

8-2017

Investigating the Neural Basis of Audiovisual Speech Perception with Intracranial Recordings in Humans

Muge O. Sertel

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations



Part of the [Cognitive Neuroscience Commons](#), and the [Systems Neuroscience Commons](#)

Recommended Citation

Sertel, Muge O., "Investigating the Neural Basis of Audiovisual Speech Perception with Intracranial Recordings in Humans" (2017). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 797.
https://digitalcommons.library.tmc.edu/utgsbs_dissertations/797

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

**INVESTIGATING THE NEURAL BASIS OF AUDIOVISUAL SPEECH
PERCEPTION WITH INTRACRANIAL RECORDINGS IN HUMANS**

by

Müge Özker Sertel, M.Sc.

APPROVED:

Michael S. Beauchamp, Ph.D. Advisory Professor

William Mattox, Ph.D. Onsite Advisor

Michael Beierlein, Ph.D.

Raymond Cho, M.D., M.Sc.

David Ress, Ph.D.

APPROVED:

Dean, The University of Texas

MD Anderson Cancer Center UT Health Graduate School of Biomedical
Sciences at Houston

**INVESTIGATING THE NEURAL BASIS OF AUDIOVISUAL SPEECH
PERCEPTION WITH INTRACRANIAL RECORDINGS IN HUMANS**

A
DISSERTATION

Presented to the Faculty of
The University of Texas
MD Anderson Cancer Center UT Health
Graduate School of Biomedical Sciences

In Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

by
Müge Özker Sertel, M.Sc.
Houston, Texas
August, 2017

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor, Dr. Michael Beauchamp, for his continuous guidance and support. His immense knowledge in neuroscience, enthusiasm for research, caring, patience and generosity in providing all resources I have needed for my research has been invaluable for fostering my academic advancement and set an outstanding example as a role model for me.

I would like to express my gratitude to my advisory committee: Dr. William Mattox, Dr. Michael Beierlein, Dr. David Ress and Dr. Raymond Cho. Their insightful comments and suggestions have been very helpful in improving my research. I have always felt very motivated after our enriching discussions about science at our meetings.

I would like to give special thanks to my lab family, Debshila, Lin, John, Johannes and Inga for making the work environment feel like home. We have shared so much over the years. They have filled my days with joy and laughter. In the most stressful days of graduate school life, their support and empathy made me stronger. I feel so lucky that our paths intersected in the Beauchamp Lab and I have gained such great friends for life.

I would also like to thank our electrocorticography team at Baylor College of Medicine: Dr. Daniel Yoshor, Dr. Bill Bosking, Dr. Brett Foster and Ping Sun. We have worked for many long hours together doing experiments. They are all excellent scientists. I have enjoyed working with them and learned so much from them.

I am grateful to GSBS and the neuroscience program for being such a supportive community with all the students, faculty and staff.

I am extremely grateful to my family in Turkey who has worked so hard to give me all of the opportunities in life. They have always believed in me and encouraged me to follow my dreams even if it meant being apart and longing.

Finally, I can't thank enough to my wonderful husband Tefvik for the sacrifices he has made. He came to Houston with me leaving his family, friends and job behind so that I can pursue my PhD. I cannot even imagine coming this far without him by my side. His love gives me all the strength I need in this life.

ABSTRACT

INVESTIGATING THE NEURAL BASIS OF AUDIOVISUAL SPEECH PERCEPTION WITH INTRACRANIAL RECORDINGS IN HUMANS

Müge Özker Sertel, M.Sc.

Advisory Professor: Michael S. Beauchamp

Speech is inherently multisensory, containing auditory information from the voice and visual information from the mouth movements of the talker. Hearing the voice is usually sufficient to understand speech, however in noisy environments or when audition is impaired due to aging or disabilities, seeing mouth movements greatly improves speech perception. Although behavioral studies have well established this perceptual benefit, it is still not clear how the brain processes visual information from mouth movements to improve speech perception. To clarify this issue, I studied the neural activity recorded from the brain surfaces of human subjects using intracranial electrodes, a technique known as electrocorticography (ECoG). First, I studied responses to noisy speech in the auditory cortex, specifically in the superior temporal gyrus (STG). Previous studies identified the anterior parts of the STG as unisensory, responding only to auditory stimulus. On the other hand, posterior parts of the STG are known to be multisensory, responding to both auditory and visual stimuli, which makes it a key region for audiovisual speech perception. I

examined how these different parts of the STG respond to clear versus noisy speech. I found that noisy speech decreased the amplitude and increased the across-trial variability of the response in the anterior STG. However, possibly due to its multisensory composition, posterior STG was not as sensitive to auditory noise as the anterior STG and responded similarly to clear and noisy speech. I also found that these two response patterns in the STG were separated by a sharp boundary demarcated by the posterior-most portion of the Heschl's gyrus. Second, I studied responses to silent speech in the visual cortex. Previous studies demonstrated that visual cortex shows response enhancement when the auditory component of speech is noisy or absent, however it was not clear which regions of the visual cortex specifically show this response enhancement and whether this response enhancement is a result of top-down modulation from a higher region. To test this, I first mapped the receptive fields of different regions in the visual cortex and then measured their responses to visual (silent) and audiovisual speech stimuli. I found that visual regions that have central receptive fields show greater response enhancement to visual speech, possibly because these regions receive more visual information from mouth movements. I found similar response enhancement to visual speech in frontal cortex, specifically in the inferior frontal gyrus, premotor and dorsolateral prefrontal cortices, which have been implicated in speech reading in previous studies. I showed that these frontal regions display strong functional connectivity with visual regions that have central receptive fields during speech perception.

TABLE OF CONTENTS

APPROVAL PAGE	i
TITLE PAGE	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
TABLE OF CONTENTS	vi
LIST OF FIGURES	vii
LIST OF TABLES	viii
CHAPTER 1: INTRODUCTION	1
The Impact of Visual Speech on Speech Perception.....	2
Brain Regions Involved in Speech Perception.....	4
Electrocorticography	9
Goals.....	11
CHAPTER 2: PROCESSING OF NOISY SPEECH IN AUDITORY CORTEX....	14
Introduction.....	15
Methods.....	17
Results.....	26
Discussion	48
CHAPTER 3: PROCESSING OF SILENT SPEECH IN VISUAL CORTEX	59
Introduction.....	60

Methods.....	61
Results.....	67
Discussion.....	75
CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS	81
REFERENCES	89
VITA.....	110

LIST OF FIGURES

Figure 2.1	Audiovisual speech stimuli with clear and noisy components	18
Figure 2.2	Location of electrodes on the superior temporal gyrus	27
Figure 2.3	Response amplitudes in all STG electrodes	30
Figure 2.4	Response amplitudes in single STG electrodes.....	31
Figure 2.5	Response amplitudes with respect to location on the STG.....	33
Figure 2.6	Response variability in all STG electrodes	37
Figure 2.7	Response variability in single STG electrodes	38
Figure 2.8	Response variability with respect to location on the STG	39
Figure 2.9	Response timing in STG electrodes.....	40
Figure 2.10	Response duration, response onset and time-to-peak in STG electrodes	42
Figure 2.11	Response amplitudes in STG with varying auditory noise levels ..	46
Figure 2.12	Bayesian model of audiovisual speech with auditory noise	50
Figure 3.1	Retinotopic organization of speech responses in visual cortex.....	68
Figure 3.2	Speech responses in frontal cortex	71
Figure 3.3	Trial-by-trial correlation for a single frontal-visual electrode pair ...	72
Figure 3.4	Functional connectivity between frontal and visual cortices.....	73
Figure 3.5	Latency of the response enhancement	74

LIST OF TABLES

Table 2.1	Linear mixed-effects model of the response amplitude	28
Table 2.2	Linear mixed-effects model of the response variability	35
Table 2.3	Linear mixed-effects model of the response duration	41
Table 2.4	Linear mixed-effects model of the response onset.....	43
Table 2.5	Linear mixed-effects model of the response peak time	43
Table 2.6	Linear mixed-effects model of the effect of accuracy on response amplitude	44
Table 2.7	Linear mixed-effects model of the effect of accuracy on response variability.....	45
Table 2.8	Linear model of the effect of varying auditory noise levels on response amplitude	47

CHAPTER 1: INTRODUCTION

The Impact of Visual Speech on Speech Perception

Speech perception is a multisensory process that involves audition and vision. When we converse with someone, we receive both auditory information from the talker's voice, and visual information from the talker's facial movements and combining the two helps us better understand what is being said. Previously speech perception was considered only an auditory function that results from the processing of speech sounds in the auditory cortex. Although hearing the voice can be sufficient to understand speech, seeing the mouth movements of the talker greatly improves speech perception under adverse listening conditions. For example, we can understand speech when we hear someone on the telephone or when we listen to the radio. However when we are in a noisy environment, when multiple people talk at the same time or when we hear someone with a strong accent, seeing mouth movements of the talker can be very helpful since we can use the shape of the mouth to identify the spoken words (1-3). In the presence of background noise, seeing mouth movements was shown to improve speech intelligibility equivalent to a signal-to-noise ratio increase of 15dB (4).

Visual speech information is especially beneficial for people with hearing impairments. Studies showed that viewing facial movements greatly helps individuals with partial hearing loss to recognize what they hear. Also, their lip-reading abilities are enhanced as they increasingly rely on visual speech information. For example older adults with hearing impairments were shown to be better at visual identification of words compared with older adults with normal hearing (5). In the case of profound deafness, lip reading by itself can be

sufficient for speech perception (6). When individuals with profound deafness are implanted with cochlear implants, they continue to benefit from lip-reading to aid their partially recovered hearing (7).

Visual speech does not only influence what we hear under adverse listening conditions. For example, the McGurk effect is a well-known perceptual phenomenon that demonstrates how visual speech can alter what is being heard even under normal listening conditions. McGurk and colleagues showed that when an auditory syllable 'ba' is paired with an incongruent visual syllable 'ga', neither the auditory nor the visual syllable is perceived but instead the two syllables fuse into a completely different syllable and perceived as 'da' (8). Subsequent psychophysical studies investigated the temporal constraints on the McGurk effect and showed that the illusion only occurs if auditory and visual syllables are presented synchronously or within a tolerable range of temporal delay. For example, if the auditory syllable was presented with more than ~300 ms delay, subjects would no longer get the fused percept but rather report perceiving the auditory syllable (9). This suggests that visual speech can alter auditory speech perception as long as the two appear to originate from the same source.

Another perceptual phenomenon that illustrates how strongly visual speech can effect heard speech is the ventriloquist effect. Ventriloquism is an old stage act, in which the performer talks without moving his/her lips while simultaneously moving a puppet's mouth, thus making his/her voice appear to be coming from the puppet. Ventriloquist effect demonstrates that visual speech can

dominate auditory speech so that the perceived location of auditory speech is shifted towards the direction of visual speech (10).

All these examples provide evidence for the perceptual interaction of auditory and visual speech. They show that visual speech can improve, alter or dominate auditory speech. Although the perceptual effects of visual speech have been widely studied and understood on a behavioral level, there are still many unknowns about how and where in the brain these multisensory interactions occur.

Brain Regions Involved in Speech Perception

Multisensory nature of speech perception involves a distributed network of brain regions: 1) Auditory cortex including Heschl's gyrus, planum temporale and surrounding auditory association areas on superior temporal gyrus (STG) and superior temporal sulcus (STS), 2) Primary visual cortex and lateral extrastriate areas including the motion sensitive middle temporal area hMT/V5 and posterior superior temporal sulcus, 3) Frontal cortex including inferior frontal gyrus and premotor cortex 4) Parietal cortex including temporoparietal junction and inferior parietal lobule.

Auditory cortex consists of functionally specialized subregions that are hierarchically structured. As one goes up in the hierarchy, specifically from the auditory core to belt and parabelt areas, response preference changes from simple stimuli such as tones to more complex stimuli like band-pass noise bursts and speech (11). The posterior third of the superior temporal gyrus (STG), including Brodmann areas 41, 42 and 22, is defined as the lateral parabelt

auditory cortex in human (12). This region contains the Wernicke's area, which has been strongly associated with speech comprehension since early clinical neurology findings (13-16).

Posterior superior temporal cortex, including both the gyrus and the sulcus (STG/S), is a key brain region for audiovisual interactions during speech perception. It is strategically located between the auditory and visual cortices and receives inputs from both auditory and extrastriate visual cortices as shown in anatomical studies in monkeys (17, 18). Both electrophysiological studies in monkeys and neuroimaging studies in humans demonstrated that posterior STG/S responds strongly to both auditory and visual stimuli, however responds the most when the two stimuli are presented together (19-22).

Posterior STG/S is particularly sensitive to vocal sounds and facial movements. A PET study with monkeys showed that species-specific calls such as coos and screams elicit larger responses in the posterior STG than non-biological sounds such as musical instruments or environmental sounds (23). Similarly, a human fMRI study demonstrated that posterior STS shows greater neural activity for vocal sounds compared with non-vocal environmental sounds (24). Another fMRI study showed that this region also responds when subjects view eye and mouth movements, but does not respond when they view checkerboard patterns (25).

Numerous studies implicated posterior STG/S as a critical region for audiovisual interactions during speech perception. Neural activity in this region was shown to be greater when auditory syllables or words were presented with

corresponding articulatory gestures compared with when they were presented alone (26, 27). In the STS, multisensory responses were shown to exhibit sub-additive properties with responses to multisensory stimulus being larger than the maximum or the mean of the responses to unisensory stimuli (28). This multisensory gain obtained by presenting the two sensory stimuli together is considered a signature of multisensory integration.

Another characteristic of multisensory integration is called the principle of inverse effectiveness, which predicts that the multisensory gain should be greater when one of the sensory modalities is degraded with noise. In line with this principle, fMRI studies showed that the response enhancement for audiovisual speech in posterior STG/S is larger when the auditory or the visual component is degraded by noise (29, 30).

A famous computational model, the Bayesian model of multisensory integration, attempts to explain the neural mechanism that underlies multisensory integration. According to the Bayesian theory, representation of the sensory world around us is not deterministic. There is internal noise in our sensory system due to factors such as probabilistic neurotransmitter release or density of receptors in the retina. For this reason when a stimulus is presented multiple times, the neural response to the exact same physical stimulus will be different at each time. Bayesian theory suggests that the brain represents sensory information probabilistically and the uncertainty is manifest in the variability of the probability distribution function. Integrating different sensory information about the same stimulus should reduce the variability of the neural response, which

would in turn increase the accuracy of sensory perception (31-33). In the case of speech perception, integrating auditory and visual information about the same speech stimulus is expected to reduce the variability of the neural response and thus improve speech perception.

A line of studies showed that posterior STG/S responded the most when auditory and visual speech components are fused together and perceived as originating from the same source. In an fMRI study, subjects were presented with audiovisual syllables in which the auditory and visual components were temporally offset. It was shown that responses in the posterior STS are greater when the presented syllables are perceived as synchronous than when they are perceived asynchronous (34). Another study demonstrated that the left posterior STS is a locus for the McGurk effect, such that disrupting the neural activity in this region by transcranial magnetic stimulation results in reduced McGurk percepts (35), thus providing direct evidence on the role of this region in the audiovisual integration of speech.

Early and late visual cortical areas are also involved in speech processing. Previous studies have reported responses to visual and audiovisual speech in the banks of the calcarine sulcus, cuneus, lingual gyrus, occipital pole and lateral occipital regions (36-39). An MEG study showed that, when subjects view verbal lip forms, responses start in the occipital cortex and progress to the superior temporal gyrus (40). Another study used fMRI to measure functional connectivity between extrastriate visual cortex and posterior STS during speech perception. They found that functional connectivity between the two regions increases when

auditory component of speech is noisy, suggesting that visual cortex supplies visual information during speech perception (41).

A widely accepted auditory processing model suggests that auditory processing is organized in a dual pathway analogous to visual processing. According to this model, auditory pathways emanate from the auditory belt area and run in two different directions. The first is the 'what' pathway, which runs anteriorly along the superior temporal gyrus and terminates at the ventrolateral prefrontal cortex, carries out sound object identification including identification of speech sounds. The second is the 'where' pathway, which runs posteriorly through inferior parietal cortex and terminates at the dorsolateral prefrontal cortex, carries out sound localization (42, 43).

According to the dual pathway model, the two auditory processing streams converge in the prefrontal cortex. Ventrolateral prefrontal cortex on the inferior frontal gyrus, which contains the Broca's area, has long known to be a motor region critical for speech production since early lesion studies (44, 45). Later neuroimaging studies showed that Broca's area is not only activated during speech production but also during speech perception (46). Moreover, it also displayed multisensory characteristics by responding to both auditory and visual speech and showing response enhancement to audiovisual speech with auditory noise (47). These findings supported the idea that Broca's area is a region with mirror system properties, where auditory speech information is matched with articulatory speech gestures during speech processing (48).

Speech responses have also been reported in other frontal regions, such as dorsal regions of the premotor cortex, however these regions did not exhibit response enhancement to audiovisual speech but rather showed greater responses to visual speech compared with audiovisual speech suggesting a role in visual-articulatory processing of speech (39, 47, 49).

Electrocorticography

The most popular technique for examining human brain function is blood oxygen level dependent functional magnetic resonance imaging (BOLD fMRI) (50), which has been widely used to study speech perception. While fMRI provides comprehensive information about the spatial details of the brain networks involved in speech processing, it is limited in terms of temporal information it can provide. BOLD signal is an indirect measure of neural activity that reflects slow blood flow and oxygen metabolism changes in the cerebral vasculature. It results in a temporal resolution of approximately 2 seconds, which is too slow to observe the neural responses to rapidly changing speech stimulus, considering that the speed of spontaneous speech can exceed 200 words per minute (51).

To measure speech responses in the human brain, we used an electrophysiological technique called electrocorticography (ECoG). ECoG is an invasive technique, in which intracranial electrodes are implanted on the cortical surface to record neural activity. It is essentially used to monitor seizure activity in epilepsy patients for clinical purposes, but at the same time it provides a unique research opportunity to obtain extremely detailed and precise information

from the human brain. It allows the measurement of neural activity at a very high temporal (~ 1 ms) and spatial (~ 10 mm) precision, a level of detail that has previously been restricted to invasive procedures on nonhuman primates.

ECoG is also known as intracranial EEG. Similar to EEG, it measures the local field potentials that reflect the summed postsynaptic electrical signals generated by the neural population underneath the electrode (can record the activity of $\sim 10^5$ neurons with a typical electrode of ~ 2 mm diameter) (52).

However ECoG has multiple advantages over EEG. First, ECoG has superior spatial resolution because the electrodes are placed on the cortical surface and they record electrical signals generated by neurons that are in close proximity. On the other hand, in EEG recordings, electrodes are placed on the scalp and they record electrical signals that are blended and distributed on the scalp surface due to the volume conduction effects (53). The approximate origin of the recorded signals can only be estimated by performing source localization analyses, limiting the spatial resolution of EEG.

Second, ECoG recordings has much higher signal-to-noise (SNR) ratio compared to EEG. In EEG recordings, signals get filtered through cerebrospinal fluid, skull and scalp. In EEG studies, responses to many trials of the same experimental condition are averaged to increase the SNR. However, ECoG can robustly measure the response for even a single trial, allowing for analyses that measure variability across trials, trial-by-trial connectivity between regions, or correlation of neural activity with behavioral responses.

Neural activity measured as local field potentials is oscillatory, meaning that it is composed of signals that oscillate at different frequencies. ECoG can reliably record high-frequency signals components (> 40 Hz), which are considered to be a good measure of population-level neural activity. It was shown that high-frequency activity is strongly correlated with both spiking rate of single-neurons and BOLD signal in human sensory cortex (54, 55). Numerous studies have used ECoG to demonstrate high-frequency responses for sensory and cognitive tasks and speech perception studies are no exception (56-59).

Goals

Visual information conveyed by mouth movements is an important component of speech perception. Behavioral studies have established that seeing mouth movements is especially beneficial when auditory speech is noisy or inaudible. However we still remain largely ignorant about the neural processes that underlie the processing of visual speech. The goal of my dissertation is to use ECoG recordings to shed light on how visual speech information modulates neural activity in the brain to improve speech perception under adverse listening conditions.

In the *first* part of my thesis (Chapter 2), I focus on speech processing in the auditory cortex and investigate how neural activity in the superior temporal gyrus (STG) is modulated when auditory speech is noisy. I hypothesized that neural activity in posterior STG will be less affected by auditory noise compared with anterior STG, because posterior STG is a multisensory region and can utilize visual speech information to compensate for auditory noise.

In the *second* part of my thesis (Chapter 3), I focus on speech processing in the visual cortex and investigate how neural activity in the visual cortex is modulated when auditory speech is inaudible. Various studies implicated frontal cortex in the modulation of visual responses during visual attention tasks (60, 61). When visual speech is the only source of information, I hypothesized that there will be a top-down influence from frontal regions on visual cortex to amplify responses, especially in visual regions with receptive field locations that correspond to the mouth of the talker.

I pursued the following specific aims to test my hypotheses:

Specific Aim 1: To determine how noisy speech modulates neural activity in the multisensory posterior STG: When auditory modality is noisy, seeing mouth movements improves speech intelligibility. I predicted that a key neural mechanism for this perceptual improvement is reduced variability: visual speech reduces the uncertainty caused by noisy auditory speech and this reduction in uncertainty is manifest in the neural response as reduced variability. Because posterior STG is multisensory and can process visual speech information to counteract the effects of noisy auditory speech, auditory noise should not affect the neural response in the posterior STG. To test this prediction, subjects were presented with repeated trials of clear and noisy speech stimuli and high-frequency broadband responses in the STG were measured. I found that noisy speech decreased the amplitude and increased the trial-to-trial variability of neural responses in anterior but not in posterior STG.

