimulus with rich spectrotemporal content, such as speech, from the more stationary noises that are commonly encountered in naturalistic acoustic environments\textsuperscript{31,33,34}.

In summary, our findings provide insights on the dynamic and adaptive properties of speech processing in the human auditory cortex that enables a listener to suppress the deleterious effects of environmental noise and focus on the foreground sound, therefore making speech a reliable and robust means of communication.
7 NapLib: Neural acoustic processing library

Portions excerpted from:

7.1 Toolbox Description

In recent years, many non-invasive studies utilizing EEG and MEG have focused on understanding speech processing for applications to various hearing and language disorders. Additionally, invasive recordings can be used to study auditory neuroscience, as well as performing brain mapping, an essential clinical procedure for epilepsy patients who go under surgical resection of seizure foci. Several approaches for brain mapping have already been developed, including electrical cortical stimulation (ECS), corticocortical evoked potentials (CCEP), and mapping based on high gamma activity.

While these methods are effective for their intended applications, they do not provide information about tuning properties of electrodes or characterize the neural encoding of speech. Additionally, ECS can induce seizures in subjects. Traditionally, quantification of the tuning properties of auditory brain regions is performed by calculating spectrotemporal receptive fields (STRFs), which are linear maps between stimulus and response that quantify a neuron’s or neural population’s ideal stimulus. However, STRFs suffer from several drawbacks. First, STRFs assume a linear relationship between stimulus and response, an assumption which has been proven false, particularly in higher-level processes. Additionally, STRFs are dependent on the particular algorithm chosen for regularization (e.g., norm and sparsity constraints), which can limit their interpretability. Finally, solving the linear regression typically requires the computation of the inverse of large matrices, making them computationally intractable for real-time systems.

In this chapter, we present the Neural Acoustic Processing Library (NAPLib), a library for studying brain regions involved in speech processing. Recent studies have shown the encoding of acoustic-phonetic features in speech cortices; since each phonemic category has unique spectrotemporal properties, studying the responsiveness of neural activity to these categories
informs us about spectro-temporal properties of responsive regions. These methods do not make linear model assumptions, and they are computationally efficient meaning that they can be implemented in real-time. We include both real-time and offline processing tools, and we demonstrate the use of this toolbox in both noninvasive and invasive neural recordings.

NAPLib is comprised of two main libraries for real-time and offline processing. The offline toolbox is developed in both MATLAB and Python, and provides functionality for source selectivity analysis, quantification of response delay, and analysis of phoneme similarity patterns in neural and acoustic space. The real time toolbox, developed in Simulink, provides quantification of electrode responses to speech and shows the selectivity of sources to segmental units, such as phonemes. Additionally, we provide a small, open-source corpus of American English.

**Speech stimuli:**

NAPLib quantifies spatial and temporal properties of neural responses to phoneme categories as subjects listen to continuous speech. In order to implement this technique, the continuous speech signal must be temporally aligned with the corresponding phoneme sequence. With the library, we provide a small, open-source corpus of American English with forced alignments generated using the Penn Phonetics Lab Forced Aligner. We provide 25 minutes of speech, consisting of 148 utterances, 8450 phonemes and two speakers (a male and a female). NAPLib is generalizable to any phonetically aligned corpus (e.g. TIMIT). Additionally, there are many open-source toolkits that can be used to generate forced alignments for existing and custom corpora.

**Offline processing:**

The offline toolbox is developed in both MATLAB and in Python. It contains three modules: data preprocessing, noise reduction and artifact rejection, and phoneme analysis.
**Preprocessing:**

The preprocessing module aligns the phoneme labels (or other segmental unit) of the stimulus with the neural recording, allows the user to choose a scalp map (EEG) or electrode locations (ECoG) for visualization purposes, and performs filtering. For EEG, we provide zero-lag, FIR bandpass filter with cut-off frequencies of 2 and 15 Hz. For ECoG, we provide a filter bank to extract high gamma activity (envelop of 70 to 150 Hz), high gamma is correlated with underlying neuronal spiking activity and encodes phonetic feature information. In offline processing, filters are non-causal and zero-phase.

The library consists of different noise reduction techniques including common average referencing, principal component analysis decomposition, and trial rejection based on visual inspection and setting a threshold.

**Phoneme analysis:**

After preprocessing, denoising, and artifact rejection, the data can be fed into the phoneme analysis pipeline. Phoneme analysis can be used to perform brain mapping for speech selective regions, finding response delay and phonetic selectivity of electrodes, and quantifying the degree to which acoustic variability is reflected in neural data.

