The Modulatory Effects of Visual Speech on Auditory Speech Perception: A Multi-Modal Investigation of How Vision Alters the Temporal, Spatial and Spectral Components of Speech

by

Karthikeyan Ganesan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Psychology and Scientific Computing)
in the University of Michigan
2022

Doctoral Committee:

Assistant Professor David Brang, Chair
Dr. Andrew Jahn
Associate Professor Zhongming Liu
Professor Chandra Sripada
Professor Daniel Weissman

Karthikeyan Ganesan

gkarthik@umich.edu

ORCID iD: 0000-0002-3029-2273

Table of Contents

List of Tables

List of Figures

Abstract

Visual speech information, especially that provided by the mouth and lips, is important during face-to-face communication. This has been made more evident by the increased difficulty of speech perception because mask usage has become commonplace in response to the COVID-19 pandemic. Masking obscures the mouth and lips, thus eliminating meaningful information from visual cues that are used to perceive speech correctly. To fully understand the perceptual benefits afforded by visual information during audiovisual speech perception, it is necessary to explore the underlying neural mechanisms involved. While several studies have shown neural activation of auditory regions in response to visual speech, the information represented by these activations remain poorly understood. The objective of this dissertation is to investigate the neural bases for how visual speech modulates the temporal, spatial, and spectral components of audiovisual speech perception, and the type of information encoded by these signals.

Most studies approach this question by using techniques sensitive to one or two important dimensions (temporal, spatial, or spectral). Even in studies that have used intracranial electroencephalography (iEEG), which is sensitive to all three dimensions, research conventionally quantifies effects using single-subject statistics, leaving group-level variance unexplained. In Study 1, I overcome these shortcomings by investigating how vision modulates auditory speech processes across spatial, temporal and spectral dimensions in a large group of epilepsy patients with intracranial electrodes implanted (n = 21). The results of this study demonstrate that visual speech produced multiple spatiotemporally distinct patterns of theta, beta, and high-gamma power changes in auditory regions in the superior temporal gyrus (STG).

While study 1 showed that visual speech evoked activity in auditory areas, it is not clear what, if any, information is encoded by these activations. In Study 2, I investigated whether these distinct patterns of activity in the STG, produced by visual speech, contain information about what word is being said. To address this question, I utilized a support-vector machine classifier to decode the identities of four word types (consonants beginning with *'b', 'd', 'g',* and *'f'*) from activity in the STG recorded during spoken (phonemes: basic units of speech) or silent visual speech (visemes: basic units of lipreading information). Results from this study indicated that visual speech indeed encodes lipreading information in auditory regions.

Studies 1 and 2 provided evidence from iEEG data obtained from patients with epilepsy. In order to replicate these results in a normative population and to leverage improved spatial resolution, in Study 3 I acquired data from a large cohort of normative subjects (n = 64) during a randomized event-related functional magnetic resonance imaging (fMRI) experiment. Similar to that of Study 2, I used machine learning to test for classification of phonemes and visemes (/fafa/, /kaka/, /mama/) from auditory, auditory-visual, and visual regions in the brain. Results conceptually replicated the results of Study 2, such that phoneme and viseme identities could both be classified from the STG, revealing that this information is encoded through distributed representations. Further analyses revealed similar spatial patterns in the STG between phonemes and visemes, consistent with the model that viseme information is used to target corresponding phoneme populations in auditory regions. Taken together, the findings from this dissertation advance our understanding of the neural mechanisms that underlie the multiple ways in which vision alters the temporal, spatial and spectral components of audiovisual speech perception.

Chapter 1 Introduction

With rapid development in healthcare technologies and improvement in standards of living, many countries around the world face an unprecedented increase in their elderly population (Atella et al., 2019) with a corresponding increase in health complications (Fontana et al., 2014). These complications include risk factors for progression to chronic conditions that affect the quality of life of these individuals (Christensen et al., 2009). Age-related hearing loss is a major condition that significantly affects over a billion people worldwide and this number has been consistently rising over the years (Olusanya et al., 2014; Vos et al., 2016).

While the nature of hearing loss in individuals is varied, it is usually agreed to be progressive (Cruickshanks et al., 2010) with a gradual decline in the ability to perceive speech in various environments such as in the presence of noise (Agrawal et al., 2008). While degraded speech perception effects differ across individuals, understanding the nature of how individuals compensate for this decline in hearing abilities might help in the designing of interventions that improve quality of life for a large section of the population. One major way the brain improves hearing is through integrating visual speech signals with what is heard. This dissertation focuses on understanding the neural bases of how individuals integrate auditory speech signals with visual information during face-to-face communication. Investigating the neural mechanisms subserving the integration of information from both the auditory and visual modalities during spoken speech could help build better treatment and hearing-aid technologies for individuals with degraded speech perception abilities. Developing these treatments and technologies is

critical since it would help promote healthy aging, leading to improved quality of life, independent living, and reduced potential health costs for individuals concerned.

Background

The perceptual integration of auditory and visual cues is an important aspect of social communication. During natural speech, auditory speech signals are conveyed rapidly (3-7 syllables per second; Chandrasekaran et al., 2009), making the identification of individual speech sounds a computationally challenging task (Elliott and Theunissen, 2009). The integration of audiovisual information during face-to-face communication can therefore help to predict and constrain perceptual inferences about speech signals in both a bottom-up and top-down manner (Bernstein and Liebenthal, 2014; Lewis and Bastiaansen, 2015; Peele and Sommers, 2015). Several studies have shown that the integration of congruent speech information from the auditory and visual modalities results in perceptual enhancements for spoken speech compared to when auditory speech is presented alone (Hickock et al., 2018; Erber, 1969, Ross et al., 2007). To fully understand the ways that visual information benefits speech perception, it is necessary to explore the neural mechanisms involved.

Mounting evidence suggests that audiovisual speech interactions may occur at multiple functional-anatomical stages in the auditory cortex (Okada et al., 2013). In one such stage, it is thought that neuronal activity is evoked within the primary auditory areas during audiovisual interactions (Kayser et al., 2008, Werner-Reiss et al., 2003, Schroeder et al., 2005). However, very little is known about the kind of information that is encoded in these interactions. Moreover, it is difficult to localize the precise cortical origins of early multisensory integration.

To understand the processes underlying these audiovisual interactions, it is essential to investigate the types of information being transformed and integrated in auditory and

multisensory regions. Converging evidence overwhelmingly suggest that the superior temporal gyrus (STG) (Mesgarani et al., 2014) and posterior superior temporal sulcus (pSTS) (Olasagasti et al., 2015; Kayser and Logothetis., 2009) play a central role in the interaction and transformation of audiovisual information. This dissertation focuses on these two regions of interest to understand the neural basis of how visual speech influences auditory speech perception.

The neural mechanisms of audiovisual speech integration have traditionally been investigated using functional magnetic resonance imaging (fMRI) and, less often, intracranial electroencephalography (iEEG). Each of these methods has proven useful in providing evidence for the different ways in which audiovisual integration supports human speech perception. While iEEG provides high temporal resolution and the ability to investigate individual frequency bands associated with distinct activity, fMRI provides more precise spatial localization of effects and better generalizability to the broader population.