Specific Aim 2: To determine if frontal cortex modulates visual cortex in a retinotopically specific manner to enhance responses to visual speech:

Responses in the visual cortex are enhanced when speech contains a noisy or entirely absent auditory component (36). I hypothesized that this response enhancement is a result of top-down modulation from frontal cortex. Because mouth is the most important facial region for transmitting visual speech information, people naturally fixate on the mouth of the talker to perceive speech under adverse listening conditions (62). I therefore predicted that the top-down modulation would be more pronounced for visual regions with central receptive fields. To test these hypotheses, subjects were presented with repeated trials of audiovisual and visual speech stimuli and high-frequency broadband responses were measured simultaneously in the visual and frontal cortices. Functional connectivity between all frontal-visual electrode pairs was examined using trial-by-trial power correlation analysis. In a separate receptive field mapping experiment, receptive field locations of visual electrodes were determined (63). I found that response enhancement and functional connectivity with frontal electrodes was greater for visual electrodes with central receptive field locations compared to visual electrodes with peripheral receptive field locations.

CHAPTER 2: PROCESSING OF NOISY SPEECH IN AUDITORY CORTEX

Introduction

Note: This chapter is based upon: Ozker, M., I. M. Schepers, J. F. Magnotti, D. Yoshor, and M. S. Beauchamp. 2017. A Double Dissociation between Anterior and Posterior Superior Temporal Gyrus for Processing Audiovisual Speech Demonstrated by Electrocorticography. *Journal of Cognitive Neuroscience* 29: 1044-1060. Reprinted with the permission from MIT.

Human speech perception is multisensory, combining auditory information from the talker's voice with visual information from the talker's face. Visual speech information is particularly important in noisy environments in which the auditory speech is difficult to comprehend (1-3). While visual speech can substantially improve the perception of noisy auditory speech, little is known about the neural mechanisms underlying this perceptual benefit.

Speech varies on a timescale of milliseconds, requiring the brain to accurately integrate auditory and visual speech with high temporal fidelity. However, the most popular technique for measuring human brain activity, blood oxygen level dependent functional magnetic resonance imaging (BOLD fMRI) is an indirect measure of neural activity with a temporal resolution on the order of seconds, making it difficult to accurately measure the rapidly changing neural responses to speech with BOLD fMRI. In order to overcome this limitation, we recorded from the brains of subjects implanted with electrodes for the treatment of epilepsy. This technique, known as electrocorticography (ECoG), allows for the direct measurement of activity in small populations of neurons with millisecond precision. We measured activity in electrodes implanted over the superior temporal gyrus (STG), a key brain area for speech perception (59, 64),

as subjects were presented with audiovisual speech with either clear or noisy auditory or visual components.

The STG is functionally heterogeneous. Regions of anterior STG lateral to Heschl's gyrus are traditionally classified as unisensory auditory association cortex (65). In contrast, regions of posterior superior temporal gyrus and superior temporal sulcus are known to be multisensory, responding to both auditory and visual stimuli including faces and voices, letters and voices, and recordings and videos of objects (19, 26, 34, 66-68).

Based on this distinction, we hypothesized that anterior and posterior regions of STG should differ in their electrocorticographic response to clear and noisy audiovisual speech. We expected that auditory association areas in anterior STG should respond strongly to speech with clear auditory component but show a reduced response to the reduced information available in speech with noisy auditory component. Multisensory areas in posterior STG should be able to use the clear visual speech information to compensate for the noisy auditory speech, resulting in similar responses to speech with speech with clear and noisy auditory component.

A related set of predictions comes from theoretical models of Bayesian integration. In these models, sensory noise and the resulting neural variability is independent in each modality. Combining the modalities through multisensory integration results in a decreased neural variability (and improved perceptual accuracy) relative to unisensory stimulation (31, 33). Bayesian models predict that unisensory areas, such as those in anterior STG, should have greatly

increased variability as the sensory noise in their preferred modality increases. Multisensory areas, like those in posterior STG, should be less influenced by the addition of auditory noise, resulting in similar variability for speech with clear and noisy auditory component.

Methods

Subject Information

All experimental procedures were approved by the Institutional Review Board of Baylor College of Medicine. Five human subjects with refractory epilepsy (3F, mean age 31) were implanted with subdural electrodes guided by clinical requirements. Following surgery, subjects were tested while resting comfortably in their hospital bed in the epilepsy monitoring unit.

Stimuli, Experimental Design and Task

Visual stimuli were presented on an LCD monitor positioned at 57 cm distance from the subject and auditory stimuli were played through loudspeakers positioned next to the subject's bed. Two video clips of a female talker pronouncing the single syllable words "rain" and "rock" with clear auditory and visual components (AV) were selected from the Hoosier Audiovisual Multi-Talker Database (69). The duration of each video clip was 1.4 seconds and the duration of the auditory stimulus was 520 milliseconds for "rain" and 580 milliseconds for "rock". The auditory word onset was 410 milliseconds for "rain" and 450 milliseconds for "rock" after the video onset. The face of the talker subtended approximately 15 degrees horizontally and 15 degrees vertically. Speech stimuli were consisted of four conditions: Speech with clear auditory and visual

components (AV), clear visual but noisy auditory components (AnV), clear auditory but noisy visual components and finally noisy auditory and noisy visual components (AnVn) (Figure 2.1).

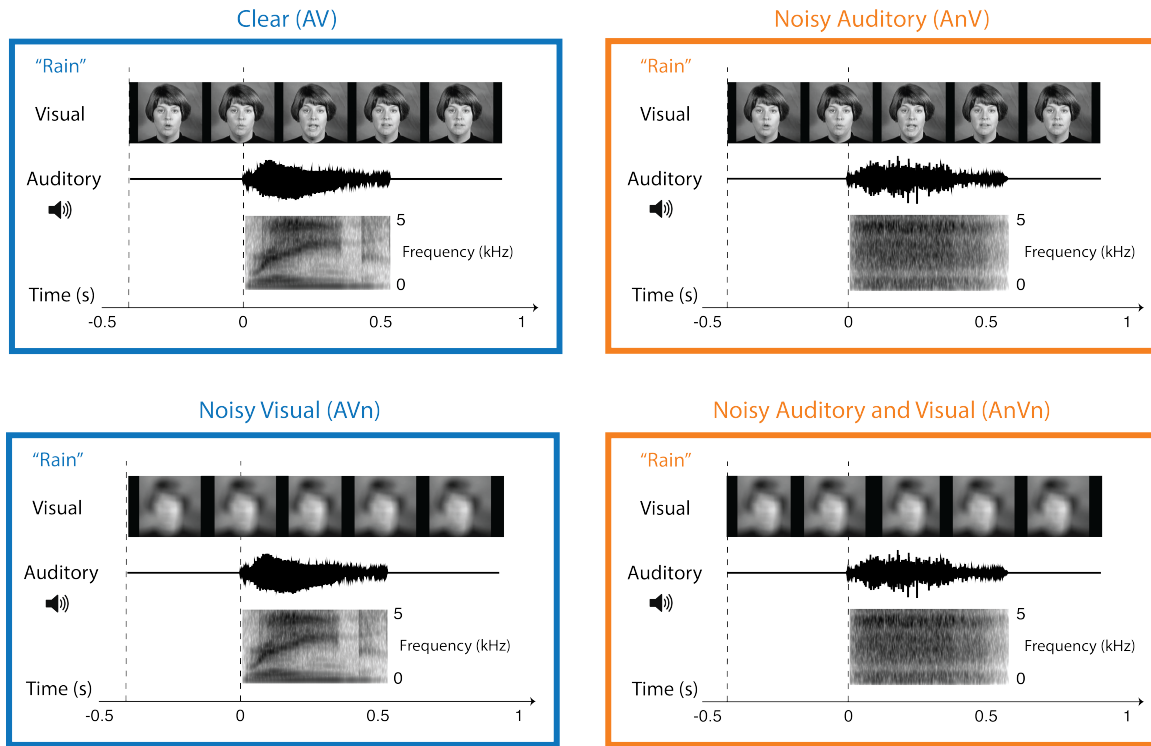


Figure 2.1 Audiovisual speech stimuli with clear and noisy components

Clear audiovisual speech (AV) consisted of a movie of a talker pronouncing the word “rain” or “rock”. Visual stimulus (top row) shows sample frames from the video. Auditory stimulus is shown as sound pressure level (middle row) and spectrogram (bottom row). Black vertical dashed lines indicate visual and auditory stimulus onsets. For noisy auditory speech (AnV), the auditory component was replaced with speech-specific noise of equal power to the original auditory speech. For noisy visual speech (AVn), the visual component was blurred using a low-pass Gaussian filter. For noisy auditory, noisy visual speech (AnVn), the auditory component was replaced with speech-specific noise and the visual component was blurred.

To create speech stimuli with a noisy auditory component, the auditory component of the speech stimulus was replaced with noise that matched the spectrotemporal power distribution of the original auditory speech. The total

power of this speech-specific noise was equated to the total power of the original auditory speech (70). This process generated speech-like noise that is difficult to recognize.

To create speech stimuli with a noisy visual component, the visual component of the speech stimulus was blurred using a 2-D Gaussian low-pass filter (Matlab function *fspecial* was used to create the filter, filter size = 30 pixels in each direction). Each video frame (image size = 200 x 200 pixels) was filtered separately using two-dimensional correlation (Matlab function *imfilter* was used to filter the images). Values outside the bounds of the images were assumed to equal the nearest image border. After filtering, images were combined back into a video. These filter settings resulted in highly blurred videos, where only the contours of the face and head and some rudimentary mouth movements were visible.

Thirty-two to fifty-six repetitions of each condition were presented in random sequence. Each 5.4 second trial consisted of a single 1.4 second video clip followed by an interstimulus interval of 4 seconds during which a fixation cross on a gray screen was presented. Subjects pressed a mouse button to report which word was presented.

Electrode Localization and Recording

Before surgery, T1-weighted structural magnetic resonance imaging scans were used to create cortical surface models with FreeSurfer (71, 72) and visualized using SUMA (73). Subjects underwent a whole-head CT after the electrode implantation surgery. The post-surgical CT scan and pre-surgical MR

scan were aligned using AFNI (74) and all electrode positions were marked manually on the structural MR images. Electrode positions were then projected to the nearest node on the cortical surface model using the AFNI program *SurfaceMetrics*. Resulting electrode positions on the cortical surface model were confirmed by comparing them with the photographs taken during the implantation surgery.

A 128-channel Cerebus amplifier (Blackrock Microsystems, Salt Lake City, UT) was used to record from subdural electrodes (Ad-Tech Corporation, Racine, WI) that consisted of platinum alloy discs embedded in a flexible silicon sheet. Electrodes had an exposed surface diameter of 2.3 mm and were located on strips or grids with inter-electrode distances of 10 mm. An inactive intracranial electrode implanted facing the skull was used as a reference for recording. Signals were amplified, filtered (low-pass: 500 Hz, Butterworth filter with order 4; high-pass: 0.3 Hz, Butterworth filter with order 1) and digitized at 2 kHz.

Electrophysiological Data Analysis

Data were analyzed in MATLAB 8.5.0 (MathWorks Inc. Natick, MA) using the FieldTrip toolbox (75). To remove common artifacts, the average signal across all electrodes was subtracted from each individual electrode's signal (common average referencing). The continuous data stream was divided into trials. Line noise at 60, 120, 180 Hz was removed and the data was transformed to time–frequency space using the multitaper method (3 Slepian tapers; frequency window from 10 to 200 Hz; frequency steps of 2 Hz; time steps of 10 ms; temporal smoothing of 200 ms; frequency smoothing of ± 10 Hz).

Our primary measure of neural activity was the broadband response in the high-gamma frequency band, ranging from 70 to 110 Hz. This frequency range is thought to reflect the frequency of action potentials in nearby neurons (54, 55, 76, 77). For each trial, the high-gamma response was measured in a window from 0 to 500 ms following auditory stimulus onset (reflecting the ~500 ms duration of the auditory stimulus) and converted to percent signal change measure by comparing the high-gamma response to a within-trial baseline window encompassing -500 to -100 ms before auditory stimulus onset. For instance, a 100% signal change on one trial would mean the power in the high-gamma band doubled from the pre-stimulus to the post-stimulus interval. For each electrode, the mean percent signal change in the high-gamma band across all trials of a given condition was calculated ($\mu\mu$).

Our second analysis focused on neural variability across repeated presentations of identical stimuli. One obvious measure of variability is variance (defined as the square of the standard deviation across all observations). However, the variance of neural responses is known to increase with increasing response amplitude (32, 78), and our initial analysis demonstrated differences in response amplitude between speech with clear and noisy auditory components (Table 1). To search for variability differences without the confound of these amplitude differences, we used a different measure of variability known as the coefficient of variation (CV) which normalizes across amplitude differences by dividing the standard deviation of the response across trials by the mean response amplitude ($CV = \sigma/\mu$) (79, 80). The CV assumes that variance covaries

linearly with amplitude. We tested this assumption by calculating the Pearson correlation between the mean and variance of the high-gamma response across all anterior and posterior STG electrodes and found it to be reasonable for the four different stimulus conditions (AV: $r = 0.96$, $p = 10^{-16}$; AnV: $r = 0.86$, $p = 10^{-8}$; AVn: $r = 0.97$, $p = 10^{-16}$; AnVn: $r = 0.91$, $p = 10^{-11}$). Although CV has the advantage of accounting for the known correlation between amplitude and variance, it has the disadvantage that it becomes undefined as response amplitude approaches zero. For this reason, response amplitudes of less than 15% were excluded from the CV analysis, affecting 3/16 anterior electrodes in Figure 2.6 and 8/216 condition-electrode pairs in Table 2.2 and Table 2.7.

Anatomical Classification and Electrode Selection

The superior temporal gyrus (STG) was segmented on each subject's cortical surface model. The posterior margin of the most medial portion of the transverse temporal gyrus of Heschl was used as a landmark to separate the STG into anterior and posterior portions. All of the STG anterior to this point (extending to the temporal pole) was classified as anterior STG. All of the STG posterior to this point was classified as posterior STG.

The cortical surface atlases supplied with FreeSurfer were used to automate ROI creation. The entire segmented STG was obtained from the Destrieux atlas (right hemisphere STG atlas value = 152, left hemisphere = 78) (81) and the anterior and posterior boundaries of the posterior STG were obtained from the Desikan-Killiany atlas (RH = 44, LH = 79) (82).

A total of 527 intracranial electrodes were recorded from. Of these, 55 were located on the STG. These were examined for stimulus-related activity, defined as significant high-gamma responses to audiovisual speech compared with pre-stimulus baseline ($p < 10^{-3}$, equivalent to ~40% increase in stimulus power from baseline). A total of 27 electrodes met both anatomical and functional criteria and were selected for further analysis. To simplify future meta-analyses and statistical comparisons between experiments, we do not report p-values as inequalities but instead report actual values (rounded to the nearest order of magnitude for p-values less than 0.001).

Response Timing Measurements

For each electrode, we calculated the response onset, time to peak and duration of the high gamma signal. To calculate the response onset, we found the first time point after the auditory speech onset at which the high-gamma signal deviated three standard deviations from baseline. To calculate the time to peak, we measured the time after the auditory speech onset at which the signal reached its maximum value. We also calculated the duration of the response curves. As a measure of response duration, we used full width at half maximum (FWHM), which was calculated by finding the width of the response curve at where the response is at 50% of the peak amplitude. We calculated the response onset, time to peak and response duration for each trial and then averaged across trials for each electrode.

Linear Mixed Effects Modeling

We used the *lme4* package (83) available for the R statistical language (R Core Team, 2015) to perform a linear mixed effect (LME) analysis of the relationship between the neural response and both fixed and random factors that may influence the response. For the main LME analyses (Tables 2.1 to 2.5), the fixed factors were the location of each electrode (Anterior or Posterior) the presence or absence of auditory noise and the presence or absence of visual noise. The random factors were the mean response of each electrode across all conditions and the stimulus exemplar. The use of stimulus exemplar as a random factor accounts for differences in response to individual stimuli and allows for inference beyond the levels of the factors tested in the particular experiment (*i.e.* generalization to other stimuli).

For each fixed factor, the LME analysis produced an estimated effect in units of the dependent variable and a standard error relative to a baseline condition (equivalent to beta weights in linear regression). For the main LME analyses, the baseline condition was always the response to AV speech in anterior electrodes. The full results of all LME analyses and the baseline condition for each analysis are shown in the tables and table legends.

Additional Experiment: Varying Levels of Auditory Noise

In an additional control experiment, we recorded responses to audiovisual speech with varying levels of auditory noise. Similar to the main experiment, for each auditory word, noise that matched the spectrotemporal power distribution of the auditory speech was generated, then noise and the original auditory speech

were added together with different weights while keeping the total power constant (70). We parametrically increased the amount of auditory noise in 11 steps from 0% to 100% in 10% increments. Forty-two to forty-four repetitions were presented for each noise level. The subject's task was to discriminate between four different words: Rain, Rock, Neck and Mouth.

Model Creation

A simple Bayesian model was constructed to aid in interpretation of the data (Figure 2.12) using a recently developed model of human multisensory speech perception (84). Briefly, the high dimensional neuronal response vector is conceptualized as a point in two-dimensional space. In this space, the x-axis represents auditory feature information and the y-axis represents visual feature information. Speech tokens are located at a fixed point in this space (shown in Figure 2.12 as the black dot at the center of each ellipse). For each presentation of an audiovisual speech stimulus, the brain encodes the auditory and visual information with noise. Over many trials, we characterize the distribution of the encoded speech stimulus as an ellipse. The axes of the ellipse correspond to the relative precision of the representation along each axis. Modalities are encoded separately, but through extensive experience with audiovisual speech, encoding a unisensory speech stimulus provides some information about the other modality. Although the results are robust across a range of parameters, for demonstration purposes, we assume that the variability of the preferred to non-preferred modality for audiovisual speech with clear auditory component is 2:1 (shown in Figure 2.12 as the asymmetry of the ellipses in the auditory and visual

representations). The integrated representation is formed according to Bayes rule which combines the two modalities into a single representation that has smaller variance than either of the component modalities: $S_{AV} = (S_A^{-1} + S_V^{-1})^{-1}$ (85). For audiovisual speech with noisy auditory component, we assume that the variability in the auditory representation increases by 150% while keeping the relative variability at the same ratio of 2:1 (shown in Figure 2.12 as larger ellipse). We model the visual representation of speech with noisy auditory component as being either identical to the representation of speech with clear auditory component or with a gain term that reduces variability by 50% (with the relative variability remaining at 2:1). The multisensory representation is calculated in the same fashion with and without gain.

Results

Across subjects, a total of 27 speech-responsive electrodes were identified on the STG. Using the posterior border of Heschl's gyrus as an anatomical landmark, 16 of these electrodes were located over anterior STG and 11 electrodes were located over posterior STG (Figure 2.2).

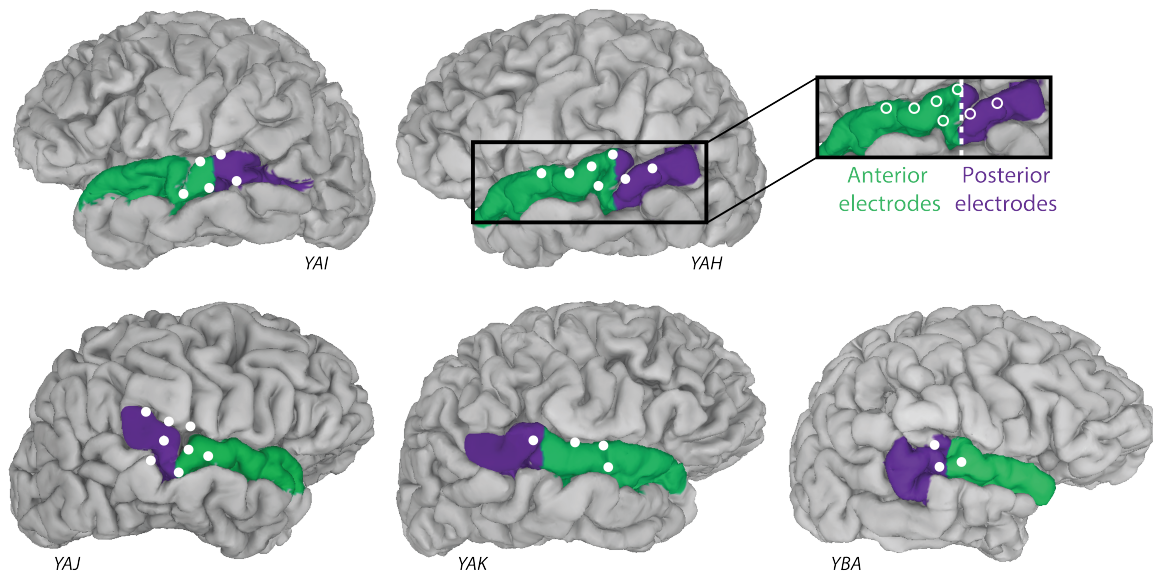


Figure 2.2 Location of electrodes on the superior temporal gyrus

Cortical surface models of the brains of five subjects (with anonymized subject ID). White circles show the location of implanted electrodes with a significant response to speech stimuli in the left hemisphere (top row) and right hemisphere (bottom row). In each hemisphere, the STG was parcellated into anterior (green) and posterior (purple) portions, demarcated by the posterior-most portion of Heschl's gyrus.

We hypothesized that the presence of noise in the speech stimulus might differentially affect responses in anterior and posterior electrodes. To test this hypothesis, we used the response amplitude in the gamma band as the dependent measure and fit a linear mixed-effects (LME) model with electrode location (Anterior vs. Posterior), the presence or absence of auditory noise in the stimulus (Clear A vs. Noisy A) and the presence or absence of visual noise in the stimulus (Clear V vs. Noisy V) as fixed factors. To account for overall differences in response amplitude across electrodes and stimulus exemplars, these were added to the model as random factors.

Amplitude of the Responses to Clear and Noisy Speech

As shown in Table 2.1, there were three significant effects in the LME model. There was a small but significant effect of electrode location ($p = 0.01$) driven by a smaller overall response in posterior electrodes (Anterior vs. Posterior: $136\% \pm 27\%$ vs. $101\% \pm 24\%$, mean signal change from baseline averaged across all stimulus conditions \pm SEM) and two larger effects: the main effect of auditory noise ($p = 10^{-14}$) and the interaction between auditory noise and the location of the electrode ($p = 10^{-10}$).