In addition to individual phonemes, we include functionality for grouping phones based on phonetic features (manner of articulation, place of articulation), phone length, and speakers. This allows for the study of acoustic, phonetic, and speaker features. In addition, this unit selection is easily generalizable and users can generate their own method for creating segmental units (e.g., syllables). When performing phonetic analyses, we recommend using individual phonemes for ECoG and EEG group analysis, while clustering labels into manner of articulation for single subject EEG due to noise concerns.
Average electrode response to phonemes:

The average response elicited by each phoneme is an important tool for visualizing the feature selectivity of an electrode. The average response $\overline{R}_{e,k}$ of electrode $e$ for phoneme $k$ occurring at time points of $T_1, T_2, ..., T_{N_k}$ in the stimulus is given by:

$$\overline{R}_{e,k}(t) = \frac{1}{N_k} \sum_{n_k=1}^{N_k} r(T_{n_k} + t, e),$$

where $r(t_{n_k}, e)$ is the neural response at time of phone onset and $t$ defines the temporal window over which the average is computed.

Response delay:

The latency between speech stimulus $S$ and neural response $R$ varies based on the brain region from which it was recorded. We quantify the distinction between phonemes at each time point using the F-statistic (between-group variability divided by within-group variability) and define response delay as the time point that shows maximum distinction between categories. Consider $R_{e,k,n}(t)$ as the response of electrode $e$ to the $n^{th}$ instance of phoneme category $k$, where $t$ denotes the sample time after the onset of phoneme. The response time is given by:

$$T_e = \arg\max_t \left( \frac{\sum_k N_k (\overline{R}_{e,k}(t) - \overline{R}_e(t))^2/(K - 1)}{\sum_{n_k} (R_{e,k,n}(t) - \overline{R}_{e,k}(t))^2/(N - K)} \right),$$

where $K$ denotes the number of phoneme categories, $N$ denotes the total number of phones in the corpus (all categories), $R_{e,k}(t)$ is defined in (1), and $\overline{R}_e(t)$ is the global mean of responses regardless of phoneme category. Categories are by default individual phonemes, but this can be generalized to any specified segmental unit.

Phoneme selectivity:
In order to characterize the selectivity of the neural response to phonemic categories at individual electrodes, we calculate the phoneme selectivity index (PSI) vector. Each electrode is characterized by a [Kx1] vector, with each element corresponding to the PSI of one phoneme; each PSI has a value ranging from 0 to K that quantifies the number of phonemes that elicit a statistically distinguishable response from the target phoneme (Wilcoxon rank-sum test). Quantification of distinction between phonemes. We calculate the distance (default is Euclidean) between the responses to phonemes for each phoneme pair at every time lag, yielding a time-varying pairwise phoneme distance matrix.

This analysis focuses on the similarities and distinctions between categories rather than on individual items. We also provide functionality to visualize the distance matrices in two and three dimensions using multi-dimensional scaling (MDS) and t-SNE.

**Comparison of phoneme properties between stimuli and response:**
Speech is a continuous signal that changes over time; even within a single category, the acoustic properties change from the start to the end of the phone. In order to find how neural responses and acoustic properties of speech sounds are related through time, we define a neural acoustic covariance matrix. This is a two-dimensional matrix that demonstrates the similarity between patterns of phones in the acoustic space and the corresponding neural responses at each time point. The acoustic similarity matrix is calculated using the acoustic spectrogram of phones.

**Functional connectivity of electrodes:**
The functional connectivity of recording regions is quantified by finding the covariance between distinction patterns of different electrodes.

**Group analysis:**
We provide an option of group analysis specifically recommended for analysis of EEG data when one subject does not provide sufficient signal to noise ratio. In this case, the response to the same phone is averaged between different subjects, after which all of the other analyses can be utilized.

The pipeline of phoneme analysis to find the responsive electrodes can be implemented in real time as shown in Figure 7.1.

Figure 7.1 Pipeline of phoneme analysis can be implemented in real-time.
7.2 Experiments

To show the efficacy of our toolbox, we demonstrate analyses from neural recordings in both EEG and ECoG.

Neural Recordings:
We recorded neural activity from subjects as they listened to the provided NAPLib corpus. We recorded from 22 EEG participants with a 62-channel recording system. The three ECoG participants were undergoing neurological assessment for epilepsy surgery; one patient had a high-density microelectrode grid array over temporal lobe, while the other had stereo EEG electrodes implanted. All subjects provided written informed consent. The Institutional Review board (IRB) of Columbia University at Morningside Campus approved all procedures.