However, a multimodal investigation of audiovisual speech processes combining the advantages of both these modalities has yet to be performed. This dissertation focuses on the use of both methods to identify the discrete neural processes and networks involved in audiovisual speech perception, with a focus on the broader primary auditory cortex.

Visual influences on auditory speech perception

In audiovisual integration, multiple features extracted from visual signals can bias or enhance auditory speech perception processes, including lip shapes, rhythmic articulatory movements, and speaker identity, among others (Chandrasekaran et al., 2009; Erber, 1975; Chen and Rao, 1998; Van Wassenhove et al., 2005). While the net result is improved speech

perception, each of these features may influence cortical auditory processes through distinct mechanisms. For example, visual speech is thought to influence the temporal structure of auditory speech processing. Auditory speech signals that are temporally correlated with lip closure elicit stronger neural activity, resulting in the modulation of cortical excitability in auditory regions (Schroeder et al., 2008).

Converging behavioral and neurophysiological evidence also suggests that perceptual enhancements from audiovisual speech (e.g., better detection and faster reaction times) and visual recovery of auditory phoneme information (the smallest unit of speech; for example, the sound /fa/) are subserved by two distinct mechanisms (Eskelund et al., 2011; Plass et al., 2014). This distinction may reflect a neural dissociation between predictive multisensory interactions that optimize feedforward encoding of auditory information and later feedback processes that alter auditory representations generated in the auditory regions (Arnal et al., 2009; Arnal et al., 2011, Reale et al., 2007). In support of this view, both visual speech (Besle et al., 2004; Arnal et al., 2009; Van Wassenhove et al., 2005) and other anticipatory visual cues (Vroomen and Stekelenburg, 2010) can speed up and reduce the magnitude of early physiological responses associated with auditory feedforward processing. This could potentially reflect optimization of auditory encoding in accordance with temporal or acoustic constraints imposed by visual information.

Neural processes involved in audiovisual speech processing

These early feedforward effects, which are insensitive to audiovisual congruity in speech, are temporally, spatially, and spectrally distinct from later (>300 ms) responses that are specific to audiovisual incongruent speech (Arnal et al., 2011; Van Wassenhove et al., 2005). These later incongruity-specific interactions point to a hierarchical feedback regime in which unisensory

speech processing is altered in accordance with integrated audiovisual information from the pSTS (Olasagasti et al., 2015; Kayser and Logothetis., 2009) and the general speech perception areas in the STG (Mesgarani et al., 2014). It should also be noted that some of these patterns of activities might likely be due to attentional or non-specific effects.

Fine-grained analysis of audio-visual speech integration, with respect to the processes that enable multisensory speech perception, has focused on the left pSTS, which sits at the intersection of auditory, visual, and parietal regions (Beauchamp et al., 2010). These studies indicate that the pSTS is responsible for integrating contextual information between the fusiform face area (FFA) and early auditory regions (Ghanzafar et al., 2010, Ghanzafar et al., 2008, Zhu & Beauchamp, 2017). Furthermore, diffusion tensor imaging (DTI) studies in humans have demonstrated the presence of connections between the FFA and pSTS regions, suggesting that a structural network may support functional interactions between these regions (Blank et al., 2011). Consistent with behavioral evidence that visual information can alter what is heard (such as in the McGurk effect), visual signals strongly modulate the response of auditory neurons to sounds (Ghanzafar et al., 2010, Ghanzafar et al., 2008, Zhu & Beauchamp, 2017). However, recent evidence suggests that phoneme-viseme correspondences are insufficient to account for the perceptually detailed information provided by visual speech (Bernstein et al., 2014), calling into question whether the pSTS is the sole region responsible for auditory-visual speech interactions.

Specifically, both normal-hearing and deaf observers can readily distinguish between visual words composed of the same visemes, suggesting that, like auditory speech, visual speech conveys additional fine-grained information beyond what is encoded in coarse categorical representations (Bernstein et al., 2014). Because visual speech facilitates perception for both

spectral details and temporal dynamics in speech (Plass et al., 2020), it could plausibly enhance perception through multiple distinct influences on not just the pSTS, but also the STG which is specialized for different aspects of the auditory speech perception. Importantly, prior research indicates that some audiovisual speech processes are associated with neural activity in distinct frequency bands, suggesting that they likely correspond to unique integrational functions across the sensory hierarchy (Arnal et al., 2009; Kaiser 2005; Kaiser 2006; Peele and Sommers, 2015).

Furthermore, early hemodynamic studies defined a broad network of brain regions believed to be involved in multisensory integration, including areas in the frontal, parietal and temporal lobes (Hall et al., 2005). Though these earlier studies showed increased activation in extrastriate cortex, inferoposterior temporal lobe, angular gyrus and superior temporal gyrus in response to silent-lip reading (Calvert et al., 1997), more recent studies with iEEG recordings have shown responses to similar stimuli in the temporo-occipital junction, posterior medial temporal gyrus (pMTG) and superior temporal gyrus (Besle et al, 2008). Similar results have also been reported using information decoding algorithms based on deep neural networks used to analyze iEEG signals (Zweig et al., 2016).

In order to gain an improved understanding of the ways in which visual information benefits speech perception, it is necessary to identify the discrete neural processes and networks involved, especially in broader auditory regions. These open questions are best addressed with a multimodal investigation of audiovisual speech processes using fMRI and iEEG. Both methods have complementary advantages. For example, while fMRI can show task-related blood-oxygen level dependent (BOLD) activation of the STG during silent lipreading (Beauchamp et al., 2004), this method does not provide information about the timing or spectral composition of these responses.

Conversely, while intracranial electroencephalography (iEEG) provides substantially more temporal and spectral information, studies on audiovisual speech integration have largely focused on high-gamma power (HGp) indexes of local population firing rates (e.g., Micheli et al., 2020; Besle et al., 2008, Reale et al., 2007, Gyol Yi et al., 2019) and iEEG provides limited spatial coverage. However, the question of how audiovisual integration occurs in other spectral components including the theta and the beta band have not yet been investigated with iEEG. Moreover, analyses of iEEG data have traditionally been performed using single-subject statistics, which severely limit inferences about normative neural processing and fail to account for variance that group-level analyses correctly model.

To investigate the neural basis for how visual speech supports speech perception, it is therefore necessary to study these processes at multiple levels including their spatial, temporal and spectral dimensions. This dissertation aims to tackle the issue of studying the temporal, spatial and spectral components involved in visual modulation of auditory speech perception using a multimodal approach, leveraging results from iEEG signals and fMRI.