Fixed effects:	Estimate	Std. Error	DF	t-value	p-value
Baseline	183.1	24.8	33.7	7.4	10^{-8}
Auditory noise (An)	-109.6	13.5	188	-8.1	10^{-13}
Posterior location x An	140.6	21.2	188	6.6	10^{-10}
Posterior location	-101	38.7	34.2	-2.6	0.01
Visual noise (Vn)	21.6	13.5	188	1.6	0.11
An x Vn	-13.3	19.1	188	-0.7	0.49
Posterior location x Vn	-8.9	21.2	188	-0.4	0.67
Posterior location x An x Vn	3.6	29.9	188	0.1	0.91

Table 2.1 Linear mixed-effects model of the response amplitude

Results of an LME model of the response amplitude. The fixed effects were the location of each electrode (Anterior vs. Posterior), the presence or absence of auditory noise (An) in the stimulus and the presence or absence of visual noise (Vn) in the stimulus. Electrodes and stimulus exemplar were included in the model as random factors. For each effect, the model estimate (in units of % signal change) for that factor is shown relative to baseline, the response in anterior electrodes to clear audiovisual speech (AV stimulus condition). The “Std. Error” column shows the standard error of the estimate. The degrees of freedom (“DF”) t-value and p-value derived from the model were calculated according to the Satterthwaite approximation, as provided by the *lmerTest* package (86). The baseline is shown first, all other effects are ranked by absolute t-value. Significant effects are shown in bold. The significance of the baseline fixed effect

is grayed-out because it was pre-specified: only electrodes responding to this condition were included in the analysis.

Speech with clear auditory component evoked a larger response than speech with noisy auditory component (Clear A, consisting of the average of the AV and AVn conditions, $151\% \pm 27\%$ vs. Noisy A, consisting of the average of the AnV and AnVn conditions, $93\% \pm 14\%$, mean \pm SEM across electrodes) driving the main effect of auditory noise. However, the response patterns were very different in anterior and posterior electrodes, leading to the significant interaction in the LME model (Figure 2.3A). Speech with clear auditory component evoked a *larger* response than speech with noisy auditory component in anterior electrodes (Clear A vs. Noisy A: $194\% \pm 39\%$, vs. $78\% \pm 16\%$, mean \pm SEM across electrodes) but speech with clear auditory component evoked a *smaller* response than speech with noisy auditory component in posterior electrodes ($88\% \pm 23\%$ vs. $115\% \pm 25\%$).

To determine if the interaction between electrode location and the response to auditory noise was consistent, we plotted the amplitude of the response to Clear A vs. Noisy A for all electrodes using one symbol per electrode (Figure 2.3B). All of the anterior electrodes lay above the line of equality, indicating uniformly larger responses for Clear A, and all of the posterior electrodes lay on or below the line of equality, indicating similar responses for Clear A and Noisy A.

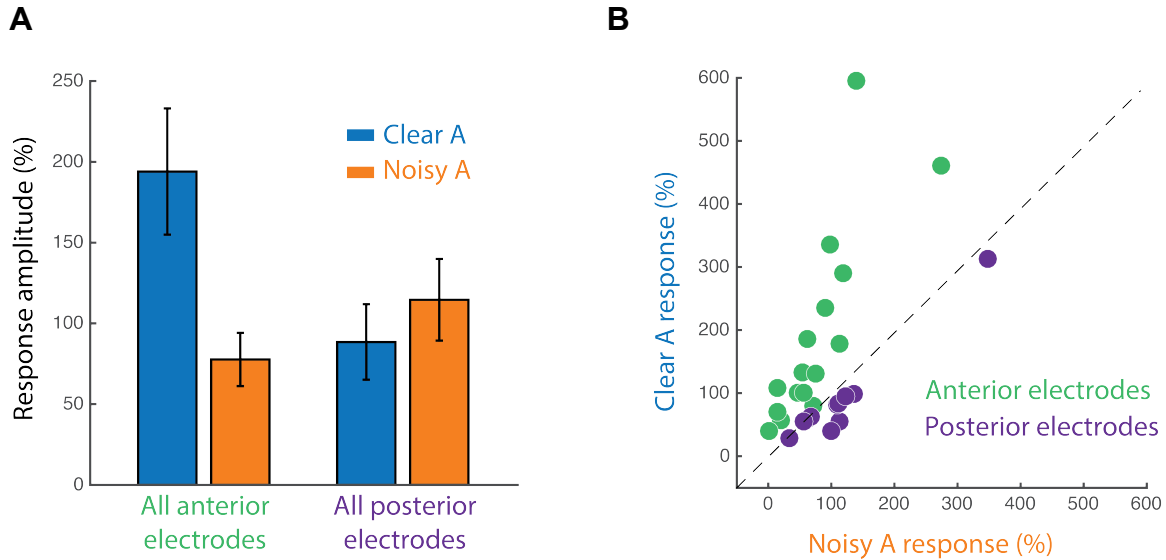


Figure 2.3 Response amplitudes in all STG electrodes

(A) The response to speech with clear auditory component (Clear A, combination of AV and AVn stimulus conditions) and noisy auditory component (Noisy A, combination of AnV and AnVn conditions) collapsed across electrodes (error bars show standard error of the mean). The response amplitude is the mean percent change in high-gamma power (70-110 Hz) in the 0 ms to 500 ms time window relative to prestimulus baseline (-500 to -100 ms).

(B) The response to Clear A vs. Noisy A speech for each individual electrode, with each anterior electrode shown as a green circle and each posterior electrode shown as a purple circle. The black dashed line represents the line of equality.

To examine the interaction between location and auditory noise in a single subject, we examined two electrodes: an anterior electrode located just anterior to the A-P boundary, and an adjacent electrode located 10 mm more posterior, just across the anterior-posterior boundary (Figure 2.4A and 2.4B). In the anterior electrode, the response to Clear A speech was much larger than the response to Noisy A speech (Clear A vs. Noisy A: $461\% \pm 35\%$ vs. $273\% \pm 21\%$, mean across trials \pm SEM; unpaired t -test across trials: $t_{147} = 4.6$, $p = 10^{-6}$) while in the adjacent posterior electrode, the response to Clear A speech was similar to the

response to Noisy A speech (Clear A vs. Noisy A: $313\% \pm 21\%$ vs. $349\% \pm 18\%$, $t_{147} = 1.3$, $p = 0.2$). Hence, two electrodes located on either side of the anterior-posterior boundary showed very different patterns of responses to Clear A and Noisy A speech.

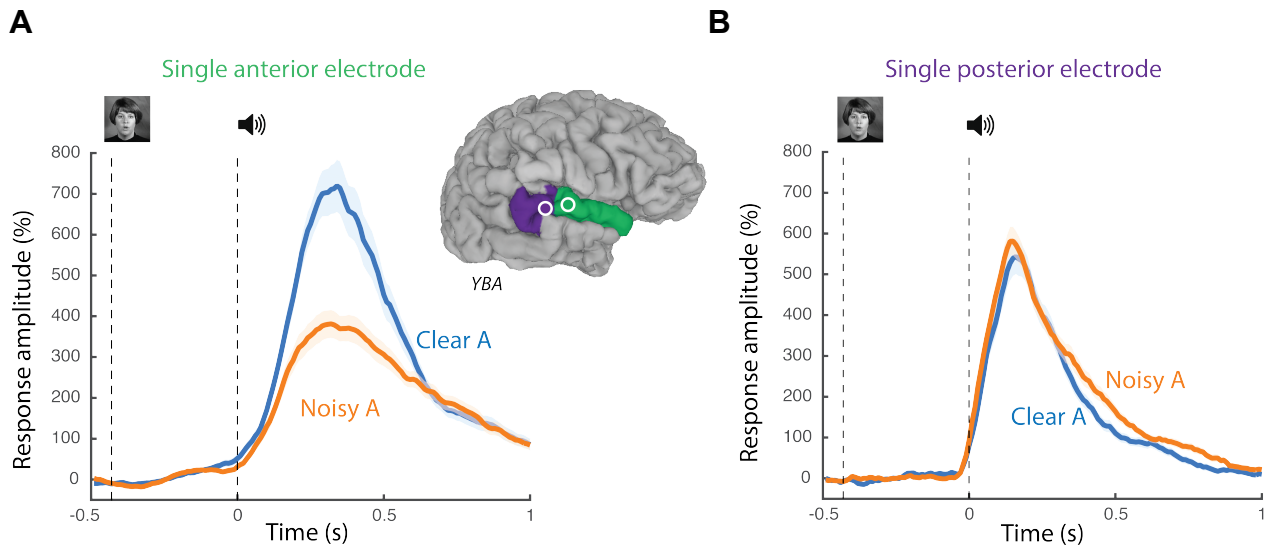


Figure 2.4 Response amplitudes in single STG electrodes

(A) High-gamma response to Clear A speech (blue trace) and Noisy A speech (orange trace) for a single anterior electrode (green electrode in inset brain). Shaded regions indicate the standard error of the mean across trials. Black vertical dashed lines indicate visual and auditory stimulus onsets, respectively.

(B) High-gamma response to Clear A and Noisy A speech in a single posterior electrode (purple electrode in inset brain).

To examine the effect of anatomical location on the response to Clear A and Noisy A speech in more detail, we calculated each electrode's location in a reference frame defined by the STG (Figure 2.5A) and the difference in the electrode's response amplitude to Clear A and Noisy A speech (Clear A – Noisy A). First, we examined electrodes sorted by their medial-to-lateral position on the STG and observed no discernible pattern (Figure 2.5B). Second, we examined

electrodes sorted by their anterior-to-posterior position on the STG (Figure 2.5C). Anterior electrodes showed uniformly positive values for Clear A – Noisy A (indicating larger responses for clear A) while posterior electrodes showed zero or negative values for Clear A – Noisy A (indicating similar or smaller responses for Clear A vs. Noisy A). However, we did not observe a gradient of responses between more anterior and more posterior electrodes, suggesting a sharp transition across the anterior-to-posterior boundary rather than a gradual shift in response properties along the entire extent of the STG. To quantify this observation, we tested two simple models. In the discrete model, there was a sharp transition between response properties on either side of the anterior-to-posterior boundary; in the continuous model, there was a gradual change in response properties across the entire extent of the STG (Figure 2.5D).

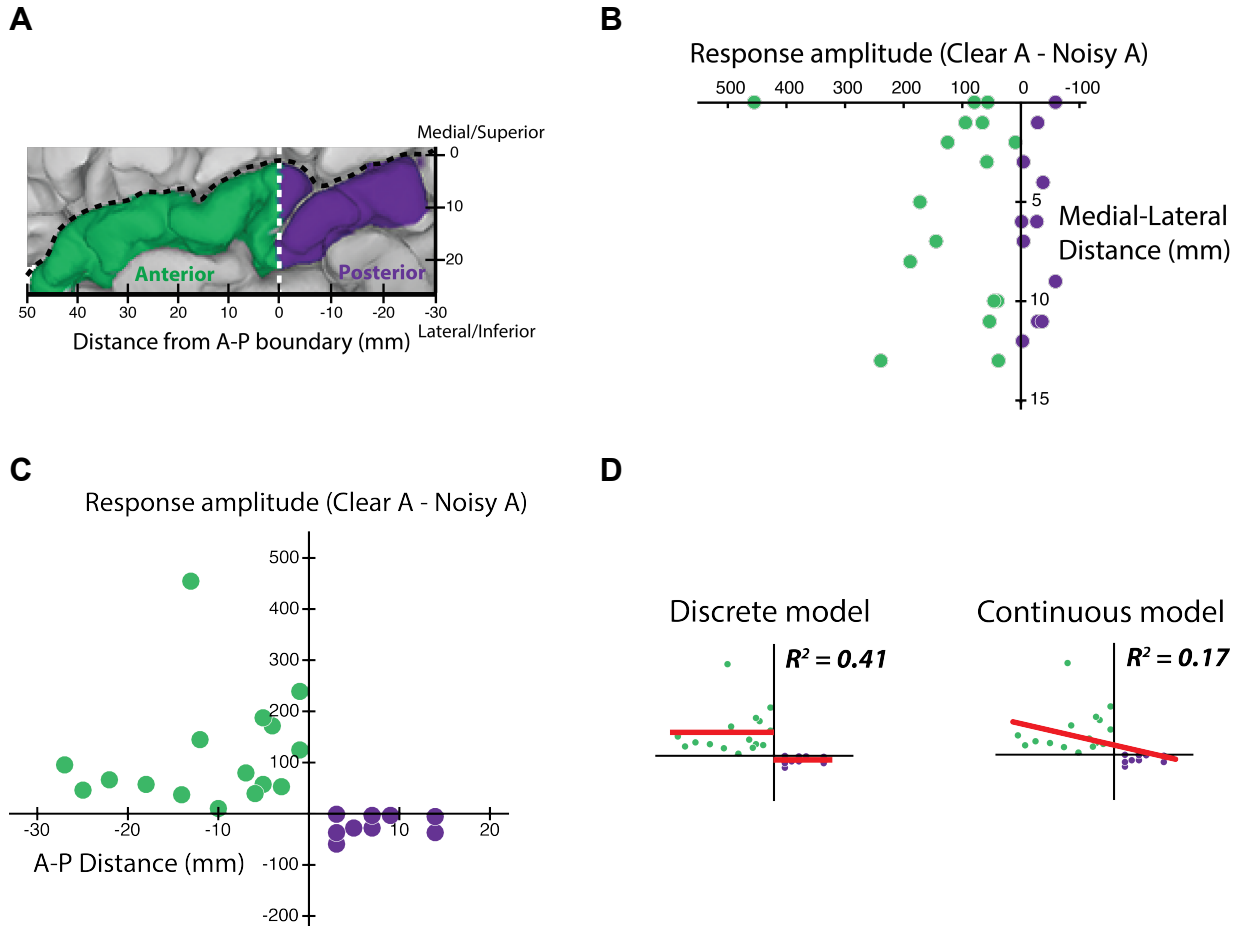


Figure 2.5 Response amplitudes with respect to location on the STG

(A) Co-ordinate system for STG measurements. Y-axis indicates distance from medial/superior border of STG (black dashed line), x-axis shows distance from the anterior-posterior border (white dashed line).

(B) The response amplitude to Clear A speech minus the response amplitude to Noisy A speech as a function of distance from the medial/superior border, one symbol per electrode (anterior electrodes in green, posterior electrodes in purple).

(C) The response amplitude to Clear A minus Noisy A speech as a function of distance from the anterior-posterior border.

(D) Discrete Model: Constant values were fit separately to the anterior and posterior electrode data in figure part **C** ($y = a$ and $y = b$) and the correlation with the data was calculated. Continuous Model: A linear model with two parameters was fit to both anterior and posterior electrodes ($y = mx+b$).

For the discrete model, we fit the amplitude vs. location points with two constants ($y = b$; horizontal lines with a fixed mean and zero slope, one mean for the anterior electrodes and one for the posterior electrodes). For the continuous model, we fit the amplitude vs. location points with a single line ($y = mx + b$). Both models fit the data using an equal number of parameters (2). The two models were compared using R^2 as a measure of the explained variance and Akaike Information Criterion (AIC) as a measure of likelihood. The discrete model fit the amplitude vs. location points much better than the continuous model ($R^2 = 0.41$ vs. 0.17) and the AIC revealed that the discrete model was more than 100 times more likely to explain the observed data ($e^{(AIC_{\text{continuous}} - AIC_{\text{discrete}})/2} = 102$).

To allow easier comparison of the A-P boundary with the functional neuroimaging literature, we converted each subject's brain into standard space and measured the co-ordinates of each electrode. The average location in standard space of the Heschl's gyrus landmark, the boundary between the anterior and posterior STG ROIs, was $y = -27 \pm 2$ (mean across subjects \pm SD). The mean position in standard space of all anterior electrodes was ($x = \pm 66$, $y = -18$, $z = 6$) while for posterior electrodes the mean position was ($x = \pm 67$, $y = -34$, $z = 12$).

Variability of the Responses to Clear and Noisy Speech

Theoretical models predict that combining the information available about speech content from the auditory and visual modalities should reduce neural variability (31, 33); see discussion and Figure 2.12 for more details. We

hypothesized that the presence of noise in the speech stimulus might differentially affect the response variability in anterior and posterior electrodes.

To test this hypothesis, we fit the same LME model used to examine response amplitude, except that response variability (CV) was used as the dependent measure. As shown in Table 2.2, there were three significant effects in the LME model, including an effect of electrode location ($p = 0.02$) driven by a larger overall response variability in posterior electrodes than in anterior electrodes (Anterior vs. Posterior: $0.85 \pm 24\%$ vs. 0.99 ± 0.1 , mean CV averaged across all stimulus conditions \pm SEM). The other two effects showed a larger effect size: the main effect of auditory noise ($p = 10^{-6}$) and the interaction between auditory noise and the location of the electrode ($p = 10^{-8}$).

Fixed effects:	Estimate	Std. Error	DF	t-value	p-value
Baseline	0.76	0.1	29.8	8	10^{-8}
Posterior location x An	-0.59	0.1	179.9	-5.7	10^{-7}
Auditory noise (An)	0.31	0.07	180.4	4.6	10^{-5}
Posterior location	0.35	0.14	39.8	2.5	0.02
Posterior location x Vn	-0.13	0.1	179.5	-1.3	0.2
Posterior location x An x Vn	0.15	0.15	179.5	1	0.31
An x Vn	0.03	0.09	179.6	0.3	0.77
Visual noise (Vn)	0.01	0.06	179.5	0.1	0.89

Table 2.2 Linear mixed-effects model of the response variability

Results of an LME model of the response variability, measures as coefficient of variation. The baseline for the model was the response in anterior electrodes to clear audiovisual speech (AV stimulus condition). Baseline is shown first, all other effects are ranked by absolute t-value. Significant effects are shown in bold.

Speech with noisy auditory component resulted in larger response variability than speech with clear auditory component (Clear A vs. Noisy A: 0.89 ± 0.06 vs. 0.93 ± 0.06 , mean \pm SEM across electrodes) driving the main effect of auditory noise in the model. However, the response patterns were very different in anterior and posterior electrodes, leading to the significant interaction (Figure 2.6A). Speech with noisy auditory component resulted in a *larger* response variability than speech with clear auditory component in anterior electrodes (Clear A vs. Noisy A: 0.73 ± 0.05 vs. 0.96 ± 0.1 , mean \pm SEM across electrodes) but speech with noisy auditory component resulted in a *smaller* response variability than speech with clear auditory component in posterior electrodes (Clear A vs. Noisy A: 1.1 ± 0.1 vs. 0.9 ± 0.1).

To determine if the interaction between electrode location and the response variability for auditory noise was consistent, we plotted the variability of the response to Clear A vs. Noisy A for all electrodes using one symbol per electrode (Figure 2.6B). Most of the anterior electrodes lay below the line of equality, indicating larger variability for Noisy A, while most of the posterior electrodes lay above the line of equality, indicating smaller variability for noisy A.

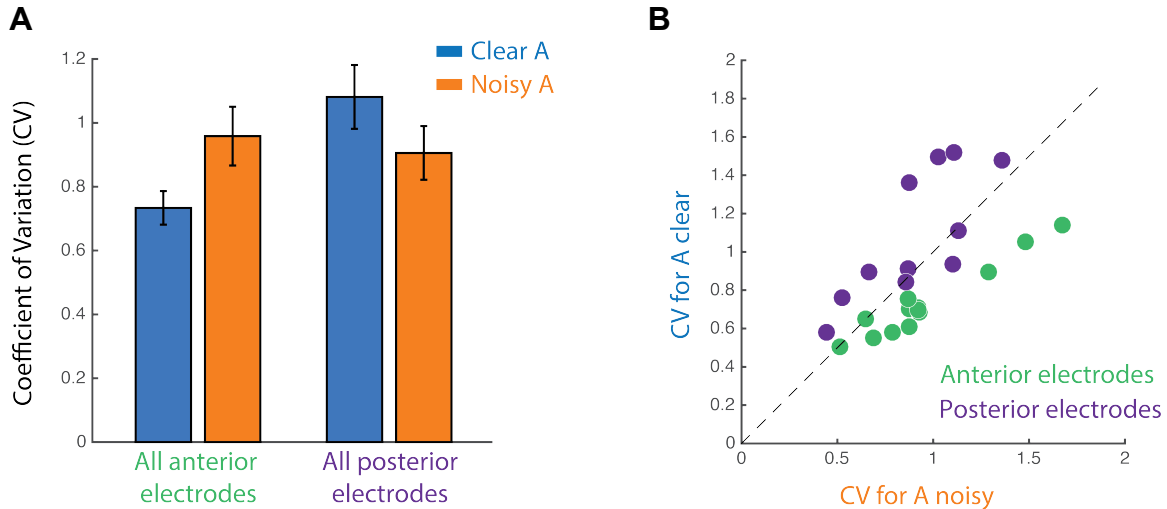


Figure 2.6 Response variability in all STG electrodes

(A) Response variability to speech with clear auditory component (Clear A, combination of AV and AVn stimulus conditions) and noisy auditory component (Noisy A, combination of AnV and AnVn conditions) collapsed across electrodes (error bars show standard error of the mean). The response variability was measured as the coefficient of variation, defined as the standard deviation of the high-gamma response divided by the mean of the high-gamma response; this measure accounts for the differences in the mean response between conditions shown in Figure 2.3.

(B) The response variability to Clear A vs. Noisy A speech for each individual electrode, with each anterior electrode shown as a green circle and each posterior electrode shown as a purple circle. The black dashed line represents the line of equality.

To demonstrate the effect at the single electrode level, we examined the interaction between location and auditory noise in a single subject, we examined two electrodes: an anterior electrode and a posterior electrode (Figure 2.7A and 2.7B). Figure 2.7A shows the normalized responses for a single anterior electrode for single trials of speech with clear and noisy auditory component. In this anterior electrode, there was variability across trials in both conditions, but the variability was much greater for speech with noisy auditory component than for speech with clear auditory component (Clear A vs. Noisy A: 1.1 vs. 1.7,

unpaired t -test across normalized trial amplitudes: $t_{221} = 5.4$, $p = 10^{-7}$). In a posterior electrode from the same subject (Figure 2.7B), the opposite pattern was observed: the variability was much greater for speech with clear auditory component than for speech with noisy auditory component (Clear A vs. Noisy A: 1.4 vs. 0.9, $t_{221} = 5$, $p = 10^{-6}$). Hence, two electrodes located on either side of the anterior-posterior boundary showed very different patterns of response variability.