Average electrode response and phoneme selectivity:
In this section we demonstrate how NAPLib can be used to visualize and quantify phonetic selectivity in both invasive and non-invasive recordings. Results shown are using the offline module, but we would like to emphasize that these analyses can also be implemented using the real-time module.

Figure 7.2A shows the average response of an example electrode (FCz) in EEG generated by group analysis including 22 subjects. Due to poor spatial resolution, it is typical to find broad responses to many phonemic categories. Because EEG recordings are also noisy, we also recommend averaging responses over subjects using group analysis. ECoG recordings provide much higher spatial resolution, resulting in the average response and corresponding PSI vector shown in Figure 7.2B from a depth electrode in Heschl’s gyrus. This electrode responds to unvoiced sibilants and
affricates (PSI > 25), which all contain strong power in high frequency channels. This suggests that this electrode has broad tuning to high frequencies, and indeed, we can see that the STRF of this electrode closely matches the average spectrogram of these phonemes (Figure 7.2C).

Response delay:

Figure 7.3 shows the F-test value at each time point based on the onset of phonemes. Phonemes are categorized based on manners of articulation. Figure illustrated the time differences between acoustic phonemes, ECoG (an electrode in Heschl’s gyrus), and EEG (FCz, 22 subjects). The local maxima are denoted with asterisks.
Figure 7.3 Finding latency using Fratio maxima.

Mapping time:
In order to quantify the duration which is needed to find a significant phoneme response, we used the ANOVA F-test (Figure 7.4). In EEG, the reported duration is based on significant distinction between manners of articulation. In ECoG, the duration for both manner of articulation and individual phoneme categories is reported. The p-value is assessed by the F distribution with correction for multiple comparisons implemented via false discovery rate ($q < 0.01$). The calculated time duration does not include the natural silences of speech.
Figure 7.4 Duration of speech required for generating phoneme mapping.

Locating speech-responsive regions:

Determining the location of response is important for a variety of both clinical and research applications. Figure 7.5A shows the percentage of time that EEG electrode responses display a statistically significant response to speech over a 10 minute duration. Figure 7.5B shows the time needed to elicit a statistically significant response to speech a patient with an implanted ECoG microelectrode array.
In this chapter, the Neural Acoustic Processing Library (NAPLib) is described, a free and open source toolbox for studying the neural representation of speech. The toolbox quantifies temporal and spectral responsiveness of electrodes based on responses to segmental linguistic categories (phonemes). Using such an approach allows for fast, efficient computations that can be implemented in real-time. As a proof of concept, we demonstrate use of the toolbox using both invasive (ECoG) and non-invasive (EEG) recordings.
8 Conclusion

This dissertation focused on dynamic and adaptive encoding of acoustic features in the human brain. Two methods of EEG and ECoG neural recordings were used to study the representation of phonetic features, the joint encoding of speaker and phonemes, and the noise reduction process through adaptation in the human brain.

Chapter 3 showed the EEG responses to continuous speech reliably encoded phonemes and phonetic features. Chapter 4 showed speaker encoding at multiple time intervals relative to the onset of the phonemes. It was demonstrated that EEG responses are primarily organized by phonetic feature, while subtler speaker variations appear within manner of articulation groups. This demonstration is consistent with previous studies that illustrated the role of phonetic over speaker characteristics in shaping the acoustic properties of phones.

Chapter 5 examined the dynamic reduction in background noise using invasive electrophysiology combined with behavioral experiments. The auditory neural responses momentarily responded to noise features when a new background noise appeared in the acoustic scene during speech perception. The auditory neural responses rapidly adapted to suppress the neural encoding of noise, resulting in enhanced neural encoding and perception of phonetic features of speech. Chapter 6 explained the role of attention and brain regions in this adaptive process. Our results showed a diversity of adaptation patterns across electrodes and cortical areas, which was largely predictable from the response properties of electrodes. Moreover, adaptation was present even when participants’ attention was focused on a secondary visual task. Chapter 7 presented a suite of tools that can be used for EEG and ECoG recordings to characterize their responses to phonemes, phonetic features, and words.
Together, this dissertation showed new evidence for dynamic and adaptive processing of speech sounds in the auditory pathway and provided computational tools to study the dynamics of speech encoding in the human brain.
9 References


10. Näätänen, R. The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology* 38, 1–21 (2001).


29. Assmann, P. & Summerfield, Q. The perception of speech under adverse conditions. in *Speech processing in the auditory system* 231–308 (Springer, 2004).


67. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation,


141. Carandini, M., Heeger, D. J. & Senn, W. A synaptic explanation of suppression in visual cortex. *J.*