The current study

Most previous studies focus on the question of how vision affects multisensory regions in the auditory cortex without considering the information missing from the methods used (Besle et al., 2004, Besle et al, 2008, Micheli et al., 2020). More specifically, they do not consider the spectral, temporal or spatial differences in how these modulations could occur. Past studies ignore the differences in how visual speech could modulate auditory speech perception in these multiple dimensions. This motivated the hypothesis for Study 1, in which I addressed the question of how visual modulations of auditory speech differ across spatial, temporal and

spectral dimensions in the primary auditory cortex from a large cohort of iEEG patients (n = 22). Crucially, a major contribution of this study was an improvement from traditional single subject statistics with iEEG and the development of a more robust group-level analysis to obtain more generalizable inferences.

In Study 2, I used iEEG to test the hypothesis that some of the visual-evoked activity in the auditory system reflects the transformation of lipread signals into phonemic information. Specifically, I used support-vector machine (SVM) classifiers across spatial and temporal dimensions to demonstrate that basic visual speech features (visemes) can be decoded from auditory areas, consistent with our model. More generally, this study adds to the growing body of evidence that it is possible to decode information from cortical activity recorded using iEEG during audiovisual speech. Furthermore, this is the first study in literature which demonstrates that it is possible to decode the identity of visemes from visual regions, let alone from auditory regions.

In Study 3, I used fMRI to further test the hypothesis that some of the visual-evoked activity in the auditory system reflects phonemic information that is extracted from lipreading. Specifically, I applied searchlight and region of interest (ROI) classifiers to identify where in the brain spatial representations encode phoneme and viseme identities. This is a complementary experiment to that examined in Study 2. First, we conceptually replicated the results from Study 2, helping to generalize the results to a large normative population (n = 64). Second, it provided whole-brain spatial coverage which allowed a hierarchical comparison of the regions in which phonemes and visemes could be classified. Third, the high-spatial resolution of STG activity enabled high-resolution representational similarity analyses (RSA) to be constructed based on

spatial patterns of activity, revealing similar distributions across corresponding visemes and phonemes.

In sum, these studies support a model in which visemes extracted from lipreading evoke distinct spatial patterns of activity in the auditory system that overlap with those of corresponding phoneme representations. This further clarifies the nature of visual speech processing in auditory regions by providing a deep understanding of the neural processes underlying audiovisual integration during speech perception. The results also improve upon previous studies by utilizing group-level analysis of iEEG signals in place of single-subject statistics.

Chapter 2 Visual Speech Differentially Modulates Beta, Theta and High Gamma Bands in The Auditory Cortex

Abstract

Speech perception is a central component of social communication. While principally an auditory process, accurate speech perception in everyday settings is supported by meaningful information extracted from visual cues (e.g., speech content, timing, and speaker identity). Previous research has shown that visual speech modulates activity in cortical areas subserving auditory speech perception, including the superior temporal gyrus (STG), potentially through feedback connections from the multisensory posterior superior temporal sulcus (pSTS).

However, it is unknown whether visual modulation of auditory processing in the STG is a unitary phenomenon or, rather, consists of multiple temporally, spatially, or functionally distinct processes. In this context, a unitary phenomenon would indicate a single pathway (either anatomically or functionally) that processes visual information during audiovisual speech processing. To explore these questions, we examined neural responses to audiovisual speech measured from intracranially implanted electrodes within the temporal cortex of 21 patients undergoing clinical monitoring for epilepsy. We found that visual speech modulated auditory processes in the STG in multiple ways, eliciting temporally and spatially distinct patterns of activity that differed across theta, beta, and high-gamma frequency bands. For the theta band, visual speech suppressed the auditory response from before auditory speech onset to well after auditory speech onset (-93 ms to 500 ms) most strongly in the posterior STG. For the beta band,

suppression was seen in the anterior STG from -311 to -195 ms before auditory speech onset and in the middle STG from -195 ms to 235 ms after speech onset. For high gamma, enhanced activity was seen from -45 ms to 24 ms only in the posterior STG.

We interpret the visual-induced changes prior to speech onset as reflecting crossmodal prediction of speech signals in these areas. In contrast, modulations after sound onset may reflect a decrease in sustained feedforward auditory activity. These results are consistent with models that posit multiple distinct mechanisms supporting audiovisual speech perception and provide a crucial map for subsequent studies to identify the types of visual features that are encoded by these separate mechanisms.

2.1 Introduction

Auditory speech signals are conveyed rapidly during natural speech (3-7 syllables per second; Chandrasekaran et al., 2009), making the identification of individual speech sounds a computationally challenging task (Elliott and Theunissen, 2009). Easing the complexity of this process, audiovisual signals during face-to-face communication help predict and constrain perceptual inferences about speech sounds in both a bottom-up and top-down manner (Bernstein and Liebenthal, 2014; Lewis and Bastiaansen, 2015; Peelle and Sommers, 2015).

Multiple features extracted from visual signals can bias or enhance auditory speech perception processes, including lip shapes, rhythmic articulatory movements, and speaker identity, among others (Chandrasekaran et al., 2009; Erber, 1975; Chen and Rao, 1998; Van Wassenhove et al., 2005). While the net result is improved speech perception, each of these features may influence cortical auditory processes through distinct mechanisms. For example, visual speech is thought to influence the temporal structure of auditory speech processing by

neurally amplifying auditory speech signals that are temporally correlated with lip closure, accomplished by modulating cortical excitability in auditory regions (Schroeder et al., 2008).

Indeed, functional dissociations are readily found in the auditory system. In the speech domain, research indicates that the superior temporal gyrus (STG) exhibits an anterior-posterior gradient in feature tuning, with anterior regions being more sensitive to spectral content and posterior regions being more sensitive to temporal information (e.g., broadband amplitude dynamics) (Hullet et al., 2016). Because visual speech facilitates perception for both spectral details and temporal dynamics in speech (Plass et al., 2020), it could plausibly enhance perception through multiple distinct influences on STG areas specialized for different aspects of the auditory speech signal. Importantly, prior research indicates that some audiovisual speech processes are associated with neural activity in distinct frequency bands, suggesting that they likely correspond to unique integrational functions across the sensory hierarchy (Arnal et al., 2009; Kaiser 2005; Kaiser 2006; Peelle and Sommers, 2015). Similarly, studies have demonstrated audiovisual speech effects at multiple time points, including during the observation of preparatory lip movements and after speech onset (Besle et al., 2008). However, identifying the specific role of each mechanism would be helped by first identifying different functional processes that are altered by visual speech (e.g., the modulatory effect of visual speech in different oscillatory frequency bands at different spatial and temporal scales).

Audiovisual speech integration studies using invasively implanted electrodes (intracranial electroencephalography; iEEG) have focused on raw signal amplitudes (Besle et al., 2008) or surrogate measures of population action potentials through high-gamma filtered power (HGp) (e.g., Micheli et al., 2020), showing early activation of auditory areas to audiovisual speech. However, these studies did not analyze the spectral composition of auditory-visual effects in

low- and high-frequency ranges, that can reflect distinct forms of information processing (Wang, 2010; Engel and Fries, 2010; Ray, Crone, Niebur, Franaszczuk, and Hsiao, 2008), and have tended to use small sample sizes and single-participant statistics (e.g., Micheli et al., 2020; Besle et al., 2008). Conversely, non-invasive EEG studies have investigated the influence of visual speech information on low-frequency signals, with strong effects on beta and theta activity at different time scales (Sakowitz et al., 2005). However, as low- and high-frequency effects were observed across separate studies and given limitations of each approach (poor spatial resolution with EEG and small sample sizes with iEEG), the interdependence of these processes remains unclear.