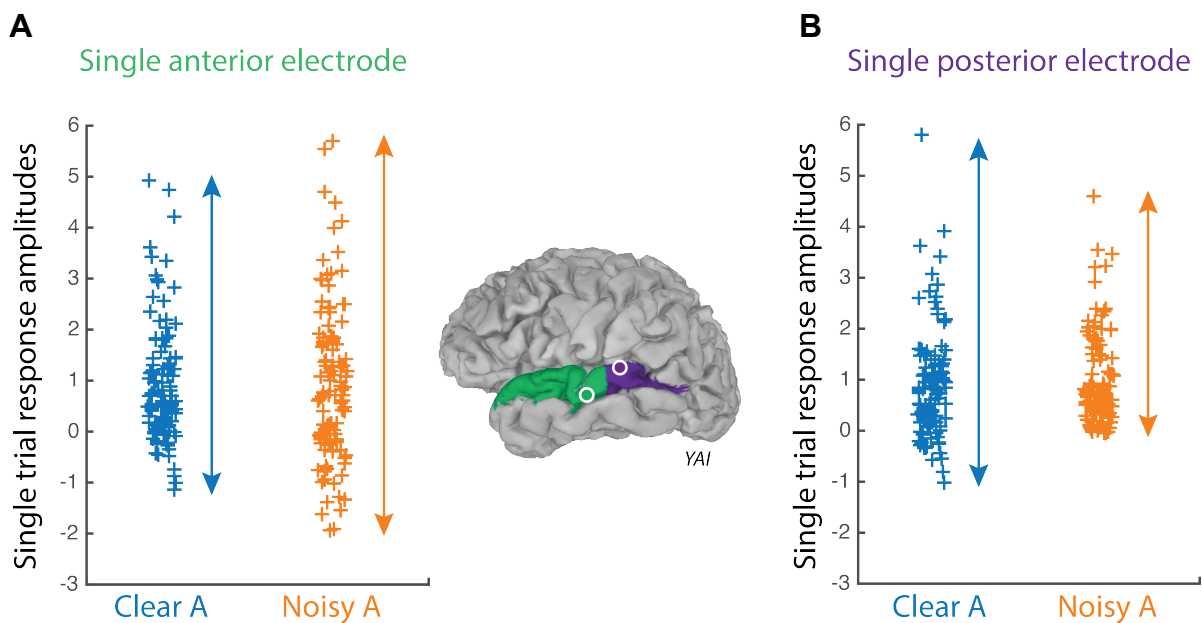


Figure 2.7 Response variability in single STG electrodes

(A) High-gamma response amplitudes to single presentations of Clear A speech (blue symbols) and Noisy A speech (orange symbols) for a single anterior electrode (green electrode in inset brain), normalized by the mean response across trials (value of one indicates a single trial response equal to the mean response across trials). Arrows illustrate coefficient of variation, a measure of variability.

(B) High-gamma response amplitudes to single presentations of speech for a single posterior electrode (purple electrode in inset brain).

To examine the effect of anatomical location on variability, we calculated the difference in each electrode's variability to Clear A and Noisy A speech (CV

for Clear A – CV for Noisy A) and plotted it against that electrode's anterior-posterior location on the STG (Figure 2.8A). Paralleling the analysis performed on response amplitude, discrete and continuous models were fit to the data (Figure 2.8B). The discrete model fit the amplitude vs. location points much better than the continuous model ($R^2 = 0.56$ vs. 0.37) and the AIC revealed that the discrete model was more likely to explain the observed data ($e^{(AIC_{continuous} - AIC_{discrete})/2} = 74$). Hence, the difference in response variability between electrodes is more accurately described as arising from two groups (Anterior and Posterior) with categorically different variability rather than as a continuous change in variability from anterior to posterior.

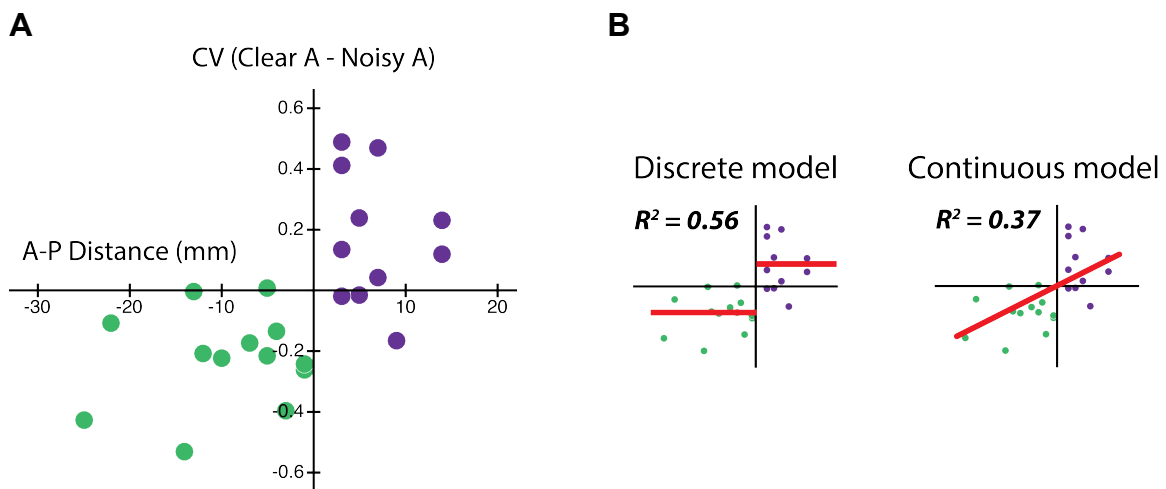


Figure 2.8 Response variability with respect to location on the STG

(A) The response variability to Clear A speech minus the response variability to Noisy A speech as a function of distance from the anterior-posterior border, one symbol per electrode (anterior electrodes in green, posterior electrodes in purple).

(B) Discrete Model: Constant values were fit separately to the anterior and posterior electrode data in figure part A ($y = a$ and $y = b$) and the correlation with the data was calculated. Continuous Model: A linear model with two parameters was fit to both anterior and posterior electrodes ($y = mx + b$).

Timing of the Responses to Clear and Noisy Speech

The high temporal resolution of ECoG allows for examination of the detailed timing of the neuronal responses. Figures 2.9A and 2.9B show the average response of anterior and posterior electrodes to Clear A and Noisy A speech. In anterior electrodes, the high-gamma response to Clear A speech started at 77 ms after auditory stimulus onset, reached half-maximum amplitude at 110 ms, peaked at 210 ms and returned to the half-maximum value at 290 ms, resulting in a total response duration (measured as the full width at half maximum, FWHM) of 190 ms.

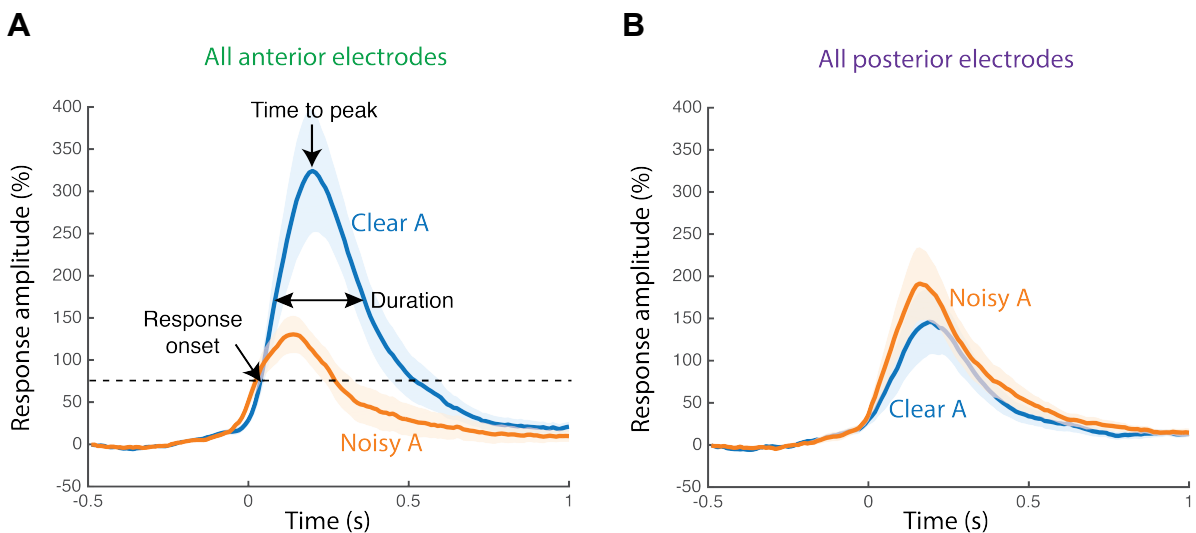


Figure 2.9 Response timing in STG electrodes

(A) High-gamma response amplitudes to Clear A and Noisy A speech averaged across all anterior electrodes, shown as percent signal change from baseline relative to time from auditory stimulus onset (error bars show standard error of the mean). Three measures of the response were calculated. Response onset time is the first time point at which the signal deviates three standard deviations from baseline. Time to peak is the time point of maximal response amplitude. Duration indicates the time between the first and last time points at which the response is equal to half of its maximum value (FWHM).

(B) High-gamma response amplitudes to Clear A and Noisy A speech averaged across all posterior electrodes.

To determine the effects of auditory noise and electrode location on the timing of the neuronal response, for each electrode we estimated response duration, onset time, and time-to-peak and separately fit three LME models with each temporal variable as the dependent measure. For the LME model with response duration as the dependent measure (Table 2.3 and Figure 2.10A) the only significant effects were the main effect of auditory noise ($p = 10^{-5}$) and the interaction between auditory noise and electrode location ($p = 10^{-5}$).

Fixed effects:	Estimate	Std. Error	DF	t-value	p-value
Baseline	206.2	9.6	41.4	21.4	10^{-16}
Posterior location x An	48.6	10.9	189	4.4	10^{-5}
Auditory noise (An)	-30.9	7	189	-4.4	10^{-5}
Posterior location	-15.1	15.1	41.4	-1	0.32
Posterior location x Vn	8.9	10.9	189	0.8	0.42
Posterior location x An x Vn	-12.2	15.5	189	-0.8	0.43
Visual noise (Vn)	-1.4	7	189	-0.2	0.84
An x Vn	-1.3	9.9	189	-0.1	0.89

Table 2.3 Linear mixed-effects model of the response duration

Results of an LME model of the response duration. The baseline for the model was the response in anterior electrodes to clear audiovisual speech (AV stimulus condition). Baseline is shown first, all other effects are ranked by absolute t-value. Significant effects are shown in bold.

These effects were driven by an overall longer response duration for Clear A speech than for Noisy A speech (Clear A vs. Noisy A: $194 \text{ ms} \pm 6 \text{ ms}$ vs. $187 \text{ ms} \pm 9 \text{ ms}$, mean across electrodes \pm SEM), with anterior electrodes showing *longer* responses for Clear A speech (Clear A vs. Noisy A: $205 \text{ ms} \pm 9 \text{ ms}$ vs.

174 ms \pm 14 ms) and posterior electrodes showing *shorter* responses for Clear A speech (Clear A vs. Noisy A: 195 ms \pm 7 ms vs. 206 ms \pm 7 ms).

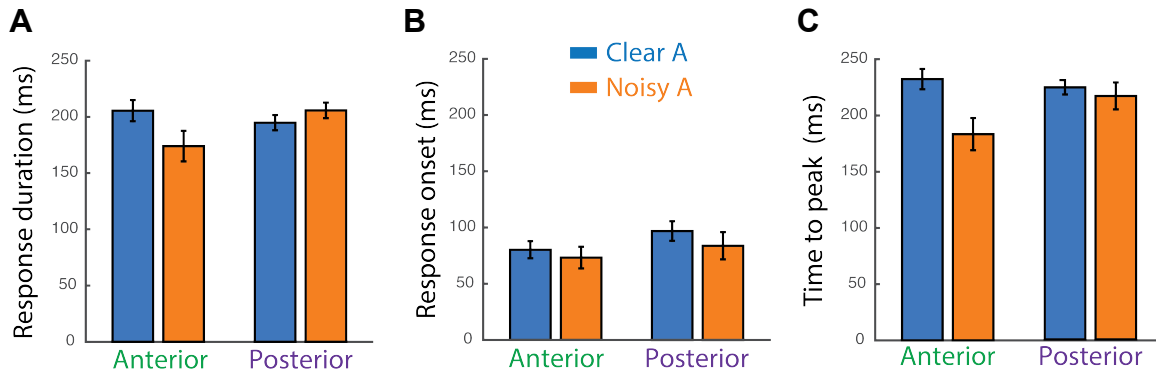


Figure 2.10 Response duration, response onset and time-to-peak in STG electrodes

(A) The response duration for Clear A vs. Noisy A speech in anterior electrodes (left) and posterior electrodes (right). Error bars show standard error of the mean.

(B) The response onset in anterior and posterior electrodes.

(C) The time to peak in anterior and posterior electrodes.

For the LME model with response onset as the dependent measure, there were no significant main effects or interactions (Table 2.4 and Figure 2.10B). For the LME model with time-to-peak as the dependent measure (Table 2.5 and Figure 2.10C), there was a significant main effect of auditory noise ($p = 10^{-9}$) and an interaction between auditory noise and electrode location ($p = 10^{-4}$) driven by a longer time-to-peak for Clear A speech (Clear A vs. Noisy A: 229 ms \pm 6 ms vs. 197 ms \pm 10 ms, mean across electrodes \pm SEM), more so in anterior electrodes (Clear A vs. Noisy A: 232 ms \pm 9 ms vs. 183 ms \pm 14 ms) than posterior electrodes (Clear A vs. Noisy A: 224 ms \pm 6 ms vs. 216 ms \pm 12 ms).

Fixed effects:	Estimate	Std. Error	DF	t-value	p-value
Baseline	81.5	9.2	27.6	8.8	10⁻⁹
Posterior location	17.6	13.6	41.3	1.3	0.2
Posterior location x An	-9.1	9.8	187.9	-0.9	0.35
An x Vn	-7.1	8.8	187.9	-0.8	0.42
Auditory noise (An)	-2.6	6.3	187.9	-0.4	0.68
Visual noise (Vn)	-2.6	6.3	187.9	-0.4	0.68
Posterior location x An x Vn	5	13.9	187.9	0.4	0.72
Posterior location x Vn	-1.3	9.8	187.9	-0.1	0.9

Table 2.4 Linear mixed-effects model of the response onset

Results of an LME model of the response onset. The baseline for the model was the response in anterior electrodes to clear audiovisual speech (AV stimulus condition). Baseline is shown first, all other effects are ranked by absolute t-value. No factors were significant.

Fixed effects:	Estimate	Std. Error	DF	t-value	p-value
Baseline	234.3	10.4	36	22.6	10⁻¹⁶
Auditory noise (An)	-46.5	7.4	187.9	-6.3	10⁻⁹
Posterior location x An	45.5	11.5	187.9	3.9	10⁻⁴
Posterior location	-12.5	15.8	41.6	-0.8	0.44
Posterior location x Vn	8.7	11.5	187.9	0.8	0.45
Visual noise (Vn)	-3.9	7.4	187.9	-0.5	0.6
Posterior location x An x Vn	-8.4	16.3	187.9	-0.5	0.61
An x Vn	-4.9	10.4	187.9	-0.5	0.64

Table 2.5 Linear mixed-effects model of the response peak time

Results of an LME model of the response peak time. The baseline for the model was the response in anterior electrodes to clear audiovisual speech (AV stimulus condition). Baseline is shown first, all other effects are ranked by absolute t-value. Significant effects are shown in bold.

Relationship Between Neuronal Responses and Perceptual Accuracy

Subjects performed a task which required them to respond to the identity of the word present in each trial. Across subjects, only AnVn trials consistently generated enough errors to compare correct and incorrect trials (AV: $99 \pm 3\%$, AVn: $98 \pm 3\%$, AnV: $81 \pm 20\%$, AnVn: $63 \pm 15\%$; % correct, mean across subjects \pm SD). To determine the relationship between neuronal response amplitude and behavioral accuracy within AnVn trials, an LME model was constructed with response amplitude as the dependent measure, electrode location (Anterior vs. Posterior) and behavioral accuracy (Correct vs. Incorrect) as fixed factors, and stimulus exemplar, subject, and electrode (nested within subject) as random factors (Table 2.6). In the LME model, the only significant effect was an interaction between electrode location and behavioral accuracy ($p = 0.01$) driven by *smaller* amplitudes in correct trials for anterior electrodes (Correct vs. Incorrect: $84\% \pm 15\%$ vs. $93\% \pm 20\%$, mean gamma power signal change relative to baseline across electrodes \pm SEM) but *larger* amplitudes in correct trials for posterior electrodes (Correct vs. Incorrect: $122\% \pm 27\%$ vs. $106\% \pm 26\%$). A similar model with CV as the dependent measure did not show any significant effects (Table 2.7).

Fixed effects:	Estimate	Std. Error	DF	t-value	p-value
Baseline	105.2	36.1	4.2	2.9	0.04
Incorrect responses x Posterior location	-25.6	10.2	65.8	-2.5	0.01
Incorrect responses	11.3	6.6	66.1	1.7	0.09
Posterior location	19.6	21.8	22.8	0.9	0.38

Table 2.6 Linear mixed-effects model of the effect of accuracy on response amplitude

Results of an LME model on the relationship between response amplitude and behavioral accuracy for auditory noise, visual noisy audiovisual speech (AnVn stimulus condition). The fixed effects were the location of each electrode (Anterior vs. Posterior) and the behavioral accuracy of the subject's responses (Correct vs. Incorrect). Subjects, electrodes nested in subjects and stimulus exemplar were included in the model as random factors. The baseline for the model was the response in anterior electrodes for correct behavioral responses. Baseline is shown first, all other effects are ranked by absolute t-value. Significant effects are shown in bold.

Fixed effects:	Estimate	Std. Error	DF	t-value	p-value
Baseline	1	0.19	7	5.4	10⁻³
Posterior location	-0.13	0.2	31.3	0.6	0.53
Incorrect responses	0.02	0.11	67.1	0.2	0.86
Incorrect responses x Posterior location	0.01	0.17	66.5	0.1	0.95

Table 2.7 Linear mixed-effects model of the effect of accuracy on response variability

Results of an LME model on the relationship between response variability (CV) and behavioral accuracy for auditory noise, visual noisy audiovisual speech (AnVn stimulus condition). The baseline for the model was the response in anterior electrodes for correct behavioral responses. Baseline is shown first, all other effects are ranked by absolute t-value. No factors were significant.

Potential Confound: Intelligibility

We observed very different neuronal responses to audiovisual speech with noisy auditory component in anterior compared with posterior electrodes, attributing this difference to the differential contributions of anterior and posterior STG to multisensory integration. However, we used only high levels of auditory noise in our audiovisual speech stimuli. To determine how the level of auditory noise influenced the effect, in one patient we presented audiovisual speech with

eleven different levels of auditory noise and examined the neural response in two electrodes located on either side of the anterior-posterior boundary (Figure 2.11A).

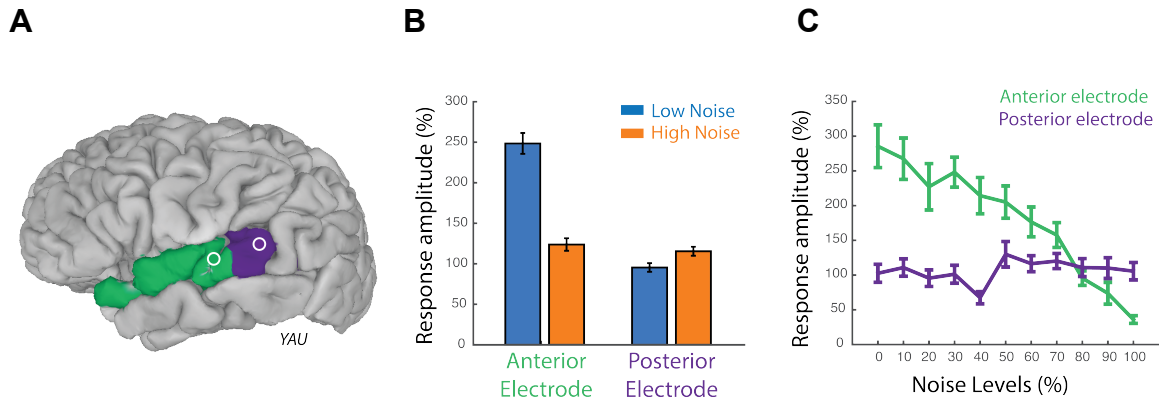


Figure 2.11 Response amplitudes in STG with varying auditory noise levels

- (A) The location of an anterior and a posterior electrode in a single subject.
- (B) The response amplitude in the anterior electrode (left bars) and posterior electrode (right bars) to audiovisual speech with low levels of auditory noise (Low Noise: 0% to 40%) and high levels of auditory noise (High Noise: 50% to 100%) averaged across trials (error bars show standard error of the mean).
- (C) Response amplitude for the anterior and posterior electrodes at each of 11 different auditory noise levels (0% to 100%) averaged across trials (error bars show standard error of the mean).

First, we examined how this data compares to our previous results by collapsing the eleven different levels of noise into just two categories “low noise” (0% - 40% noise levels) and “high noise” (50% - 100% noise levels) similar to our initial analysis of Clear A and Noisy A audiovisual speech. The responses were similar to that observed with just two levels of noise (compare Figure 2.11B and Figure 2.3A). An LME model fit to the data across the different noise levels (Table 2.8) showed significant effects of noise level ($p = 10^{-16}$), electrode location

($p = 10^{-16}$), and an interaction between noise level and location ($p = 10^{-16}$), driven by significantly greater response in anterior electrodes to low noise stimuli (Low vs. High: $248\% \pm 13\%$ vs. $124\% \pm 8\%$, mean across trials \pm SEM) and similar responses in posterior electrodes to low and high noise conditions (Low vs. High: $95\% \pm 5\%$ vs. $115\% \pm 5\%$).