Thus, at present the field lacks a unified framework for how visual speech information alters responses within auditory regions. This study sought to fill this gap by examining the interdependence of spatial, temporal, and spectral effects during audiovisual speech perception in a large cohort of patients with iEEG recordings (745 electrodes implanted in auditory areas of 21 individuals) who performed an audiovisual speech task while undergoing clinical monitoring for epilepsy. Specifically, we examined visual effects on auditory speech processes across multiple frequency bands associated with both subthreshold oscillations and neural firing. Moreover, to integrate statistical results across participants, we used linear mixed-effects models to perform statistical inference at the group level, facilitating generalization, and compared observed effects to those seen at the single participant level. Analyzing these data using group-level statistics, we found that visual speech produced multiple spatiotemporally distinct patterns of theta, beta, and high-gamma power throughout the STG. These results are consistent with the view that visual speech enhances auditory speech processes through multiple functionally distinct mechanisms and provides a map for investigating the information represented in each process.

2.2 Materials and Methods

2.2.1 Participants, implants and recordings

Data were acquired from 21 patients with intractable epilepsy undergoing clinical evaluation using iEEG. Patients ranged in age from 15-58 years (mean = 37.1, SD = 12.8) and included 10 females. iEEG was acquired from clinically implanted depth electrodes (5 mm center-to-center spacing, 2 mm diameter) and/or subdural electrodes (10 mm center-to-center spacing, 3 mm diameter): 13 patients had subdural electrodes and 17 patients had depth electrodes (Figure 11). Across all patients, data was recorded from a total of 1367 electrodes (mean = 65, SD = 25.3, range = 24 - 131 per participant). The number, location, and type of electrodes used were based on the clinical needs of the participants. iEEG recordings were acquired at either 1000 Hz (n = 5), 1024 Hz (n = 11 participants), or 4096 Hz (n = 5 Participants) due to differences in clinical amplifiers. All participants provided informed consent under an institutional review board (IRB)-approved protocol at the University of Chicago, Rush University, University of Michigan, or Henry Ford hospital.

2.2.2 MRI and CT acquisition and processing

Preoperative T1-weighted magnetic resonance imaging (MRI) and a postoperative computed tomography (CT) scans were acquired for all participants. Registration of the pre-operative MRI to postoperative CT was performed using the 'mutual information' method contained in SPM12 (Viola and Wells, 1997; Penny et al., 2006); no reslicing or resampling of the CT was used. Electrode localization was performed using custom software (Brang et al., 2016; available for download online https://github.com/towle-lab/electrode-registration-app/). This algorithm identifies and segments electrodes from the CT based on intensity values and

projects subdural electrodes to the dura surface using the shape of the electrode disk to counteract postoperative compression. The Freesurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/; Dale, Fischl, and Sereno 1999; Fischl, Sereno, and Dale, 1999) was used for subsequent image processing procedures including cortical surface reconstruction, volume segmentation, and anatomical labelling (http://surfer.nmr.mgh.harvard.edu/; Dale, Fischl, and Sereno 1999; Fischl, Sereno, and Dale, 1999).

2.2.3 Tasks and Stimuli

Participants were tested in the hospital at their bedside using a 15-inch MacBook Pro computer running Psychtoolbox (Kleiner et al., 2007). Auditory stimuli were presented through a pair of free-field speakers placed approximately 15 degrees to each side of the patients' midline, adjacent to the laptop. Data were aggregated from three audiovisual speech perception paradigms (using different phonemes spoken by different individuals across tasks) to ensure generalizability of results and an adequate sample for group-analyses: 7 participants completed variant A, 8 participants variant B, and 6 participants variant C. Each task presented participants with auditory and visual speech stimuli in various combinations. As this study examines the modulatory role of visual information on auditory processes, only the auditory-only and audiovisual (congruent auditory/visual signals) conditions were analyzed from each task variant.

On each trial a single phoneme was presented to the participant (variant A: /ba/ /da/ /ta/ /tha/, variant B: /ba/ /da/ /ga/, variant C: /ba/ /ga/ /ka/ /pa/). Figure 1 shows the timing and structure of an example trial from task variant B. Trials began with a fixation cross against a black screen that served as the intertrial interval (ITI), presented for an average of 750 ms (random jitter plus or minus 250 ms, uniformly sampled). In the audiovisual condition, the face

appeared either 750 ms before sound onset (task variant B) or 500 ms before sound onset (variants A and C); across all three variants, face motion began at 500 ms before sound onset. In the auditory-only condition, either the fixation cross persisted until sound onset (variant A) or a uniform gray square (mean contrast of the video images and equal in size) was presented for either 750 ms before sound onset (variant B) or 500 ms before sound onset (variant C). Trials were presented in a random order and phonemes were distributed uniformly across conditions. While conditions were matched in terms of trial numbers, participants completed a variable number of trials (based on task variant and the number of blocks completed): mean = 68 trials per condition (SD = 23, range = 32-96). Onset of each trial was denoted online by a voltage isolated transistor-transistor logic (TTL) pulse.

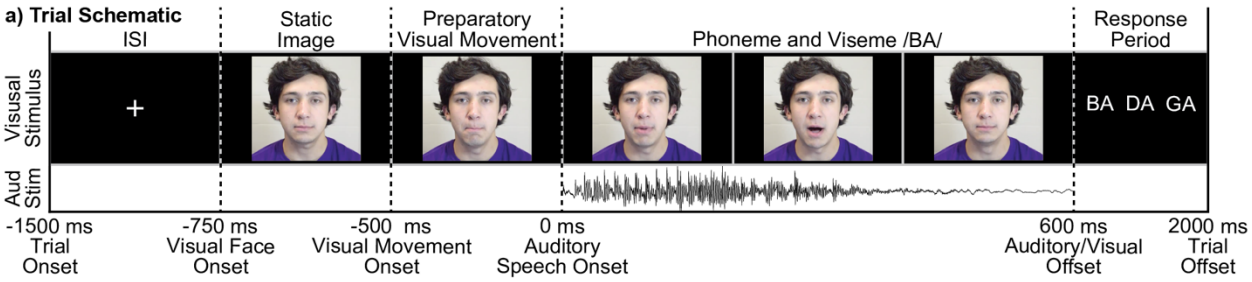


Figure 1. ***Schematic of the task.*** *Task Variant B trial schematic. All trials began with a fixation cross 1500 ms before the onset of an auditory stimulus, lasting for an average of 750 ms (plus or minus 250 ms jitter). In the auditory-only condition a blank screen followed the fixation cross for 750 ms. In the audiovisual condition the face appeared at the offset of the fixation (750 ms before sound onset), with preparatory visual movement beginning 250 ms later. Auditory phonemes (/ba/, /da/, or /ga/) onset at 0 ms in both conditions.*

In variants A and B, following each trial participants were prompted to identify which phoneme they had heard either aloud or via button press. In variant C, participants were cued to identify a phoneme on only 20% of the trials (data not analyzed). As auditory stimuli were presented without additional noise, we anticipated high levels of accuracy. Consistent with this, in variants A and B accuracy did not differ across auditory-only and audiovisual conditions

(behavioral data was unavailable for one participant): auditory-only mean accuracy = 95.3% (SD = 6.0%), audiovisual mean accuracy = 95.8% (SD = 6.4%), $t(13) = 0.518$, $p = .61$.

2.2.4 iEEG Data Preprocessing

Data were referenced in a bipolar fashion (signals subtracted from each immediately adjacent electrode in a pairwise manner) to ensure that the observed signals were derived from maximally local neuronal populations. Only electrodes meeting anatomical criteria within auditory areas were included in analyses. Anatomical selection required that an electrode be proximal to an auditory temporal lobe region as defined by the Freesurfer anatomical labels superiortemporal, middletemporal, and supramarginal in MNI space, resulting in 765 bipolar electrode pairs. Excessively noisy electrodes (either manually identified or due to variability in the raw signal greater than 5 SD compared to all electrodes) were removed from analyses, resulting in 745 remaining electrodes; across participants the mean proportion of channels rejected was 3.3% (SD = 8.7%, Range = 0 to 37.5%).

Slow drift artifacts and power-line interference were attenuated by high-pass filtering the data at .1 Hz and notch-filtering at 60 Hz (and its harmonics at 120, 180, and 240 Hz). Each trial was then segmented into a 2-second epoch centered around the onset of the trial. Individual trials were then separately filtered into three frequency ranges using wavelet convolution and then power transformed: theta (3 - 7 Hz, wavelet cycles varied linearly from 3-5), beta (13 - 30 Hz, wavelet cycles varied linearly from 5-10), HGp (70 - 150 Hz in 5 Hz intervals, wavelet cycles = 20 at 70 Hz, and increased linearly to maintain the same wavelet duration across frequencies); data were then resampled to 1024 Hz. Theta, beta, and HGp were selected based on previous reported findings of auditory-visual speech integration effects in these ranges (e.g., Arnal et al., 2009; Kaiser 2005; Kaiser 2006; Peelle and Sommers, 2015; Micheli et al., 2020). Within each

frequency range and evaluated separately at each electrode, we identified outliers in spectral power at each time point that were 3 scaled median absolute deviations from the median trial response. Outlier values were replaced with the appropriate upper or lower threshold value using the 'clip' option of the Matlab command 'filloutliers'. Across participants, a mean of .2% of values were identified as outliers (SD = .1%, Range = .1 to .5%).

Though electrodes were implanted in both the left and right hemispheres, electrodes were projected into the left hemisphere for visualization and analyses. This was accomplished through registering each participant's skull-stripped brain to the cvs_avg35_inMNI152 template image through affine registration using the Freesurfer function mri_robust_register (Reuter, Rosas, Fischl, 2010). Right-hemisphere electrode coordinates were then reflected onto the left hemisphere across the sagittal axis.

Functional selection was evaluated separately for each of the three frequency bands of interest (theta, beta, and HGp) to identify auditory-responsive electrodes: accordingly, different electrode numbers were included across each of the frequency analyses. To ensure orthogonality with the examined condition differences, the functional localizer required electrodes to demonstrate a significant post-stimulus response (0 - 500 ms) regardless of condition relative to zero using one-sample *t*-tests after correcting for multiple comparisons using false discovery rate (FDR). Beta and theta selection applied two-tailed t-tests while HGp applied one-tailed t-tests (as meaningful auditory HGp responses were predicted to elicit HGp increases (Beauchamp., 2016). Only electrodes meeting both anatomical and functional criteria were included in analyses (*n* = 465).

2.2.5 Group-Level Analyses

Traditionally, iEEG studies have focused on individual-participant analyses utilizing fixed-effect statistics (e.g., Micheli et al., 2018; Besle et al., 2008; Chang et al, 2010; Plass et al., 2020). The main reasons for this are small sample sizes and difficulty in transforming data into a common reference plane such as the MNI space. While these approaches are valid for estimating parameters and effect sizes within a single individual, they do not provide estimates across participants and thus lack generalizability across epilepsy patients, making inferences to the general population more difficult. Moreover, some studies mix between- and within-participant statistics by aggregating data from all participants without modeling participant as a random effect, violating independence assumptions (e.g., Lega, Germi, Rugg, 2017). This approach has been discussed extensively under the title of 'pseudoreplication' and can lead to spurious and poorly generalized results (for a discussion see Aarts et al., 2014; Lazic 2010; Lazic et al., 2018). These concerns for iEEG research have been raised and theoretically addressed previously by other groups using variants of a mixed-effects model (Kadipasoglu et.al., 2014; Kadipasoglu et al., 2015). To overcome these limitations, we employed two separate analysis approaches.

2.2.6 Group-Level Spatial Analyses

To identify regions of the auditory temporal lobe that responded differently to auditory-only versus audiovisual stimuli, we conducted individual-participant statistics and aggregated data across participants using an approach from the meta-analysis literature (treating each participant as an independent replication). Specifically, each 'virtual' bipolar electrode (calculated as the average coordinates between the associated pair of electrodes) was transformed into MNI space (Freesurfer cvs_avg35_inMNI152) and linked to neighboring vertices (within 10 mm Euclidean distance) on the Freesurfer MNI cortical pial surface (decimated from 1 mm to 4

mm); this one-to-many approach mitigates the imperfection of cross-participant spatial registration. Next, statistics were evaluated separately at each vertex for each participant using independent-sample t-tests, to compare auditory-only and audiovisual trials between -1000 to 500 ms (auditory-onset at 0 ms; data were averaged across 100 ms time-windows prior to statistical analyses). Within-participant statistics were adjusted for multiple comparisons across vertices and time using FDR (Groppe, Urbach, and Kutas, 2011). The approach yielded individual-participant *p*-value maps at each of the 15 time-points. *P*-value maps were then aggregated across participants using Stouffer's Z-score method (Stouffer et al., 1949).

2.2.7 Group-Level Regional Time-Series Analyses

While the meta-analysis approach establishes the strength of an effect at the group-level, it fails to provide group-level estimates and cannot effectively model data from both within and between participants (as is necessary in the evaluation of interactions across time, space, and analyzed frequency ranges). To model more general group-level differences between auditory-only and audiovisual conditions we used linear mixed-effects models. Because appropriately fitted models require more data than is often present at a single vertex, we created three regions of interest (ROIs) within the STG. ROIs were divided into three equal partitions from the "superiortemporal" label in Freesurfer, comprising anterior, middle, and posterior regions, similar to the division of the STG used previously (Smith et al., 2013). Electrodes within 10 mm of these labels were linked to the closest of the three (no electrode was linked to multiple labels). Our focus on the STG was motivated by previous demonstrations of strong effects of lipreading in this region (e.g., Smith et al., 2013). A numerical breakdown of the number of electrodes and participants in each of the three regions of the STG is provided in Table 1.