Fixed effects:	Estimate	Std. Error	t-value	p-value
Baseline	248.5	8.6	28.9	10^{-16}
Posterior location	-153.1	12.2	-12.6	10^{-16}
High auditory noise	-124.8	11.6	-10.7	10^{-16}
High auditory noise x Posterior location	144.8	16.5	8.8	10^{-16}

Table 2.8 Linear model of the effect of varying auditory noise levels on response amplitude

Results of a linear model of the response amplitude for varying auditory noise levels in a single subject. Responses in individual trials were used as samples. Electrode location (Anterior vs. Posterior) and noise level (Low vs. High) were used as factors. The baseline for the model was the response in anterior electrodes to audiovisual speech with low auditory noise. Baseline is shown first, all other effects are ranked by absolute t-value. Significant effects are shown in bold. The significance of the baseline fixed effect is grayed-out because it was pre-specified: only electrodes responding to this condition were included in the analysis.

Next, we examined the response to each different level of auditory noise. In the anterior electrode, increasing levels of auditory noise led to smaller responses while in the posterior electrode, increasing levels of auditory noise led to similar or slightly larger gamma band responses (Figure 2.11C). We quantified this by fitting a line to the anterior and posterior electrode responses at 11 different auditory noise levels. The anterior electrode fit was significant ($R^2 = 0.9$,

$p = 10^{-6}$) with a negative slope ($m = -24$) while the posterior electrode fit was not significant ($R^2 = 0.07$, $p = 0.4$) with a slightly positive slope ($m = 1.32$).

The subject performed at a high level of accuracy even in trials with a high level of auditory noise (zero errors) demonstrating that the visual speech information was able to compensate for the increased levels of auditory noise.

Discussion

We observed a double dissociation in the responses to audiovisual speech with clear and noisy auditory components for both amplitude and variability measures. In anterior STG, the amplitude of the high-gamma response was greater for speech with clear auditory component than for speech with noisy auditory component, while in posterior STG responses were similar or slightly greater for speech with noisy auditory component. In anterior STG, the coefficient of variation across single trials was greater for speech with noisy auditory component, while in posterior STG it was greater for speech with clear auditory component.

These data are best understood within the framework of Bayes-optimal models of multisensory integration (87, 88) and speech perception (85, 89). In these models, different sensory modalities are posited to contain independent sources of environmental and sensory noise. Because of the independence of noise sources across modality, Bayesian integration results in a multisensory representation that has smaller variance than either of the unisensory variances (31, 33).

Recently, a Bayesian model of causal inference in audiovisual speech perception was proposed (84). Figure 2.12 shows an application of this model to our data. We assume that anterior STG contains a unisensory representation of auditory speech, that extrastriate visual areas contain a representation of visual speech and that posterior STG contains a representation of multisensory speech formed by integrating inputs from unisensory auditory and visual areas (41, 90). The neural implementation of Bayes-optimal integration is thought to rely on probabilistic population codes (32, 91) in which pools of neurons encode individual stimuli in a probabilistic fashion. These population codes are modeled as Gaussians in which amplitude and variability are inversely related. A smaller, more focal Gaussian indicates larger amplitude and less variability in the population code, while a larger Gaussian indicates smaller amplitude and more variability.

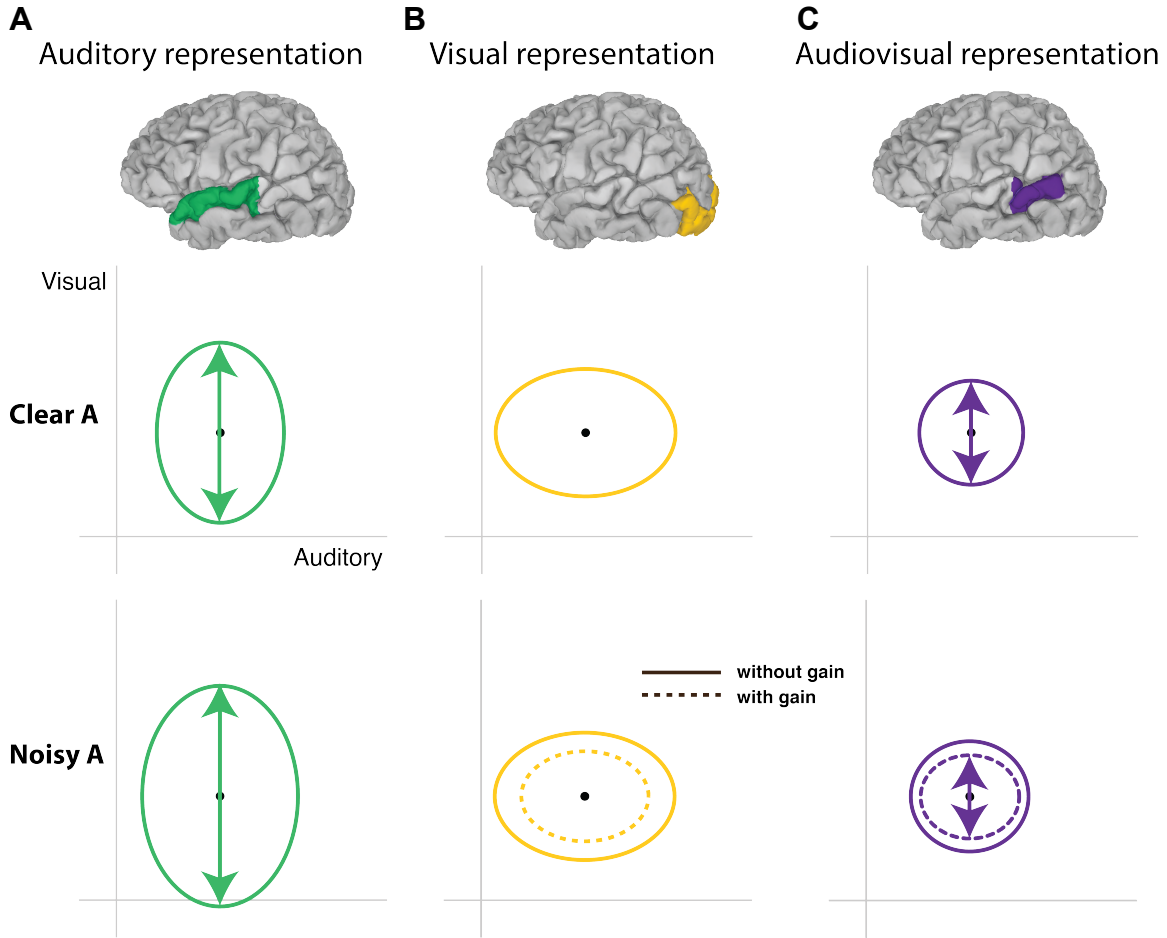


Figure 2.12 Bayesian model of audiovisual speech with auditory noise

(A) The model assumes that a neural representation of the auditory component of audiovisual speech exists in anterior STG (top row: brain region colored green). The high-dimensional neural representation is projected onto a two-dimensional space (middle and bottom rows) in which the x-axis represents auditory feature information and the y-axis represents visual feature information. The stimulus representation is shown as an ellipse indicating the cross-trial variability in representation of an identical physical stimulus due to sensory noise. For audiovisual speech with clear auditory component (Clear A) in anterior STG (green ellipse in middle row) there is less variability along the auditory axis and more variability along the visual axis, indicated by the shape of the ellipse. For audiovisual speech with noisy auditory component (Noisy A) in anterior STG (green ellipse in bottom row), there is greater variability along both axes due to the added stimulus noise (see Methods for details).

(B) The model assumes that a neural representation of the visual component of audiovisual speech exists in lateral extrastriate visual cortex (top row: brain region colored yellow). In the visual representation, there is less variability along the visual axis and more variability along the auditory axis, indicated by the

shape of the ellipse. For audiovisual speech with noisy auditory component (Noisy A) the visual component of the speech is identical, so the representation should be identical (yellow ellipse in bottom row). However, evidence from Schepers and colleagues (36) demonstrates that response in visual cortex to Noisy A speech is actually greater than to Clear A speech, suggesting an increase in gain due to attentional modulation or other top-down factors. The representation with gain modulation is shown with the dashed yellow ellipse.

(C) The model assumes that a neural representation that integrates both auditory and visual components of audiovisual speech exists in posterior STG (top row: brain region colored purple). Due to the principles of Bayesian integration, this representation has smaller variability than either the auditory representation or the visual representation (compare size of purple ellipse in each row with green and yellow ellipses). Assuming gain modulation, the integrated representation of Noisy A speech (dashed purple ellipse in bottom row) has smaller variability than the representation of Clear A speech (purple ellipse in middle row).

For audiovisual speech with a clear auditory component (Clear A), the neural population code in anterior STG has a given amplitude and variability.

When auditory noise is added (Noisy A), the population code amplitude decreases and the variability increases (32), an accurate description of the response in anterior STG for noisy compared with clear auditory speech.

For the visual representation in lateral extrastriate cortex, the visual information is the same in the Clear A and Noisy A conditions, predicting similar population codes for both conditions. For the multisensory representation in posterior STG, the population code is calculated as the optimal integration of the response in auditory and visual representations. The visual information serves to compensate for the increased auditory noise in the Noisy A condition, so that the population code for the integrated representation is only slightly broader for Noisy A than Clear A speech, a match to the observation that the amplitude and

variability of the response to Noisy A and Clear A speech are much more similar in posterior STG than they are in anterior STG.

A close inspection of the data shows that, contrary to Bayesian models, the response in posterior STG was slightly *more* focal (30% greater amplitude and 16% reduced variability) for Noisy A compared with Clear A conditions. While counter-intuitive, this result is consistent with evidence that visual cortex responds more to noisy than clear audiovisual speech (36). This enhancement may be attributable to top-down modulation from higher-level areas that increase the gain in visual cortex, similar to attentional modulation in which representations in visual cortex are heightened and/or sharpened by spatial or featural attention (92, 93). This gain increase would be adaptive because it would increase the likelihood of decoding speech from visual cortex under conditions of low or no auditory information, at the cost of additional deployment of attentional and neural resources. We implemented this gain modulation in our Bayesian model as reduced variance in the visual representation for Noisy A compared with Clear A speech. When this reduced variance visual representation is integrated with the noisy auditory representation, the resulting multisensory representation becomes more focal for Noisy A than Clear A speech, a fit to the observed increased amplitude and reduced variability for Noisy A compared with Clear A speech in posterior STG.

While the Bayesian model provides a conceptual framework for understanding how multisensory integration could affect the amplitude and variance of neuronal population responses, it is agnostic about the actual

stimulus features important for integration. We did not observe a main effect of visual noise (or an interaction between visual noise and auditory noise) in the LME analysis on amplitude and variance (Table 2.1 and Table 2.2). Most of the relevant information provided by the visual signal during auditory-visual speech perception is related to the timing of mouth opening and closing relative to auditory speech. The blurring procedure used to generate the noisy visual speech may leave this timing information intact, rendering it less noisy than expected.

Our Bayesian model also does not make explicit predictions about the latency or duration of the neuronal response. However, we observed the same pattern of double dissociation between anterior and posterior STG for response duration as in other response measures. At the high levels of auditory noise used in our experiments, the auditory representation contains little useful information, so it would be adaptive for top-down modulation to decrease both the amplitude and duration of activity in the anterior STG auditory representation for Noisy A speech. Interestingly, the absolute duration of the response in posterior STG during Noisy A speech was the same as the absolute duration of the response in anterior STG during Clear A speech (210 ms), raising the possibility that this is the time-frame of the selection process in which the competing unisensory and multisensory representations are selected for perception and action.

An interaction between electrode location and response amplitude was also observed in an analysis of perceptual accuracy (only speech with both noisy auditory and noisy visual component generate enough errors for this analysis). In

anterior electrodes, responses were larger for incorrect trials, while in posterior electrodes responses were larger for correct trials. This supports the idea that posterior regions are particularly important in the perception of noisy speech, with larger amplitude indicating a more focal peak of activity in the population code and less uncertainty about the presented stimulus.

Anterior vs. Posterior Anatomical Specialization

There was a strikingly sharp boundary between the anterior and posterior response patterns, suggesting that anterior and posterior STG are functionally distinct. Although the posterior two thirds of the STG is classically defined as Brodmann area 22, a previous study that combines cytoarchitectonic and receptorarchitectonic mapping identified a distinct cortical area on the posterior border of the STG, which is called the area Te3 (94). Supporting our finding, this study provided anatomical evidence for an anterior posterior specialization within the STG.

We divided STG at the posterior border of Heschl's gyrus (mean $y = -27$), a landmark that also has been used in previous neuroimaging studies of speech processing (95-97). A functional division in STG near Heschl's gyrus is consistent with the division of the auditory system into two processing streams, one of which runs anterior-ventral from Heschl's gyrus and one of which runs posterior-dorsal (65, 98). These two streams are often characterized as specialized for processing "what" or object identity features (anterior-ventral) and "where" or object location features (posterior-dorsal) by analogy with the different streams of

visual processing (99). However, these labels do not neatly map onto an anterior preference for clear speech and a posterior preference for noisy speech (100) and may reflect preferences for different rates of spectrotemporal modulation (101).

While we are not aware of previous studies examining changes in the neural variability to Clear A and Noisy A audiovisual speech, a number of neuroimaging studies have reported anterior-to-posterior differences in the amplitude of the neural response to Clear A and Noisy A audiovisual speech. Stevenson and James (102) presented clear audiovisual speech and audiovisual speech with noise added to both modalities (noisy auditory + noisy visual), contrasting both against a standard baseline condition consisting of simple visual fixation. Anterior regions of STG/STS showed greater responses to clear than noisy audiovisual speech (Figure 5C and Table 3 in their paper, $y = -20$ compared with $y = -18$ in the present study, mean across left and right hemispheres) while posterior regions (Figure 5D and Table 1 in their paper, $y = -37$, compared with $y = -34$ in the present study) showed similar responses to clear and moderately noisy audiovisual speech. This single dissociation differs from our finding of a double dissociation and results from the relatively weak responses to noisy speech observed by Stevenson and James in posterior STG/STS. This could be explained by their use of noisy auditory + noisy visual speech vs. our use of noisy auditory + clear visual speech: if posterior regions respond to both auditory and visual speech information, degraded visual

information might be expected to reduce response amplitudes in posterior regions.

Consistent with these results, Lee and Noppeney (103) found that anterior STG/STS ($y = -16$, their Table 2) showed significant audiovisual interactions only for clear speech while posterior STG/STS (mean $y = -36$, Table 2 in their paper) showed interactions for both clear and noisy audiovisual speech.

Bishop and Miller (104) reported greater responses to clear vs. noisy audiovisual speech in anterior regions of STG (Table 1 in their paper, $y = -13$ mean across left and right hemispheres) while McGettigan and colleagues (105) reported greater responses for clear than noisy audiovisual speech in both anterior STG ($y = -12$, Table 1 in their paper) and posterior STG ($y = -42$).

While most neuroimaging studies have reported greater responses to clear than noisy audiovisual speech, two studies have reported the opposite result of greater responses to noisy speech in the STG (30, 38). However, the interpretation of these studies is complex. Sekiyama and colleagues tested clear and noisy speech consisting of McGurk syllables and incongruent audiovisual speech in which the auditory and visual components do not match (including McGurk syllables) are known to evoke responses in STS that are both different from congruent syllables and vary markedly from subject-to-subject (106, 107). Callan and colleagues performed an analysis in which they first subtracted the response to auditory-only clear speech from the response to audiovisual clear speech; then subtracted the response to auditory-only noisy speech from the response to audiovisual noisy speech; and finally subtracted the two differences.

Without a direct comparison between clear and noisy audiovisual speech, it is possible that the reported preference for noisy audiovisual speech was driven by the intermediate analysis step in which the auditory-only response was subtracted from the audiovisual response. For instance, even if clear and noisy audiovisual speech evoked the exact same response, a weak response to auditory-only noisy speech and a strong response to auditory-only clear speech (a pattern observed in a number of studies, see below) would result in the reported greater response to noisy audiovisual speech.

The idea of an anterior-to-posterior double dissociation is also generally supported by the neuroimaging literature examining brain responses to clear and noisy auditory-only speech, although the many differences in the stimulus materials, task manipulations, and data analysis strategies makes direct comparisons difficult. Obleser and colleagues (108) reported a double dissociation, with posterior regions ($y = -26$, Table 1 in their paper) preferring noisy speech to clear speech, while anterior regions ($y = -18$) preferred clear speech to noisy speech. A double dissociation was also reported by Du and colleagues (109): anterior regions of STG ($y = -15$, Table S2 in their paper) showed greater BOLD amplitude with less auditory noise while posterior regions ($y = -32$) showed greater BOLD amplitude with more auditory noise. Similarly, Wild and Johnsrude (110) found that anterior regions of STG ($y = -12$, Table 1 in their paper) preferred clear to noisy speech, while posterior regions ($y = -30$) preferred noisy speech to clear speech.

Single dissociations consistent with an anterior preference for clear speech are also common in the literature. Scott and colleagues (111) found that anterior regions ($y = -12$) showed greater response amplitudes for clear speech while posterior regions ($y = -38$, Figure 2A in their paper) showed similar response amplitudes. Giraud and colleagues (112) also reported greater response amplitudes for clear than noisy speech in anterior STG (Table 1 in their paper, $y = -4$ mean across left and right hemispheres) but not posterior STG.

CHAPTER 3: PROCESSING OF SILENT SPEECH IN VISUAL CORTEX

Introduction

Speech perception is multisensory: humans combine visual information from the talker's face with auditory information from the talker's voice to aid in perception. However, the contribution of visual information to speech perception is influenced by two factors. First, if the auditory information is noisy or absent, visual speech is more important than if the auditory speech is clear. Current models of speech perception assume that top-down processes serve to incorporate this factor into multisensory speech perception. For instance, visual cortex shows enhanced responses to audiovisual speech containing a noisy or entirely absent auditory component (36) raising an obvious question: since visual cortex presumably cannot assess the quality of auditory speech, where is the top-down modulation that enhances visual speech processing originating? Neuroimaging studies have shown that speech reading (perception of visual-only speech) leads to strong responses in frontal regions including the inferior frontal gyrus, premotor cortex, frontal eye fields and dorsolateral prefrontal cortex (37, 47, 103, 113, 114). Especially, frontal eye fields and dorsolateral prefrontal cortex constitute the major components of the dorsal attention network and play a key role in the visual spatial attention. These regions activate when attention is overtly or covertly directed to a specific location in space (61, 115, 116). Directed functional connectivity studies showed that frontal eye fields exert top-down influence on visual cortex during spatial attention (117). Concurrent TMS-fMRI studies provided more causal evidence by demonstrating that applying TMS over frontal eye fields modulates activity in the visual cortex (118). We predicted that

these frontal regions that comprise the dorsal attention network serve as the source of a control signal, enhancing activity in visual cortex when auditory speech is noisy or absent.

Second, visual information about the content of speech is not distributed equally throughout the visual field. Some regions of the talker's face are more informative about speech content, with the mouth of the talker carrying the most information. When presented with noisy speech, humans foveate the mouth of the talker to enhance comprehension (62, 119). Therefore, if frontal cortex enhances visual responses during audiovisual speech perception, one should expect this enhancement to occur preferentially in the central portion of the visual field.

The relationship between these two factors within the neural substrates of speech perception is unknown. We sought to link these two factors by testing the hypothesis that when visual speech is critical to extract meaning from speech, top-down circuitry is engaged to enhance visual cortex responses to the mouth of the talker, helping to make speech intelligible.

Methods

Subject Information

All experimental procedures were approved by the Institutional Review Board of Baylor College of Medicine. Eight human subjects with refractory epilepsy (5F, mean age 38, 5L hemisphere) were implanted with subdural electrodes guided by clinical requirements. Following surgery, subjects were

tested while resting comfortably in their hospital bed in the epilepsy monitoring unit.

Experiment Setup

Visual stimuli were presented on an LCD monitor (Viewsonic VP150, 1024 x 768 pixels) positioned at 57 cm distance from the subject, resulting in a display size of 30.5° x 22.9°.

Receptive Field Mapping Procedures

Mapping stimulus consisted of a square checkerboard pattern (3° x 3° size) briefly flashed (rate of 2 Hz and a duty cycle of 25 %) in different positions on the display monitor to fill a grid over the region of interest in the visual field (63 positions, 7 x 9 grid). 12-30 trials for each position were recorded.

Subjects fixated at the center of the screen and performed a letter detection task to ensure that they were not fixating on the mapping stimulus. Different letters were randomly presented at the center of the screen (2° in size presented at a rate of 1-4 Hz) and they were required to press a mouse button whenever the letter “X” appeared. The mean accuracy was 88 ± 14 % with a false alarm rate of 8 ± 14 % (mean across subjects \pm SD; responses were not recorded for one subject).

Speech Experiment Procedures

Four video clips of a female talker pronouncing the single syllable words “drive”, “known”, “last” and “meant” were presented under audiovisual (AV), visual (*Vis*) and auditory (*Aud*) conditions. Visual stimuli were presented using

the same monitor used for receptive field mapping, with the face of the talker subtending approximately 13 degrees horizontally and 21 degrees vertically. Speech sounds were played through loudspeakers positioned next to the subject's bed. The average duration of the video clips was ~1500 ms (drive: 1670 ms, known: 1300 ms, last: 1500 ms, meant: 1400 ms). In *AV* and *Vis* trials, mouth movements started at ~200 ms after the video onset on average (drive: 200 ms, known: 233 ms, last: 200 ms, meant: 200 ms). Sound duration was ~480 ms on average (drive: 500 ms, known: 400 ms, last: 530 ms, meant: 500 ms).

The three different conditions were randomly intermixed, separated by interstimulus intervals of 2.5 s. 32-64 repetitions for each condition were presented. Subjects were instructed to fixate either the mouth of the talker (during *Vis* and *AV* trials) or a white fixation dot presented at the same location as the mouth of the talker on a gray background (during *Aud* trials and the interstimulus intervals). To ensure attention to the stimuli, subjects were instructed to press a mouse button on 20% of trials in which a catch stimulus was presented, consisting of the *AV* word "PRESS". The mean accuracy was $88 \pm 18\%$, with a false alarm rate of $3 \pm 6\%$ (mean across subjects \pm SD; for one subject, button presses were not recorded).

Electrode Localization and Recording

Before surgery, T1-weighted structural magnetic resonance imaging scans were used to create cortical surface models with FreeSurfer (71, 72) and visualized using SUMA (73). Subjects underwent a whole-head CT after the electrode implantation surgery. The post-surgical CT scan and pre-surgical MR

scan were aligned using AFNI (74) and all electrode positions were marked manually on the structural MR images. Electrode positions were then projected to the nearest node on the cortical surface model using the AFNI program *SurfaceMetrics*. Resulting electrode positions on the cortical surface model were confirmed by comparing them with the photographs taken during the implantation surgery.

A 128-channel Cerebus amplifier (Blackrock Microsystems, Salt Lake City, UT) was used to record from subdural electrodes (Ad-Tech Corporation, Racine, WI) that consisted of platinum alloy discs embedded in a flexible silicon sheet. Two types of electrodes were implanted, containing an exposed surface of either 2.3 mm or 0.5 mm; an initial analysis did not suggest any difference in the responses recorded from the two types of electrodes so they were grouped together for further analysis. An inactive intracranial electrode implanted facing the skull was used as a reference for recording. Signals were amplified, filtered (low-pass: 500 Hz, fourth-order Butterworth filter; high-pass: 0.3 Hz, first-order Butterworth) and digitized at 2 kHz. Data files were converted from Blackrock format to MATLAB 8.5.0 (MathWorks Inc. Natick, MA) and the continuous data stream was divided into trials. All analyses were conducted separately for each electrode.

Receptive Field Mapping Analysis

The voltage signal in each trial (consisting of the presentation of a single checkerboard at a single spatial location) was smoothed using a Savitzky-Golay polynomial filter (“sgolayfilt” function in Matlab) with polynomial order set to 5

and frame size set to 11. If the raw voltage exceeded a threshold of 3 standard deviations from the mean voltage, suggesting noise or amplifier saturation, the trial was discarded; < 1 trial per electrode discarded on average. The filtered voltage response at each spatial location was averaged, first across trials and then across time-points (from 100 to 300 ms post-stimulus) resulting in a single value for response amplitude; these values were then plotted on a grid corresponding to the visual field. A two-dimensional Gaussian function was fit to the responses to approximate the average receptive field of the neurons underlying the electrode. For any given electrode, a high correlation between the fitted Gaussian and the raw evoked potentials indicated an accurate localization of the receptive field. A conservative threshold of $r > 0.7$ was used to find only electrodes with high-amplitude, focal receptive fields (63). The center of the fitted Gaussian was used as the estimate of the RF center of the neurons underlying the electrode.

Speech Analysis

While for the RF mapping analysis, we used raw voltage as our measure of neural response, speech stimuli evoke a long lasting response that is not well captured by evoked potentials. Therefore, our primary measure of neural activity was the broad-band (non-synchronous) response in the high-gamma frequency band, ranging from 70 to 150 Hz. This response is thought to reflect action potentials in nearby neurons (54, 55, 76, 77).

To calculate broadband high-gamma response, the average signal across all electrodes was subtracted from each individual electrode's signal (common

average referencing), line noise at 60, 120, 180 Hz was filtered and the data was transformed to time–frequency space using the multitaper method available in the FieldTrip toolbox (75) with 3 Slepian tapers; frequency window from 10 to 200 Hz; frequency steps of 2 Hz; time steps of 10 ms; temporal smoothing of 200 ms; frequency smoothing of ± 10 Hz.

The high-gamma response (70 – 150 Hz) at each time point following stimulus onset was measured as the percent change from baseline, with the baseline calculated over all trials and all experimental conditions in a time window from –500 to –100 ms before stimulus onset. To reject outliers, if at any point following stimulus onset the response was greater than ten standard deviations from the mean calculated across the rest of the trials, the entire trial was discarded (average of 10 trials were discarded per electrode, range from 1 to 16).

To determine if electrodes responded to visually-presented faces, the mean high-gamma response (100 to 500 ms after stimulus onset) was compared with the pre-stimulus response (–500 to –100 before stimulus onset) across all AV and Vis trials using an unpaired t-test. Electrodes exceeding a significance threshold of $q < 0.01$, false-discovery rate corrected were considered responsive.

Electrode Selection and Linear Mixed Effects Modeling

Out of 154 total occipital lobe electrodes, we selected 73 electrodes that had well-demarcated spatial receptive fields and responded to talking faces. In each of the 8 subjects, we selected the single prefrontal electrode (located on or near precentral gyrus) that showed the largest response to AV speech.

We used the *lme4* package (83) available for the R statistical language (R Core Team, 2015) to perform a linear mixed effect (LME) analysis of the neural responses in each electrode. For each fixed factor, the LME analysis produces an estimated effect in units of the dependent variable (equivalent to beta weights in a linear regression) that is relative to an arbitrary baseline condition (defined in our analysis as the response to AV speech) and a standard error.

Functional Connectivity Analysis

The average high-gamma power in the 200-1500 milliseconds was calculated for each trial. This time interval corresponds to the period where mouth movements occur in the speech stimuli. After calculating the average broadband (70-150 Hz) power for each trial, functional connectivity between the 73 total frontal-visual cortex electrode pairs was measured by calculating the trial-by-trial Spearman rank correlation across trials of the same speech condition (AV, Vis or Aud).

Results

In eight patients, we measured the receptive fields of neurons underlying electrodes implanted over visual cortex by presenting small checkerboards at different visual field locations and determining the location with the maximum evoked response (Figure 3.1A and 3.1B).

In the talking face stimulus, the mouth subtended approximately 5 degrees of visual angle (Figure 3.1C). Electrodes were classified into two groups by their receptive field centers: central electrodes ($<5^\circ$) that would be expected to represent the talkers face ($n = 49$) and peripheral electrodes ($>5^\circ$) that would not

($n = 24$). Finally, we compared the responses to speech (Figure 3.1D) using a linear mixed-effects (LME) model with the broadband response amplitude as the dependent measure, the RF location of each visual electrode (central or peripheral) and stimulus condition (AV, *Vis* or *Aud*) as fixed factors, and the central response to AV speech as the baseline.

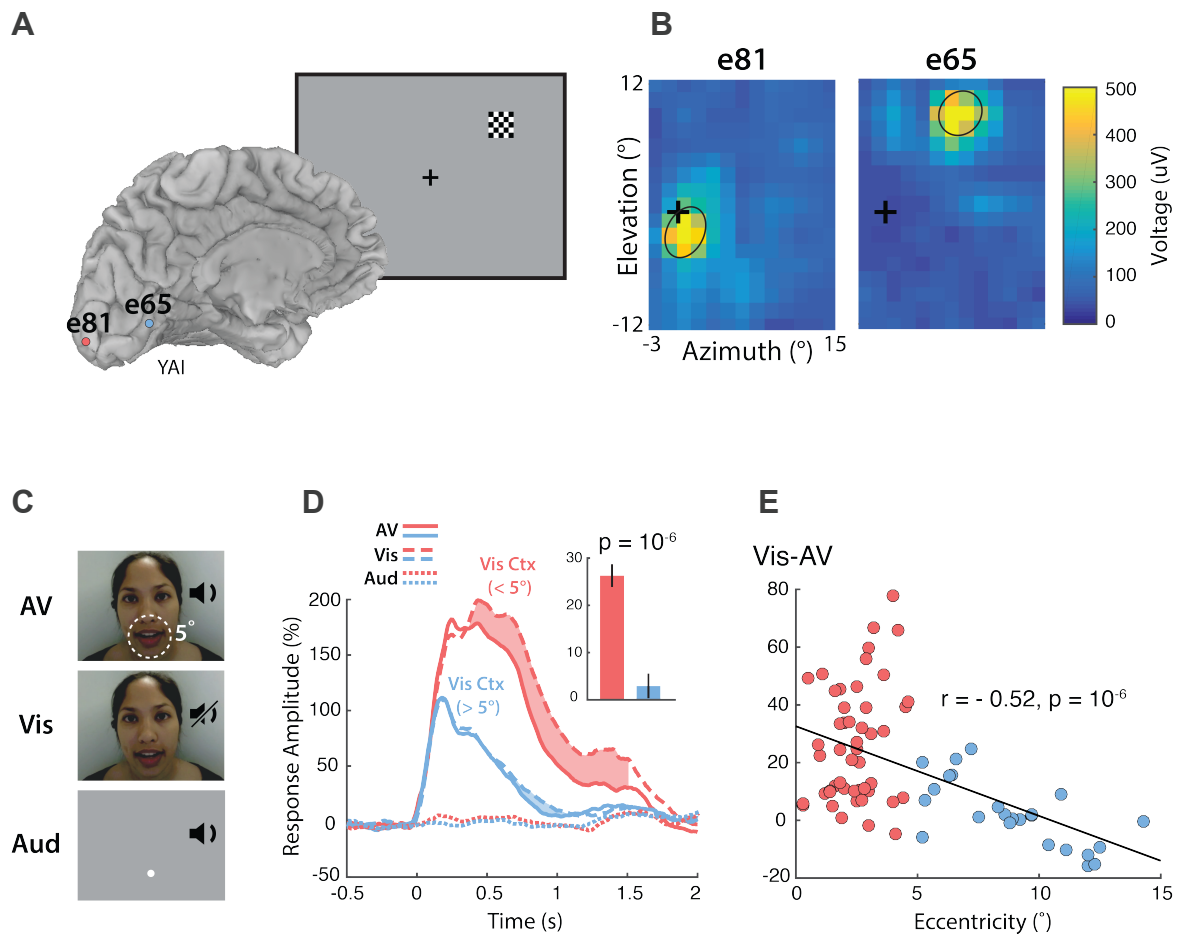


Figure 3.1 Retinotopic organization of speech responses in visual cortex

(A) (Left) Medial view of a cortical surface model of the left hemisphere brain of a single subject (anonymized subject ID YAI). Posterior electrode **e81** (red circle) was located superior to the calcarine sulcus on the occipital pole (red circle) while anterior electrode **e65** (blue circle) was located inferior to the calcarine on the medial wall of the hemisphere. (Right) The receptive field mapping stimulus consisted of a small checkerboard presented at random screen locations while subjects performed a letter detection task at fixation (not shown).

(B) The responses evoked by the mapping stimulus in electrodes **e81** (left panel) and **e65** (right panel). Color scales corresponds to the amplitude of the visual evoked response at each location in the visual field, with the crosshairs showing the center of the visual field and the black ellipse showing the half-maximum of a two-dimensional Gaussian fitted to the response. Electrode **e81** had a central receptive field (eccentricity at RF center of 2.5°) while electrode **e65** had a peripheral receptive field (eccentricity 10.9°).

(C) The speech stimuli consisted of audiovisual recordings of a female talker speaking words (*AV*) edited so that only the visual portion of the recording was presented (*Vis*) or only the auditory portion of the recording was presented (*Aud*). The mouth region of the talker's face subtended 5° (white dashed circle in top panel). Subjects were instructed to fixate the talker's mouth (*AV* and *Vis* conditions) or a fixation point presented at the same screen location as the talker's mouth (*Aud* condition).

(D) Broadband responses (70-150 Hz) to audiovisual (solid line), visual-only (dashed line) and auditory-only (dotted line) speech averaged across visual electrodes with central (red) and peripheral (blue) receptive field locations. Shaded regions show the periods throughout the speech stimuli when response to visual-only speech was greater than the response to audiovisual speech. Inset bar graph shows the average enhancement (200-1500 ms) for central (red bar) and peripheral (blue bar) visual electrodes. Error bars indicate standard error of the mean.

(E) Response enhancement (*Vis-AV*) in each visual electrode is shown with respect to the eccentricity of that electrode's receptive field (Central visual electrodes are shown with red circles, peripheral visual electrodes are shown with blue circles). Black line depicts the negative correlation between connectivity and eccentricity.

There was a main effect of electrode location, with significantly greater response in central than peripheral electrodes (central vs. peripheral: 79 ± 7 % vs. 22 ± 6 %; $p = 10^{-5}$) and a main effect of stimulus condition, with a significantly greater response for *Vis* speech compared with *AV* speech (*Vis* vs. *AV*: 95 ± 10 % vs. 77 ± 8 %; main effect of *V* speech $p = 10^{-6}$) and a significantly weaker response for *Aud* speech (*Aud* vs. *AV*: -3 ± 1 % vs. 77 ± 8 %; $p = 10^{-16}$).

Critically, there was a significant interaction between RF location and stimulus

condition. Central electrodes showed a large difference between the responses to *Vis* and *AV* speech while peripheral electrodes showed almost no difference (*Vis* – *AV*, central: 26 ± 3 % vs. peripheral: 2 ± 2 %; $t = 5.4$, $p = 10^{-6}$, unpaired t-test). We confirmed the relationship between eccentricity and enhanced responses to visual speech with an analysis in which RF location was treated as a continuous variable. There was a significant negative correlation between eccentricity and response enhancement (Figure 3.1E; Pearson's correlation: $r = -0.52$, $p = 10^{-6}$).

There was a large difference in the responses to *Vis* and *AV* speech in central visual electrodes, even though the bottom-up visual stimulus in the two conditions was identical, suggesting that that top-down influences might play a role. We investigated responses in frontal cortex as a possible source of these top-down signals. One frontal electrode was selected in each subject ($n = 8$; average Talairach co-ordinates: $x = 50$, $y = -8$, $z = 34$; range x : 35 to 50; y : -2 to 20; x : 10 to 40; locations of all electrodes shown in Figure 3.2A). As in visual cortex, frontal cortex showed the strongest responses to *Vis* speech (Figure 3.2B; *Vis* vs. *AV* = 53 ± 10 % vs. 29 ± 5 %; $p = 10^{-3}$, LME model main effect of *Vis* speech).

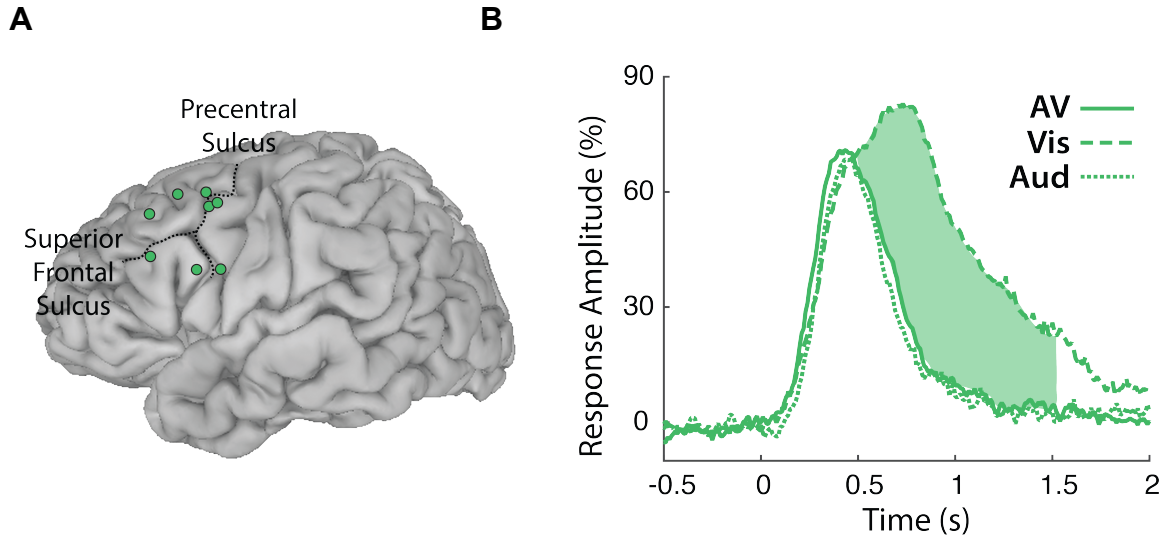


Figure 3.2 Speech responses in frontal cortex

(A) Cortical surface model showing the frontal electrodes (green circles), obtained by projecting electrodes from all subjects onto left hemisphere of a template brain.

(B) Broadband responses (70-150 Hz) to audiovisual (solid line), visual-only (dashed line) and auditory-only (dotted line) speech averaged across frontal electrodes. Shaded region show the periods throughout the speech stimuli when response to visual-only speech was greater than the response to audiovisual speech.

To determine whether the enhanced responses to *Vis* speech in visual and frontal cortex were related, we examined functional connectivity between pairs of electrodes with one member of the pair in frontal cortex and the other in visual cortex. To measure functional connectivity, the relationship between the trial-by-trial broadband power within each pair was assessed using Spearman's rank correlation (120, 121). Figure 3.3 shows the functional connectivity for a sample electrode pair with a frontal electrode located on the inferior portion of the precentral gyrus (Talairach co-ordinates: $x = 64$, $y = 0$, $z = 22$) and a visual electrode with a central receptive field (1.6° eccentricity). During presentation of

Vis speech (but not *AV* or *Aud* speech) functional connectivity was strong (*Vis*: $\rho = 0.52$, $p = 10^{-5}$; *AV*: $\rho = 0.24$, $p = 0.07$; *Aud*: $\rho = 0.17$, $p = 0.2$).

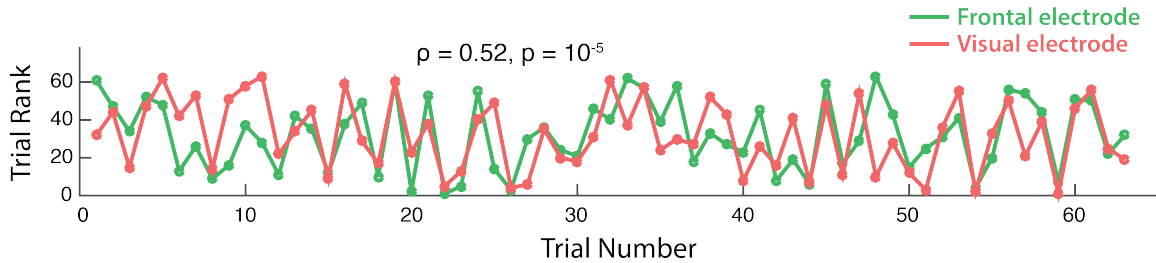


Figure 3.3 Trial-by-trial correlation for a single frontal-visual electrode pair

Average broadband responses for each visual-only speech trial (200-1500 ms, 70-150 Hz) measured simultaneously in a single frontal (green) and a single visual cortex (red) electrode. Trials are ranked based on their response amplitudes (y axis) and shown with respect to their presentation orders (x axis).

If frontal cortex was responsible for enhanced visual cortex responses to visual speech, we would expect high connectivity between frontal cortex and central visual electrodes (which showed a large difference between *Vis* and *AV* speech) and low connectivity between frontal cortex and peripheral visual electrodes (which showed little difference between *Vis* and *AV* speech). To quantitatively test this idea, we fit an LME model with the strength of the connection (ρ) between each frontal-visual electrode pair as the dependent measure; the receptive field location of the visual electrode (central or peripheral) and the stimulus condition (*AV*, *Vis* or *Aud* speech) as fixed factors; and connectivity with the central electrodes during *AV* speech as the baseline.

As predicted, there was a large main effect of RF location, with greater connectivity in central electrodes than in peripheral electrodes (Figure 3.4A; central vs. peripheral: 0.21 ± 0.02 vs. 0.01 ± 0.02 %; $p = 10^{-4}$) and this difference

was larger in the *Vis* speech condition than the *AV* or *Aud* conditions (central – peripheral ρ , *Vis*: 0.34; *AV*: 0.17; *Aud*: 0.07). Treating the RF location of each visual electrode as a continuous variable revealed a significant negative correlation between eccentricity and ρ for both *Vis* and *AV* speech (Figure 3.4B; *Vis*: $r = -0.7$, $p = 10^{-11}$; *AV*: $r = -0.36$, $p = 10^{-3}$) with a greater effect of eccentricity on connectivity during presentation of *Vis* vs. *AV* speech (-0.7 vs. -0.36 , $z = 2.9$, $p = 0.004$ by Fisher r -to- z).

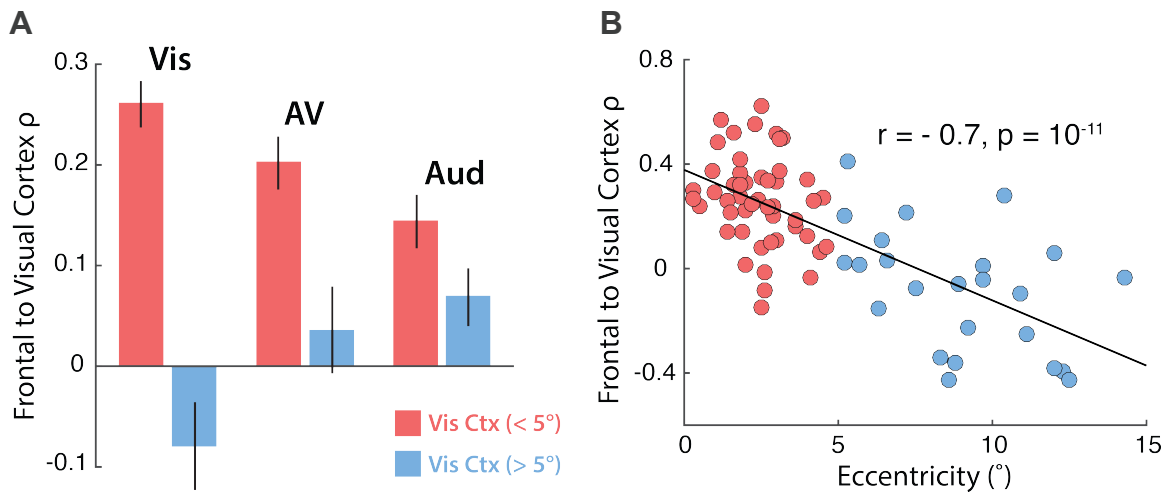


Figure 3.4 Functional connectivity between frontal and visual cortices

(A) Functional connectivity (calculated as Spearman Rank correlation ρ) for *AV*, *Vis* and *Aud* speech is averaged across all frontal-visual electrode pairs and shown separately for central (red) and peripheral (blue) visual electrodes. Error bars indicate the standard error of the mean.