Linear mixed-effects modeling was performed using the fitlme function in Matlab R2019a (Mathworks Inc., Natwick, MA). Electrodes in the same ROI from the same participant were averaged prior to analysis to reduce the complexity of the model and as neighboring electrodes share variance. Individual trials were not averaged within or across participants prior to analysis. Nine main-effect models were constructed, in which differences between auditory-only and audiovisual trials were separately evaluated at each of the three STG ROIs (anterior, middle, posterior) and three frequency bands (theta, beta, HGp) using the equation: $y_{ij} = \beta_0 + (\beta_1 + u_{1,j})participant_{ij} + u_{0,j} + \varepsilon_{i,j}$, where, *y* represents the ECoG trial value, with a fixed effects term for the trial condition and a random intercept and slope term for the participant ID. In Matlab notation, this is represented as: ECoG_Trial_Value ~ Trial_Cond + (Trial_Cond|Participant_ID). Critically, we modeled both random intercepts and random slopes for trial condition as there were multiple measurements per participant and to maintain 'maximal' models for confirmatory hypothesis testing (Barr et al., 2013). Statistics for the main-effect models were adjusted for comparisons at multiple time-points from -500 to 500 ms using FDR correction (q = .05) (Groppe, Urbach, and Kutas, 2011).

Interaction models were subsequently constructed to evaluate whether audiovisual versus auditory-only condition effects varied as a function of frequency band, ROI, and time, using the Matlab notation: ECoG_Trial_Value ~ Trial_Cond * FrequencyBand * ROI * Time + (Trial_Cond|Participant_ID). While these model parameters were selected for inclusion based on confirmatory hypothesis testing, we also justified model selection using AIC comparisons. Separate models were constructed at each 1 ms time-point for the main-effect models (shown in Figures 6-8). Data were averaged in 5 ms time-bins for the interaction models due to computational complexity and memory requirements. The inclusion of time as a random factor in

interaction models may appear to violate the assumption of independence as spectral power demonstrates autocorrelations. However, the inherent characteristics of the mixed effect model's covariance structure should account for this dependence (Riha et al., 2020; Barr et al., 2013).

More generally, calculating degrees of freedom with linear mixed-effect models is a readily acknowledged challenge (e.g., Luke 2017). Acknowledging this, model significance was estimated using residual degrees of freedom. To ensure that the likely inflated degrees of freedom did not drive our effects, we additionally examined effects using a conservative estimation of degrees of freedom, based only on the number of participants who contributed data to a particular analysis (maximum of 21); all interactions that were significant remained significant (at $p < .001$).

2.2.8 Individual Electrode Analyses

To examine individual differences in the patterns of activity evoked across electrodes and participants, individual electrode statistics were examined at representative electrodes. Unpaired t-tests were conducted separately at each time-point comparing audiovisual versus auditory HGp (random factor = trial). Statistics for the main-effect models were adjusted for comparisons at multiple time-points from -500 to 500 ms using FDR correction ($q = .05$).

To examine whether one audiovisual effect predicted another or whether audiovisual effects arose from the same electrodes, we examined the linear relationship between audiovisual effects at separate frequency bands and time-windows, measured across individual electrodes. Electrodes were localized to the anterior, middle, and posterior STG and examined separately as three regions of interest. Activity in each frequency band was averaged across time ranges to capture observed audiovisual effects based on single frequency analyses: Pre-Aud HGp, -45 to 0 ms; Post-Aud HGp, 0 to 24 ms; Pre-Aud Theta, -93 to 0 ms; Post-Aud Theta, 0 to 500 ms; Pre-

Aud Beta, -311 to 0 ms; Post-Aud Beta, 0 to 235 ms. Trials were averaged within each electrode and subtracted across conditions (auditory-only minus audiovisual) to yield audiovisual effects. Relationships were estimated using linear mixed effect models similar to those above, using the Matlab notation: ECoG_Electrode_Value_Effect1 ~ ECoG_Electrode_Value_Effect2 + (1|Participant_ID). Effect 1 and 2 in this context reflect either frequency pairs (e.g., does the pre-auditory high-gamma effect predict the post-auditory beta suppression effect?) or time ranges (e.g., is the audiovisual effect in the beta band before auditory onset related to the audiovisual effect in the beta band after auditory onset?). Adding the additional slope parameter to the model failed to explain significantly more variance.

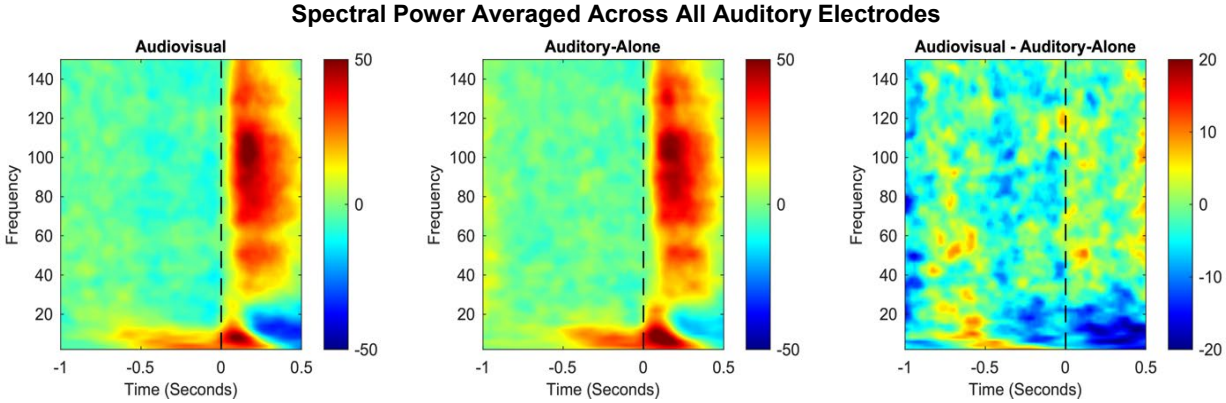


Figure 2. ***Group-level spectral plots****. Group-level plots showing event-related spectral power from 2-150 Hz. Data reflect ECoG activity from all anatomically localized electrodes (n = 745), first averaged across electrodes within each participant, then averaged across participants. Dotted lines denote auditory onset. Color scale reflects normalized power.*

2.3 Results

2.3.1 Group-Level Spatial Analyses

Figure 2 shows the spectro-temporal plot of the event related spectral power (ERSP) for audiovisual signals from all auditory electrodes across all participants. Data demonstrate that spectral power was distributed over multiple frequency bands while audiovisual stimuli were

presented: increased power in theta and high-gamma ranges, along with beta suppression. This, supported by past studies, provides justification for subsequent analyses focusing on these three frequency bands.