(B) Functional connectivity between frontal-visual electrode pairs, measured as trial-by-trial power correlation across visual-only speech trials, is shown with respect to the eccentricity of the electrode's receptive field (Connectivity with central visual electrodes are shown with red circles, connectivity with peripheral visual electrodes are shown with blue circles). Black line depicts the negative correlation between connectivity and eccentricity.

One possible confound in the frontal-visual functional connectivity analysis is synchronous changes driven by bottom-up stimulus effects, with some

stimulus exemplars simply evoking stronger responses in both areas. To control for this possibility, in each electrode we subtracted the mean response to a stimulus from all trials in which that stimulus was presented. Even after removing stimulus effects in this way, the effects noted above remained; *i.e.* greater frontal connectivity for visual cortex electrodes with central vs. peripheral receptive fields (*Vis*: central 0.22 ± 0.03 vs. peripheral -0.07 ± 0.05 ; Unpaired t-test: $t_{71} = 5$, $p = 10^{-6}$; *AV*: central 0.19 ± 0.03 vs. peripheral 0.04 ± 0.05 ; Unpaired t-test: $t_{71} = 2.4$, $p = 0.02$) and a negative correlation between functional connectivity and eccentricity (*Vis*: $\rho = -0.56$, $p = 10^{-7}$; *AV*: $\rho = -0.24$, $p = 0.04$; significantly greater effect of eccentricity on connectivity for *Vis* vs. *AV* speech, -0.56 vs. -0.24 , $z = 2.3$, $p = 0.02$ by Fisher r-to-z).

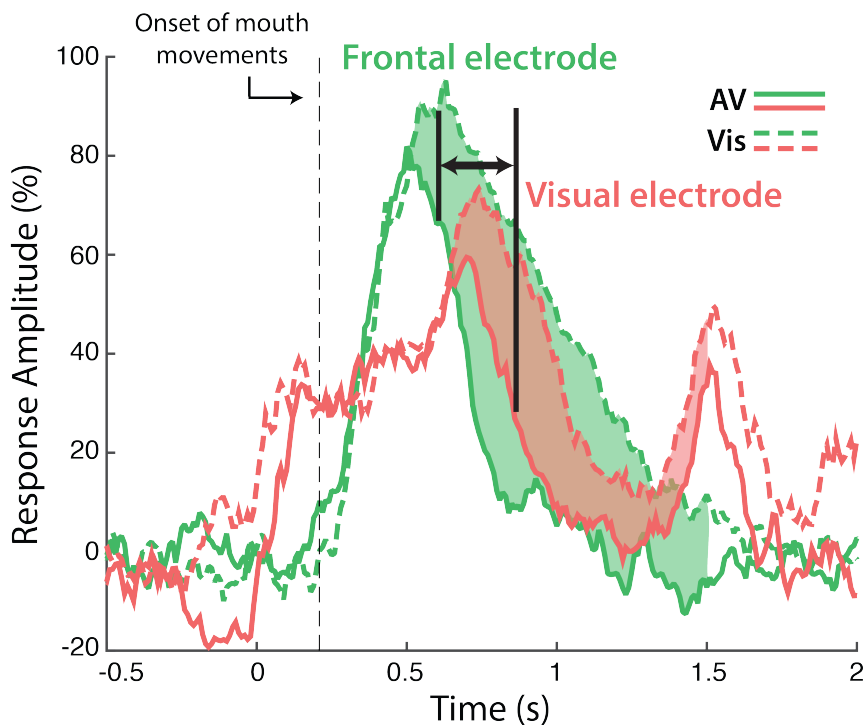


Figure 3.5 Latency of the response enhancement

Broadband responses (70-150 Hz) to audiovisual (solid line) and visual-only (dashed line) speech averaged across trials. Responses measured in a single frontal (green) and a single visual cortex (red) electrode. Dashed black line depicts the onset of mouth movements. Shaded regions show the periods throughout the speech stimuli when response to visual-only speech was greater than the response to audiovisual speech. Solid black lines indicate the time at which divergence between *Vis* and *AV* is significant for the frontal and visual electrode.

Our interpretation proposes that frontal cortex modulates central regions of visual cortex in a top-down fashion. If this is true, then one might expect frontal response differences to *precede* visual cortex response differences. To test this idea, we examined the time courses of the response to speech in a sample frontal-visual electrode pair. Both electrodes showed larger responses to *Vis* speech than *AV* speech, but the increased response for *Vis* speech occurred *earlier* in the frontal electrode, beginning at 400 ms after the onset of mouth movements, versus 660 ms for the visual electrode. This supports that idea that frontal cortex is the source of the modulation signal in visual cortex. The earlier divergence between *Vis* and *AV* speech in frontal compared with visual cortex was consistent across subjects, with an average latency difference of 174 ms (frontal vs. visual: 341 ± 125 ms vs. 515 ± 107 ms, $t_{14} = 3$, $p = 0.01$).

Discussion

Speech perception is one of the brain's most important tasks and relies on information from both the auditory and the visual modalities. The relative importance of these modalities changes, with visual speech having greater importance when auditory speech is degraded or absent. A possible neural substrate for this change is an enhanced response in visual cortex for visual

speech (36). In the current chapter, we examined whether there is a retionotopic bias for this response enhancement in the visual cortex. A previous study on eye movements during speech perception showed that mouth of the talker did not have to be at the center of the visual field to influence the perception of the McGurk illusion (122). Subjects still perceived the illusion when they fixated on the eye region of the talker or away from the talker's face. Another study showed that the McGurk illusion persisted even when the face of the talker was degraded by spatial blurring (123). These findings demonstrate that visual information supplied by the periphery or by coarse facial features can still influence speech perception. However another study showed that when subjects viewed audiovisual speech movies with a noisy auditory component they spent more time gazing at the mouth of the talker (62, 119). This suggests that when auditory speech is not informative, a natural strategy is to fixate on the mouth of the talker to extract information from visual speech. When subjects fixate on the mouth of the talker, we expected visual cortex representing the central visual field to receive the majority of visual speech information. Indeed, our results showed that only these central parts of visual cortex exhibited response enhancement to visual speech.

We also investigated the source of this response enhancement, predicting that it is caused by a top-down influence from higher order frontal regions. Because in the absence of a top-down influence, visual cortex would respond similarly to visual and audiovisual speech, since the visual stimulus is identical in the two conditions.

We suggested that frontal-visual attentional control circuits are automatically engaged during speech perception in the service of increasing perceptual accuracy for the processing of this very important class of stimuli. This allows for precise, time-varying control: as the quality of auditory information fluctuates, as auditory noise in the environment increases or decreases, frontal cortex can up or down-regulate activity in visual cortex accordingly. It also allows for precise spatial control: as the mouth of the talker contains the most speech information, frontal cortex can selectively enhance visual cortex activity that is relevant for speech perception by enhancing activity in subregions of visual cortex that represent the mouth.

Our results showed that the frontal cortex exhibits the same response pattern as the visual cortex, responding more to visual speech than audiovisual speech. Responses to visual speech in frontal cortex have also been demonstrated in various fMRI studies. These studies showed that frontal regions had larger BOLD responses during speech reading when contrasted with responses to baseline conditions such as fixation, still faces or gurning faces (37, 103, 114). In complete agreement with our results, a more recent study demonstrated that responses to visual speech in these frontal regions were larger even when contrasted with responses to audiovisual speech (47).

We demonstrated significant functional connectivity between frontal and visual cortices for all speech conditions, supporting that the coinciding response enhancement in both regions is not independent but rather a result of an interaction between the two regions. Supporting our finding, previous studies

provided plenty of evidence for an interaction between the two regions.

Anatomical studies in monkeys revealed that there are cortical connections between frontal and visual cortices that consist of both bottom-up and top-down projections (124, 125). In humans, a possible anatomical linkage supporting this processing is the inferior fronto-occipital fasciculus connecting frontal and occipital regions (126). Furthermore, neuroimaging studies in humans showed that prefrontal cortex modulates visual cortex in a top-down manner during goal-directed visual memory and visual attention tasks (116, 127). In a lesion study, patients with prefrontal cortex lesions had weaker visual responses in the ipsilesional hemisphere during a visual discrimination task compared to healthy control subjects. Also their detection rate was lower when stimuli were presented at the contralesional visual field (60). More direct evidence on frontal modulation of visual cortex came from a study, which showed that stimulating the prefrontal cortex with TMS around the frontal eye fields alters BOLD responses in the visual cortex as well as the perceived contrast of the presented visual stimuli (118, 128).

We showed that the retinotopic bias observed for response enhancement in the visual cortex was also present for functional connectivity with the frontal cortex, such that electrodes with central receptive field locations had stronger functional connectivity with the frontal cortex than electrodes with peripheral receptive field locations. This suggests that the top-down signal from the frontal cortex may be related to attention to the relevant information, explaining why response enhancement is greater at the location of the mouth rather than

anywhere else in the visual field. Previous studies demonstrated that visual attention not only operates through facilitation of visual responses at the location of the stimulus but also through inhibition of the surround (129-131). While larger responses at the location of the mouth can be explained by attentional facilitation, smaller responses observed in visual regions representing the peripheral visual field can be related to the attentional inhibition of the surround.

Although we have established the functional interaction between the frontal and visual cortices during speech processing, our functional connectivity analysis was based on trial-by-trial power correlations between the two regions, which provides no information on the direction of interaction. However we interpreted the relative timing of the responses in the two regions as evidence on the direction of the interaction. Specifically, response enhancement occurred earlier in the frontal cortex than in visual cortex, suggesting that enhancement in visual cortex can be attributed to top-down influences from frontal cortex.

These results link two distinct strands of research: visual speech processing and attention. First, although previous studies of speech perception frequently observe activity in frontal cortex during perception of a visual-only speech (37, 47, 103, 113, 114), the precise role of this frontal activity has been difficult to determine. Second, it is well known that frontal regions in an around the frontal eye fields modulate visual cortex activity during tasks that require voluntary control of spatial or featural attention (61, 116, 132, 133), however it has not been clear how these attentional networks function during other important cognitive tasks, such as speech perception.

Our results support a model in which attentional control regions of frontal cortex selectively modulate visual cortex, amplifying activity with both spatial and context selectivity to enhance speech intelligibility. Most models of speech perception focus on auditory cortex inputs into parietal and frontal cortex (42, 43). Our findings suggest that visual cortex should also be considered an important component of the speech perception network, as it is selectively and rapidly modulated during audiovisual speech perception.

CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS

Seeing mouth movements render speech more intelligible when auditory speech is noisy or inaudible, however the neural substrates that underlie this perceptual benefit are not completely understood. My thesis explores how visual information from mouth movements is processed in the human brain to improve speech perception.

The *first* part of my thesis (Chapter 2) aims to understand how noisy speech is processed in the auditory cortex, specifically in the superior temporal gyrus (STG). In human electrophysiology studies, the most common way to examine the effects of an experimental condition on the neural response is to measure response amplitudes. Examining other measures, such as response variance, requires robust responses for each trial, therefore have only been tested with single-cell electrophysiology experiments in non-human primates. In my analyses, taking advantage of the high SNR of ECoG measurements, I examined both the amplitude and the variability of the neural responses.

My results demonstrated a double dissociation in responses to speech with clear and noisy auditory component within the STG. Noisy speech caused a decrease in response amplitude and an increase in response variability in anterior STG, but not in posterior STG. To interpret the computational roles of the two regions, I considered the Bayesian model of multisensory integration, which suggests that noisy sensory information leads to high variability in the neural response, however combining sensory information from different modalities should reduce this variability (85). There has been no direct evidence from EEG or fMRI studies to confirm these predictions of the Bayesian model in neural

population level in the human brain. Supporting the Bayesian model, my results confirmed these phenomena and demonstrated that posterior STG was more resistant to noisy speech due to its multisensory characteristics.

The Bayesian model also suggests that when two sensory cues are integrated, because a noisy sensory cue is less reliable, it should have less weight in the integration (31, 33). In support of this prediction, a previous study by Nath and Beauchamp showed that functional connectivity of the STS with auditory and visual cortices depends on the reliability of the auditory and visual modalities. In other words, STS had stronger connectivity with the visual cortex when auditory speech was noisy, while it had stronger connectivity with the auditory cortex when visual speech was blurry (41). In future work, it will be important to determine the actual weights of auditory and visual modalities in the audiovisual integration of speech. By mathematical modeling of audiovisual integration, it will be possible to predict the neural response based on the physical features of the speech stimulus.

Another important finding was the distinct border between the two response patterns, which was demarcated by a landmark corresponding to the posterior margin of Heschl's gyrus. Preference for clear versus noisy speech changed sharply rather than gradually across this border, suggesting a strong functional specialization for posterior STG. However, because we recorded neural activity on the lateral cortical surface, we could only examine responses in the anterior-posterior direction along the STG. The area posterior to the Heschl's gyrus that is buried within the lateral sulcus is called the planum temporale and

constitutes an important portion of the Wernicke's area (134). In future studies, it would be important to record neural activity from planum temporale using penetrating depth electrodes and examine responses along the lateral-medial direction in order to fully characterize this region.

The *second* part of my thesis (Chapter 3) focuses on speech processing in the visual cortex to understand how responses in the visual cortex are modulated when visual speech is the only source of information. A recent ECoG study by Schepers and colleagues demonstrated that responses in the visual cortex were greater for visual speech (silent speech) than for audiovisual speech (36). This was a new finding because previous fMRI studies had not reported any such response differences, possibly due to the slow temporal resolution of fMRI (34, 41, 103, 135, 136), and it raised two important questions. The first question was the source of the response enhancement in the visual cortex. I predicted that there was a top-down influence on visual cortex to amplify responses when visual information is critical for speech perception, because otherwise one would expect the visual cortex to be insensitive to the auditory content of speech.

The second question was whether the amount of response enhancement was uniform throughout the visual cortex. Schepers and colleagues examined the responses in different subregions of the visual cortex, however they did not observe any differences in the amount of response enhancement. However, visual cortex is organized retinotopically, meaning that different locations in the visual field are represented at different parts of the visual cortex (137). I predicted that the amount of response enhancement should vary in different retinotopical

regions of the visual cortex. Since the mouth carries the majority of visual speech information, I expected that as we watch someone talk, the parts of visual cortex that have receptive field locations around the mouth of the talker should show greater response enhancement.

Using the same data set as Schepers and colleagues, I first showed that in addition to visual cortex, frontal regions, specifically the inferior frontal gyrus, premotor cortex including the frontal eye fields and dorsolateral regions of the prefrontal cortex, also showed greater response to visual speech than audiovisual speech, confirming the role of these regions in speech-reading as demonstrated by previous studies (37, 47, 103, 114). Moreover, the onset of the response enhancement in these frontal regions preceded the response enhancement in visual cortex, suggesting a top-down influence by frontal cortex on visual cortex. Next, I analyzed the receptive field mapping data collected from the same group of subjects in a separate experiment and discovered that the response enhancement observed in visual cortex was indeed retinotopically specific, with only central regions of visual cortex that represent the mouth region of the talker showing enhancement. Finally I demonstrated that these central regions of visual cortex had strong functional connectivity with frontal regions, not the peripheral regions.

Taken together, these results support a model in which frontal cortex selectively modulates visual cortex, enhancing activity with both spatial selectivity (only mouth regions of the face are enhanced) and context selectivity (enhancement is greater when visual speech is more important). An important

question that remains to be determined is the role of parietal cortex in the modulation of visual responses during speech perception. Similar to frontal regions, parietal regions including temporoparietal junction, inferior parietal lobule and intraparietal sulcus have also been shown to activate for both visual and audiovisual speech (34, 40, 138, 139). Likewise, these regions have also been implicated in the top-down modulation of visual cortex during visual attention tasks (133, 140). For future research, it would be interesting to examine the functional connectivity between frontal, parietal and visual cortices to delineate the interactions within this circuitry during speech perception.

Another exciting future direction would be to link neural activation to perception, for example to predict when an individual will perceive a noisy or silent speech stimulus correctly versus incorrectly based on the neural activity. Because speech perception is a complex task that requires the involvement of a network of brain regions, a complete understanding of the speech network will be crucial to achieve this mission. It will be necessary to use multimodal techniques that examine the whole speech network at once rather than focusing on one of its components at a time.

The findings of my dissertation not only contribute to basic neuroscience by clarifying how the brain uses visual speech information to compensate for noisy or inaudible auditory speech, but also have clinical significance especially pertaining to individuals with hearing loss. Hearing loss is a common cause of impaired speech perception and is the 3rd most prevalent health problem in the United States affecting 48 million Americans (141). A better understanding of the

neural processes that support speech perception will help patients with hearing loss by guiding therapeutic interventions ranging from speech therapy to assistive devices that will improve speech perception.

To name but a few, understanding the neural substrates of speech perception would help optimizing behavioral intervention techniques, such as computerized training paradigms used for patients with hearing aids and cochlear implants (142). It would guide alternative treatment methods such as speech therapy coupled with noninvasive stimulation of brain regions that are important for speech processing (e.g. with transcranial direct current stimulation). Likewise, it would help to improve neurofeedback measurements of brain activity, such that patients may be able to increase their use of the visual speech information by self-regulating activity in brain regions responsible for visual speech processing. Also it would be critical for the development of “speech perception” neural prosthetics that could use computerized voice recognition to decode speech and then stimulate brain areas that no longer receive normal sensory input with an artificial pattern of activity that mirrors the pattern evoked in a normal-hearing individual.

Today the most advanced computer technologies that implement speech recognition (e.g. Siri, Google Assistant, Alexa or Cortana) are nowhere near as capable as the human brain. Speech is a cognitive function unique to humans that distinguishes us from our closest animal relatives and allows us to exchange ideas and convey emotions. We are only beginning to understand our brain’s

amazing ability to communicate with others through speech, which is fundamental to our identity as human beings.

REFERENCES

1. Sumby, W. H., and I. Pollack. 1954. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america* 26: 212-215.
2. Bernstein, L. E., E. T. Auer, and S. Takayanagi. 2004. Auditory speech detection in noise enhanced by lipreading. *Speech Communication* 44: 5-18.
3. Ross, L. A., D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe. 2007. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral cortex (New York, N.Y. : 1991)* 17: 1147-1153.
4. Summerfield, Q. 1987. Some preliminaries to a comprehensive account of audio-visual speech perception.
5. Tye-Murray, N., M. S. Sommers, and B. Spehar. 2007. Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and hearing* 28: 656-668.
6. Suh, M. W., H. J. Lee, J. S. Kim, C. K. Chung, and S. H. Oh. 2009. Speech experience shapes the speechreading network and subsequent deafness facilitates it. *Brain* 132: 2761-2771.
7. Rouger, J., S. Lagleyre, B. Fraysse, S. Deneve, O. Deguine, and P. Barone. 2007. Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences* 104: 7295-7300.

8. McGurk, H., and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264: 746-748.
9. Munhall, K. G., P. Gribble, L. Sacco, and M. Ward. 1996. Temporal constraints on the McGurk effect. *Perception & Psychophysics* 58: 351-362.
10. Vroomen, J., and B. De Gelder. 2004. Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon. *The handbook of multisensory processes* 3: 1-23.
11. Wessinger, C. M., J. VanMeter, B. Tian, J. Van Lare, J. Pekar, and J. P. Rauschecker. 2001. Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *Journal of cognitive neuroscience* 13: 1-7.
12. Kaas, J. H., and T. A. Hackett. 2000. Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the United States of America* 97: 11793-11799.
13. Bogen, J. E., and G. Bogen. 1976. Wernicke's region—where is it? *Annals of the New York Academy of Sciences* 280: 834-843.
14. Geschwind, N. 1970. The organization of language and the brain. In *Science*. Citeseer.
15. Penfield, W., and L. Roberts. 1959. Speech and brain. Princeton, NJ: Princeton University Press.
16. Wernicke, C. 1874. *Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis*. Cohn.

17. Seltzer, B., and D. N. Pandya. 1994. Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. *The Journal of comparative neurology* 343: 445-463.
18. Lewis, J. W., and D. C. Van Essen. 2000. Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *The Journal of comparative neurology* 428: 112-137.
19. Beauchamp, M. S., K. E. Lee, B. D. Argall, and A. Martin. 2004. Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron* 41: 809-823.
20. Barraclough, N. E., D. Xiao, C. I. Baker, M. W. Oram, and D. I. Perrett. 2005. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of cognitive neuroscience* 17: 377-391.
21. Calvert, G. A., P. C. Hansen, S. D. Iversen, and M. J. Brammer. 2001. Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *NeuroImage* 14: 427-438.
22. Bruce, C., R. Desimone, and C. G. Gross. 1981. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of neurophysiology* 46: 369-384.

23. Gil-da-Costa, R., A. Martin, M. A. Lopes, M. Munoz, J. B. Fritz, and A. R. Braun. 2006. Species-specific calls activate homologs of Broca's and Wernicke's areas in the macaque. *Nature neuroscience* 9: 1064-1070.
24. Belin, P., R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike. 2000. Voice-selective areas in human auditory cortex. *Nature* 403: 309-312.
25. Puce, A., T. Allison, S. Bentin, J. C. Gore, and G. McCarthy. 1998. Temporal cortex activation in humans viewing eye and mouth movements. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 18: 2188-2199.
26. Reale, R. A., G. A. Calvert, T. Thesen, R. L. Jenison, H. Kawasaki, H. Oya, M. A. Howard, and J. F. Brugge. 2007. Auditory-visual processing represented in the human superior temporal gyrus. *Neuroscience* 145: 162-184.
27. Calvert, G. A., M. J. Brammer, E. T. Bullmore, R. Campbell, S. D. Iversen, and A. S. David. 1999. Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport* 10: 2619-2623.
28. Beauchamp, M. S. 2005. Statistical criteria in fMRI studies of multisensory integration. *Neuroinformatics* 3: 93-113.
29. Stevenson, R. A., M. Bushmakin, S. Kim, M. T. Wallace, A. Puce, and T. W. James. 2012. Inverse effectiveness and multisensory interactions in visual event-related potentials with audiovisual speech. *Brain topography* 25: 308-326.