2.3.2 Group-Level Spatial Analyses: Theta Power

Figure 3 shows group-level differences in theta power (3 - 7 Hz) between audiovisual and auditory-only trials. A small but significant difference (audiovisual > auditory) emerged from -700 to -600 ms before sound onset in the supramarginal gyrus (peak coordinates: x = -60.7, y = -56.2, z = 30.3 , p = 0.001) with a peak-response in this region between -600 to -500 ms before sound onset (peak coordinates: x = -60.7, y = -56.2, z = 30.3, p = 0.0003). This activation pattern reflected only a small percentage of the supramarginal gyrus (SMG) (1.7% of SMG vertices at time-point -700 to -600 ms, and 2.6% of SMG vertices at time-point -600 to -500 ms).

In contrast to this initial pattern, the majority of condition differences were observed in the middle temporal gyrus (MTG) and STG with significantly more power in auditory trials compared to audiovisual trials. This pattern emerged as early as -300 to -200 ms (peak coord: x = -47.2, y = -33, z = -4.3, p = 0.0003) and peaked during the time range 100 to 200 ms following sound onset (peak coord: x = -60.8, y = -20.5, z = 11.4, p = 9.6e-12). The greatest proportion of significant vertices were observed from 200 to 300 ms (STG = 27.5%, MTG = 12.4%, SMG = 5.4%), strongly weighted towards the middle to posterior STG. These data suggest that the majority of theta-related activity during audiovisual speech processing occurs following sound onset.

2.3.3 Group-Level Spatial Analyses: Beta Power

Figure 4 shows group-level differences in beta power (13 - 30 Hz) between audiovisual and auditory-only trials. As was observed in the theta band, a small but significant difference (audiovisual > auditory) emerged from -700 to -600 ms before sound onset in the supramarginal gyrus (peak coordinates: x = -60.1, y = -24.6, z = 15, p = 0.005; .2% of SMG vertices were significant); no other significant audiovisual > auditory differences were observed throughout the time-series. In contrast to this initial pattern, the majority of condition differences were observed in the STG with significantly more power in auditory trials compared to audiovisual trials; this observation of reduced beta power is most consistent with increased beta suppression (see Section 3.7 for additional evidence).

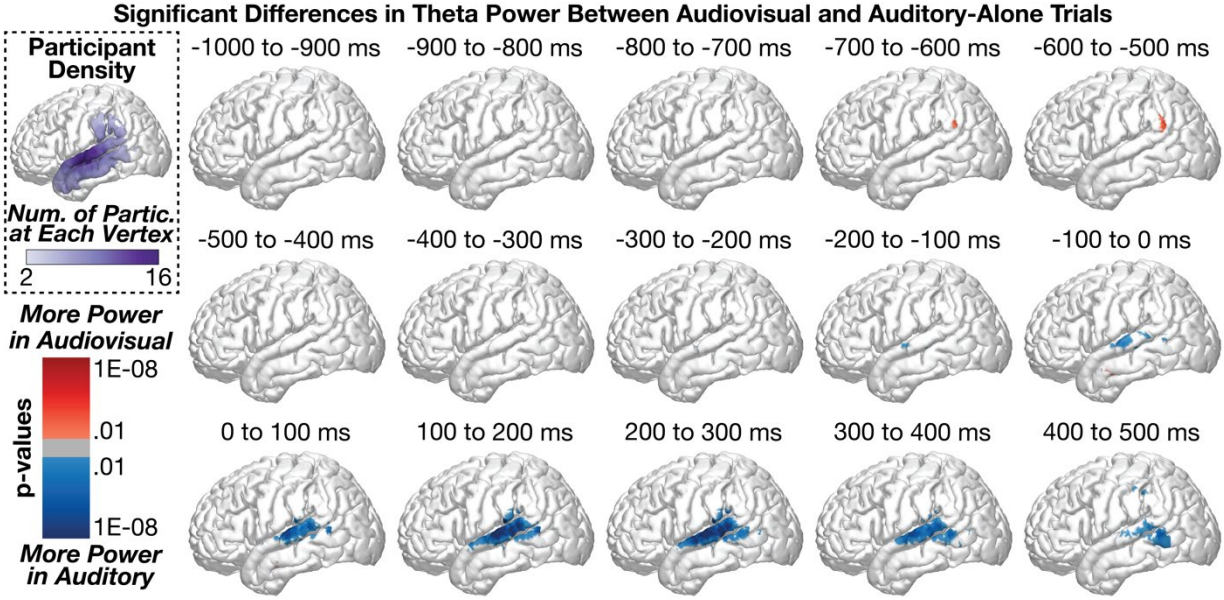


Figure 3. ***Group level plots in the theta band****. Group-level analyses comparing theta power between audiovisual and auditory-only conditions at 100 ms time-windows (sound onset at 0 ms). Statistics conducted vertex-wise at the individual participant level and aggregated across participants using Stouffer's Z-Score method. Multiple comparisons applied across time and space using FDR. Top-left plot shows the number of participants who were included at each vertex. Congruent audiovisual stimuli elicited reduced theta power at the middle to posterior STG, peaking after the onset of the speech sound.*

This pattern emerged as early as -400 to -300 ms (peak coord: x = -61.8, y = -1, z = -11.8, p = 0.0001) along the anterior to middle STG/MTG and peaked -200 to -100 ms before sound

onset (x = -65, y = -10, z = 0.9, p = 8.8e-08); the majority of significant vertices during this time range were in the STG: STG = 16.2%, MTG = 3.1%, SMG = 2.7%. Whereas the peak activation occurred from the -200 to -100 ms time-window, the greatest proportion of significant vertices were observed in the -100 to 0 ms time-window range: STG = 20.6%, MTG = 4.9%, SMG = 2.3%. These data suggest that the majority of beta-related activity during audiovisual speech processing occurs before sound onset in contrast to the spatial and temporal pattern of results observed for theta band activity. See Section 3.9 for a direct comparison of the spatiotemporal effects between theta and beta band activity. As the differences did not emerge until after face-onset but immediately prior to sound onset (i.e., during which time preparatory visual movements were observed by participants), we interpret these results to reflect predictive coding information along the STG (e.g., Bastos et al., 2012; Peelle and Sommers, 2015).