30. Callan, D. E., J. A. Jones, K. Munhall, A. M. Callan, C. Kroos, and E. Vatikiotis-Bateson. 2003. Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport* 14: 2213-2218.
31. Fetsch, C. R., A. Pouget, G. C. DeAngelis, and D. E. Angelaki. 2012. Neural correlates of reliability-based cue weighting during multisensory integration. *Nature neuroscience* 15: 146-154.
32. Ma, W. J., J. M. Beck, P. E. Latham, and A. Pouget. 2006. Bayesian inference with probabilistic population codes. *Nature neuroscience* 9: 1432-1438.
33. Knill, D. C., and A. Pouget. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in neurosciences* 27: 712-719.
34. Miller, L. M., and M. D'Esposito. 2005. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 25: 5884-5893.
35. Beauchamp, M. S., A. R. Nath, and S. Pasalar. 2010. fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 30: 2414-2417.
36. Schepers, I. M., D. Yoshor, and M. S. Beauchamp. 2015. Electrocorticography Reveals Enhanced Visual Cortex Responses to Visual Speech. *Cerebral cortex (New York, N.Y. : 1991)* 25: 4103-4110.

37. Hall, D. A., C. Fussell, and A. Q. Summerfield. 2005. Reading fluent speech from talking faces: typical brain networks and individual differences. *Journal of cognitive neuroscience* 17: 939-953.
38. Sekiyama, K., I. Kanno, S. Miura, and Y. Sugita. 2003. Auditory-visual speech perception examined by fMRI and PET. *Neuroscience research* 47: 277-287.
39. Skipper, J. I., H. C. Nusbaum, and S. L. Small. 2005. Listening to talking faces: motor cortical activation during speech perception. *NeuroImage* 25: 76-89.
40. Nishitani, N., and R. Hari. 2002. Viewing lip forms: cortical dynamics. *Neuron* 36: 1211-1220.
41. Nath, A. R., and M. S. Beauchamp. 2011. Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 31: 1704-1714.
42. Rauschecker, J. P., and S. K. Scott. 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience* 12: 718-724.
43. Hickok, G., and D. Poeppel. 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92: 67-99.

44. Keller, S. S., T. Crow, A. Foundas, K. Amunts, and N. Roberts. 2009. Broca's area: nomenclature, anatomy, typology and asymmetry. *Brain and language* 109: 29-48.
45. Broca, P. 1861. Perte de la parole, ramollissement chronique et destruction partielle du lobe antérieur gauche du cerveau. *Bull Soc Anthropol* 2: 235-238.
46. Ojanen, V., R. Mottonen, J. Pekkola, I. P. Jaaskelainen, R. Joensuu, T. Autti, and M. Sams. 2005. Processing of audiovisual speech in Broca's area. *NeuroImage* 25: 333-338.
47. Callan, D. E., J. A. Jones, and A. Callan. 2014. Multisensory and modality specific processing of visual speech in different regions of the premotor cortex. *Frontiers in psychology* 5: 389.
48. Skipper, J. I., S. Goldin-Meadow, H. C. Nusbaum, and S. L. Small. 2007. Speech-associated gestures, Broca's area, and the human mirror system. *Brain and language* 101: 260-277.
49. Flinker, A., A. Korzeniewska, A. Y. Shestyuk, P. J. Franaszczuk, N. F. Dronkers, R. T. Knight, and N. E. Crone. 2015. Redefining the role of Broca's area in speech. *Proceedings of the National Academy of Sciences of the United States of America* 112: 2871-2875.
50. Friston, K. J. 2009. Modalities, modes, and models in functional neuroimaging. *Science* 326: 399-403.

51. Miller, J. L., F. Grosjean, and C. Lomanto. 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica* 41: 215-225.
52. Ritaccio, A., D. Boatman-Reich, P. Brunner, M. C. Cervenka, A. J. Cole, N. Crone, R. Duckrow, A. Korzeniewska, B. Litt, K. J. Miller, D. W. Moran, J. Parvizi, J. Viventi, J. Williams, and G. Schalk. 2011. Proceedings of the Second International Workshop on Advances in Electrocorticography. *Epilepsy & behavior : E&B* 22: 641-650.
53. Buzsaki, G., C. A. Anastassiou, and C. Koch. 2012. The origin of extracellular fields and currents--EEG, ECoG, LFP and spikes. *Nature reviews. Neuroscience* 13: 407-420.
54. Mukamel, R., H. Gelbard, A. Arieli, U. Hasson, I. Fried, and R. Malach. 2005. Coupling between neuronal firing, field potentials, and fMRI in human auditory cortex. *Science* 309: 951-954.
55. Nir, Y., L. Fisch, R. Mukamel, H. Gelbard-Sagiv, A. Arieli, I. Fried, and R. Malach. 2007. Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. *Current biology : CB* 17: 1275-1285.
56. Crone, N. E., D. Boatman, B. Gordon, and L. Hao. 2001. Induced electrocorticographic gamma activity during auditory perception. Brazier Award-winning article, 2001. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* 112: 565-582.

57. Edwards, E., M. Soltani, L. Y. Deouell, M. S. Berger, and R. T. Knight. 2005. High gamma activity in response to deviant auditory stimuli recorded directly from human cortex. *Journal of neurophysiology* 94: 4269-4280.
58. Steinschneider, M., K. V. Nourski, H. Kawasaki, H. Oya, J. F. Brugge, and M. A. Howard, 3rd. 2011. Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cerebral cortex (New York, N.Y. : 1991)* 21: 2332-2347.
59. Mesgarani, N., C. Cheung, K. Johnson, and E. F. Chang. 2014. Phonetic feature encoding in human superior temporal gyrus. *Science* 343: 1006-1010.
60. Barcelo, F., S. Suwazono, and R. T. Knight. 2000. Prefrontal modulation of visual processing in humans. *Nature neuroscience* 3: 399-403.
61. Kastner, S., and L. G. Ungerleider. 2000. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience* 23: 315-341.
62. Vatikiotis-Bateson, E., I.-M. Eigsti, S. Yano, and K. G. Munhall. 1998. Eye movement of perceivers during audiovisualspeech perception. *Attention, Perception, & Psychophysics* 60: 926-940.
63. Yoshor, D., W. H. Bosking, G. M. Ghose, and J. H. Maunsell. 2007. Receptive fields in human visual cortex mapped with surface electrodes. *Cerebral cortex (New York, N.Y. : 1991)* 17: 2293-2302.
64. Binder, J. R., J. A. Frost, T. A. Hammeke, P. S. Bellgowan, J. A. Springer, J. N. Kaufman, and E. T. Possing. 2000. Human temporal lobe activation

- by speech and nonspeech sounds. *Cerebral cortex* (New York, N.Y. : 1991) 10: 512-528.
65. Rauschecker, J. P. 2015. Auditory and visual cortex of primates: a comparison of two sensory systems. *The European journal of neuroscience* 41: 579-585.
 66. van Atteveldt, N., E. Formisano, R. Goebel, and L. Blomert. 2004. Integration of letters and speech sounds in the human brain. *Neuron* 43: 271-282.
 67. Calvert, G. A., R. Campbell, and M. J. Brammer. 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current biology : CB* 10: 649-657.
 68. Foxe, J. J., G. R. Wylie, A. Martinez, C. E. Schroeder, D. C. Javitt, D. Guilfoyle, W. Ritter, and M. M. Murray. 2002. Auditory-somatosensory multisensory processing in auditory association cortex: an fMRI study. *Journal of neurophysiology* 88: 540-543.
 69. Sheffert, S., L. Lachs, and L. Hernandez. 1996. Research on spoken language processing progress report no. 21. *Bloomington, IN: Speech Research Laboratory, Indiana University*: 578-583.
 70. Schepers, I. M., T. R. Schneider, J. F. Hipp, A. K. Engel, and D. Senkowski. 2013. Noise alters beta-band activity in superior temporal cortex during audiovisual speech processing. *NeuroImage* 70: 101-112.

71. Dale, A. M., B. Fischl, and M. I. Sereno. 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage* 9: 179-194.
72. Fischl, B., M. I. Sereno, and A. M. Dale. 1999. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage* 9: 195-207.
73. Argall, B. D., Z. S. Saad, and M. S. Beauchamp. 2006. Simplified intersubject averaging on the cortical surface using SUMA. *Human brain mapping* 27: 14-27.
74. Cox, R. W. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and biomedical research, an international journal* 29: 162-173.
75. Oostenveld, R., P. Fries, E. Maris, and J. M. Schoffelen. 2011. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience* 2011: 156869.
76. Ray, S., and J. H. Maunsell. 2011. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS biology* 9: e1000610.
77. Jacques, C., N. Witthoft, K. S. Weiner, B. L. Foster, V. Rangarajan, D. Hermes, K. J. Miller, J. Parvizi, and K. Grill-Spector. 2016. Corresponding ECoG and fMRI category-selective signals in human ventral temporal cortex. *Neuropsychologia* 83: 14-28.

78. Tolhurst, D. J., J. A. Movshon, and A. F. Dean. 1983. The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision research* 23: 775-785.
79. Churchland, M. M., B. M. Yu, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, G. S. Corrado, W. T. Newsome, A. M. Clark, P. Hosseini, B. B. Scott, D. C. Bradley, M. A. Smith, A. Kohn, J. A. Movshon, K. M. Armstrong, T. Moore, S. W. Chang, L. H. Snyder, S. G. Lisberger, N. J. Priebe, I. M. Finn, D. Ferster, S. I. Ryu, G. Santhanam, M. Sahani, and K. V. Shenoy. 2010. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience* 13: 369-378.
80. Gur, M., A. Beylin, and D. M. Snodderly. 1997. Response variability of neurons in primary visual cortex (V1) of alert monkeys. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 17: 2914-2920.
81. Destrieux, C., B. Fischl, A. Dale, and E. Halgren. 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* 53: 1-15.
82. Desikan, R. S., F. Segonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, and R. J. Killiany. 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 31: 968-980.

83. Bates, D., M. Mächler, B. Bolker, and S. Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
84. Magnotti, J. F., and M. S. Beauchamp. 2017. A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech. *PLoS computational biology* 13: e1005229.
85. Ma, W. J., X. Zhou, L. A. Ross, J. J. Foxe, and L. C. Parra. 2009. Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PLoS One* 4: e4638.
86. Kuznetsova, A., P. B. Brockhoff, and R. H. B. Christensen. 2015. Package 'lmerTest'. *R package version: 2.0-29*.
87. Ernst, M. O., and M. S. Banks. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415: 429-433.
88. Alais, D., and D. Burr. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Current biology* 14: 257-262.
89. Bejjanki, V. R., M. Clayards, D. C. Knill, and R. N. Aslin. 2011. Cue integration in categorical tasks: insights from audio-visual speech perception. *PLoS One* 6: e19812.
90. Bernstein, L. E., and E. Liebenthal. 2014. Neural pathways for visual speech perception. *Frontiers in neuroscience* 8: 386.
91. Angelaki, D. E., Y. Gu, and G. C. DeAngelis. 2009. Multisensory integration: psychophysics, neurophysiology, and computation. *Current opinion in neurobiology* 19: 452-458.

92. Kastner, S., M. A. Pinsk, P. De Weerd, R. Desimone, and L. G. Ungerleider. 1999. Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron* 22: 751-761.
93. Maunsell, J. H. R., and S. Treue. 2006. Feature-based attention in visual cortex. *Trends in neurosciences* 29: 317-322.
94. Morosan, P., A. Schleicher, K. Amunts, and K. Zilles. 2005. Multimodal architectonic mapping of human superior temporal gyrus. *Anat Embryol (Berl)* 210: 401-406.
95. Hickok, G., and D. Poeppel. 2000. Towards a functional neuroanatomy of speech perception. *Trends in cognitive sciences* 4: 131-138.
96. Specht, K., and J. Reul. 2003. Functional segregation of the temporal lobes into highly differentiated subsystems for auditory perception: an auditory rapid event-related fMRI-task. *NeuroImage* 20: 1944-1954.
97. Okada, K., F. Rong, J. Venezia, W. Matchin, I. H. Hsieh, K. Saberi, J. T. Serences, and G. Hickok. 2010. Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. *Cerebral cortex (New York, N.Y. : 1991)* 20: 2486-2495.
98. Pickles, J. O. 2015. Auditory pathways: anatomy and physiology. *Handbook of clinical neurology* 129: 3-25.

99. Mishkin, M., and L. G. Ungerleider. 1982. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural brain research* 6: 57-77.
100. Leonard, M. K., and E. F. Chang. 2014. Dynamic speech representations in the human temporal lobe. *Trends in cognitive sciences* 18: 472-479.
101. Hullett, P. W., L. S. Hamilton, N. Mesgarani, C. E. Schreiner, and E. F. Chang. 2016. Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 36: 2014-2026.
102. Stevenson, R. A., and T. W. James. 2009. Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *NeuroImage* 44: 1210-1223.
103. Lee, H., and U. Noppeney. 2011. Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension. *The Journal of Neuroscience* 31: 11338-11350.
104. Bishop, C. W., and L. M. Miller. 2009. A multisensory cortical network for understanding speech in noise. *Journal of cognitive neuroscience* 21: 1790-1805.
105. McGettigan, C., A. Faulkner, I. Altarelli, J. Obleser, H. Baverstock, and S. K. Scott. 2012. Speech comprehension aided by multiple modalities: behavioural and neural interactions. *Neuropsychologia* 50: 762-776.

106. Erickson, L. C., B. A. Zielinski, J. E. Zielinski, G. Liu, P. E. Turkeltaub, A. M. Leaver, and J. P. Rauschecker. 2014. Distinct cortical locations for integration of audiovisual speech and the McGurk effect. *Frontiers in psychology* 5: 534.
107. Nath, A. R., and M. S. Beauchamp. 2012. A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage* 59: 781-787.
108. Obleser, J., J. Zimmermann, J. Van Meter, and J. P. Rauschecker. 2007. Multiple stages of auditory speech perception reflected in event-related FMRI. *Cerebral cortex (New York, N.Y. : 1991)* 17: 2251-2257.
109. Du, Y., B. R. Buchsbaum, C. L. Grady, and C. Alain. 2014. Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proceedings of the National Academy of Sciences of the United States of America* 111: 7126-7131.
110. Wild, C. J., M. H. Davis, and I. S. Johnsrude. 2012. Human auditory cortex is sensitive to the perceived clarity of speech. *NeuroImage* 60: 1490-1502.
111. Scott, S. K., C. C. Blank, S. Rosen, and R. J. Wise. 2000. Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123: 2400-2406.
112. Giraud, A. L., C. Kell, C. Thierfelder, P. Sterzer, M. O. Russ, C. Preibisch, and A. Kleinschmidt. 2004. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cerebral cortex (New York, N.Y. : 1991)* 14: 247-255.

113. Okada, K., and G. Hickok. 2009. Two cortical mechanisms support the integration of visual and auditory speech: a hypothesis and preliminary data. *Neuroscience letters* 452: 219-223.
114. Calvert, G. A., and R. Campbell. 2003. Reading speech from still and moving faces: the neural substrates of visible speech. *Journal of cognitive neuroscience* 15: 57-70.
115. Vossel, S., J. J. Geng, and G. R. Fink. 2014. Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry* 20: 150-159.
116. Corbetta, M., and G. L. Shulman. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews. Neuroscience* 3: 201-215.
117. Vossel, S., R. Weidner, J. Driver, K. J. Friston, and G. R. Fink. 2012. Deconstructing the architecture of dorsal and ventral attention systems with dynamic causal modeling. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 32: 10637-10648.
118. Driver, J., F. Blankenburg, S. Bestmann, and C. C. Ruff. 2010. New approaches to the study of human brain networks underlying spatial attention and related processes. *Experimental brain research* 206: 153-162.

119. Yi, A., W. Wong, and M. Eizenman. 2013. Gaze patterns and audiovisual speech enhancement. *Journal of speech, language, and hearing research : JSLHR* 56: 471-480.
120. Foster, B. L., V. Rangarajan, W. R. Shirer, and J. Parvizi. 2015. Intrinsic and task-dependent coupling of neuronal population activity in human parietal cortex. *Neuron* 86: 578-590.
121. Hipp, J. F., D. J. Hawellek, M. Corbetta, M. Siegel, and A. K. Engel. 2012. Large-scale cortical correlation structure of spontaneous oscillatory activity. *Nature neuroscience* 15: 884-890.
122. Pare, M., R. C. Richler, M. ten Hove, and K. G. Munhall. 2003. Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect. *Percept Psychophys* 65: 553-567.
123. MacDonald, J., S. Andersen, and T. Bachmann. 2000. Hearing by eye: how much spatial degradation can be tolerated? *Perception* 29: 1155-1168.
124. Yeterian, E. H., D. N. Pandya, F. Tomaiuolo, and M. Petrides. 2012. The cortical connectivity of the prefrontal cortex in the monkey brain. *Cortex; a journal devoted to the study of the nervous system and behavior* 48: 58-81.
125. Miller, E. K., and J. D. Cohen. 2001. An integrative theory of prefrontal cortex function. *Annual review of neuroscience* 24: 167-202.
126. Catani, M., and M. Mesulam. 2008. The arcuate fasciculus and the disconnection theme in language and aphasia: history and current state.

- Cortex; a journal devoted to the study of the nervous system and behavior*
44: 953-961.
127. Gazzaley, A., J. Rissman, J. Cooney, A. Rutman, T. Seibert, W. Clapp, and M. D'Esposito. 2007. Functional interactions between prefrontal and visual association cortex contribute to top-down modulation of visual processing. *Cerebral cortex (New York, N.Y. : 1991)* 17 Suppl 1: i125-135.
 128. Ruff, C. C., F. Blankenburg, O. Bjoertomt, S. Bestmann, E. Freeman, J. D. Haynes, G. Rees, O. Josephs, R. Deichmann, and J. Driver. 2006. Concurrent TMS-fMRI and psychophysics reveal frontal influences on human retinotopic visual cortex. *Current biology : CB* 16: 1479-1488.
 129. Slotnick, S. D., J. Schwarzbach, and S. Yantis. 2003. Attentional inhibition of visual processing in human striate and extrastriate cortex. *NeuroImage* 19: 1602-1611.
 130. Smith, A. T., K. D. Singh, and M. W. Greenlee. 2000. Attentional suppression of activity in the human visual cortex. *Neuroreport* 11: 271-277.
 131. Heinemann, L., A. Kleinschmidt, and N. G. Muller. 2009. Exploring BOLD changes during spatial attention in non-stimulated visual cortex. *PLoS One* 4: e5560.
 132. Gunduz, A., P. Brunner, A. Daitch, E. C. Leuthardt, A. L. Ritaccio, B. Pesaran, and G. Schalk. 2011. Neural correlates of visual-spatial attention in electrocorticographic signals in humans. *Frontiers in human neuroscience* 5: 89.

133. Miller, E. K., and T. J. Buschman. 2013. Cortical circuits for the control of attention. *Curr Opin Neurobiol* 23: 216-222.
134. Wise, R. J., S. K. Scott, S. C. Blank, C. J. Mummery, K. Murphy, and E. A. Warburton. 2001. Separate neural subsystems within 'Wernicke's area'. *Brain* 124: 83-95.
135. Wilson, S. M., I. Molnar-Szakacs, and M. Iacoboni. 2008. Beyond superior temporal cortex: intersubject correlations in narrative speech comprehension. *Cerebral cortex (New York, N.Y. : 1991)* 18: 230-242.
136. Okada, K., J. H. Venezia, W. Matchin, K. Saberi, and G. Hickok. 2013. An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex. *PLoS One* 8: e68959.
137. Wandell, B. A., A. A. Brewer, and R. F. Dougherty. 2005. Visual field map clusters in human cortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 360: 693-707.
138. Bernstein, L. E., E. T. Auer, Jr., M. Wagner, and C. W. Ponton. 2008. Spatiotemporal dynamics of audiovisual speech processing. *NeuroImage* 39: 423-435.
139. Chu, Y. H., F. H. Lin, Y. J. Chou, K. W. Tsai, W. J. Kuo, and I. P. Jaaskelainen. 2013. Effective cerebral connectivity during silent speech reading revealed by functional magnetic resonance imaging. *PLoS One* 8: e80265.
140. Bisley, J. W., and M. E. Goldberg. 2010. Attention, intention, and priority in the parietal lobe. *Annual review of neuroscience* 33: 1-21.

141. Lin, F. R., J. K. Niparko, and L. Ferrucci. 2011. Hearing loss prevalence in the United States. *Arch Intern Med* 171: 1851-1852.
142. Ingvalson, E. M., B. Lee, P. Fiebig, and P. C. Wong. 2013. The effects of short-term computerized speech-in-noise training on postlingually deafened adult cochlear implant recipients. *Journal of speech, language, and hearing research : JSLHR* 56: 81-88.

VITA

Müge Özker Sertel was born in Istanbul, Turkey on August 25, 1983, the daughter of Fevziye Özker and Taner Özker. After completing Istanbul High School in 2003, she entered Koç University in Istanbul, Turkey. She received the degree of Bachelor of Science in Electrical and Electronics Engineering in June of 2007. In September of 2007, she entered Boğaziçi University in Istanbul, Turkey and received the degree of Masters in Science in Biomedical Engineering in September of 2009. For the next year she continued her research in Boğaziçi University Biomedical Engineering Institute. In the following two years, she worked as a research assistant at Stanford University School of Medicine, Department of Neurology and Neurological Sciences. In August of 2012, she entered the PhD program in Neuroscience at The University of Texas Health Science Center at Houston.

Permanent address:

Toraman Sokak No: 14, 34467

Emirgan/Istanbul, Türkiye