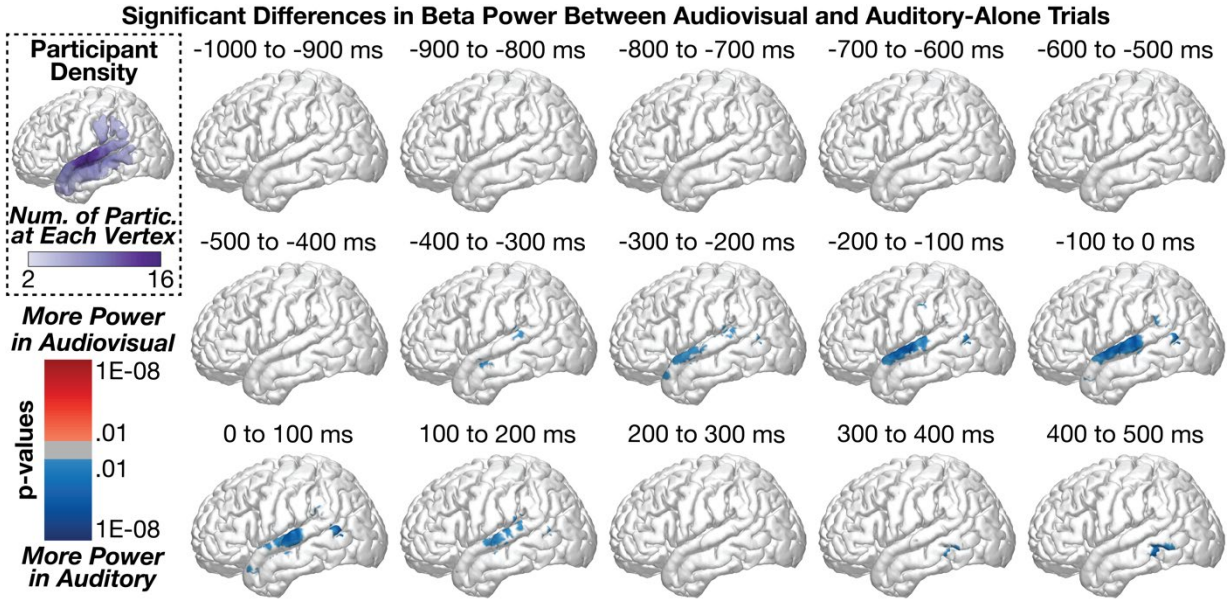


Figure 4. ***Group level plots in the beta band****. Group-level analyses comparing beta power between audiovisual and auditory-only conditions at 100 ms time-windows (sound onset a 0 ms). Top-left plot shows the number of participants who contributed data to each vertex. audiovisual stimuli elicited greater beta suppression at the posterior STG, peaking before sound onset.*

2.3.4 Group-Level Spatial Analyses: High-Gamma Power

Figure 5 shows group-level differences in high-gamma power (HGp; 70 - 150 Hz) between audiovisual and auditory-only trials. The first significant time-points in the series were observed in the MTG (audiovisual > auditory) beginning from -700 to -600 ms (peak coord: x = -55.8, y = -63.2, z = 8.4, p = 1.5e-07). Small clusters of effects were observed between -600 to -100 ms (all effects reflected less than 5% of the number of vertices in each region). Beginning from -100 to 0 ms, however, we observed a strong cluster of significant differences (audiovisual > auditory) in the MTG and STG (peak coord: x = -57.4, y = -66.6, z = 9.4, p = 8.1e-12, Region = MTG, percent significant vertices in each region: STG = 8.2%, MTG = 10.1%, SMG = 1.7%). This effect persisted throughout the time-series but shifted more inferior to the MTG by 400 to 500 ms (proportion significant vertices in each region: STG = 1.4%, MTG = 12.2%, SMG = 0%). In contrast to results in the theta and beta frequency bands, HGp effects were largely restricted to the posterior STG/MTG.

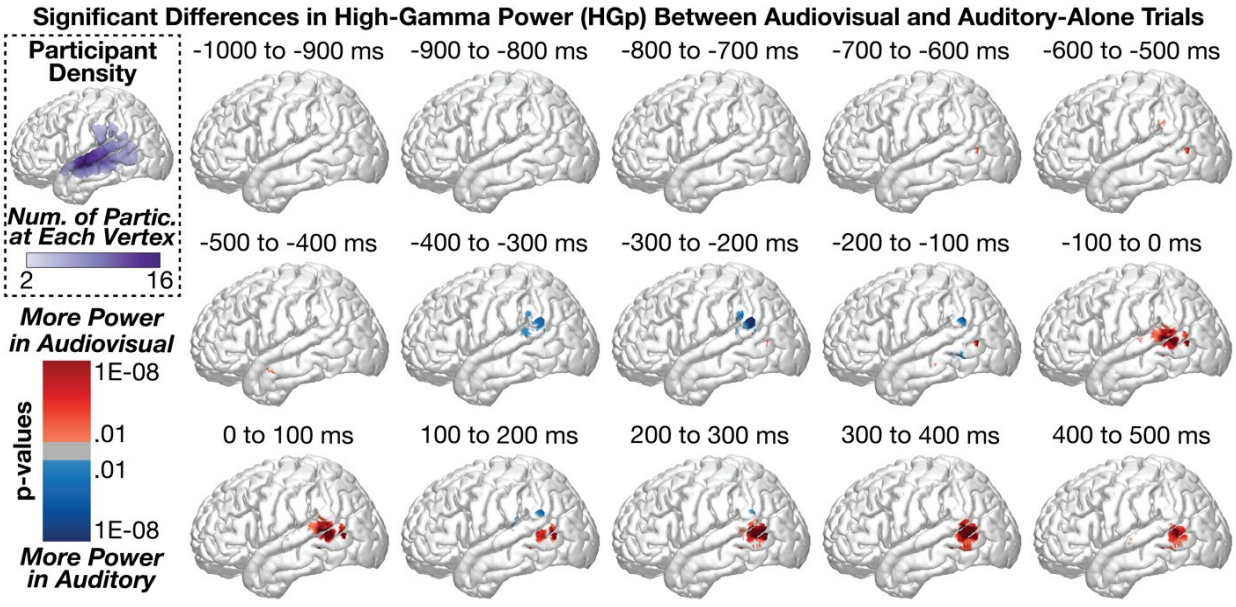


Figure 5. ***Group level plots in high gamma power****. Group-level analyses comparing high gamma power (HGp) between audiovisual and auditory-only conditions at 100 ms time-windows (sound onset a 0 ms). Top-left plot shows the number of participants who contributed data to each vertex. audiovisual stimuli elicited greater power at the posterior STG, peaking beginning before sound onset.*

2.3.5 Group-Level Regional Time-Series Analyses

While the spatial analyses demonstrated significant patterns of activity along the STG, MTG, and SMG, this approach does not effectively allow comparisons across regions or allow the examination of interactions with time and across frequency. To model the influence of visual speech information on spectral power at the group level, we used linear mixed effects models for data aggregated into three regions of the STG (anterior, middle, and posterior regions), consistent with prior studies (Smith et al., 2013). Separate models were constructed at each time point and ROI, and multiple comparison corrections were applied. Importantly, in our estimation of condition effect (auditory-only versus audiovisual), we modelled both random intercepts and slopes (Barr et al., 2013). Table 1 shows the number of electrodes and participants who contributed data to each analysis. Figure 12 shows time-series analyses separated by task variants.

Table 1. ***Number of electrodes and participants present in each of the three regions of STG analyzed****.*

	Anterior STG		**Middle STG**		**Posterior STG**	
Frequency	**N. Elecs**	**N Partic.**	**N. Elecs**	**N Partic.**	**N. Elecs**	**N Partic.**
HGp	22	10	150	18	66	12
Beta	59	14	138	18	62	12
Theta	72	16	162	18	72	12

2.3.6 Group-Level Regional Time-Series Analyses: Theta Power

Regardless of condition, theta power within the STG increased steadily beginning before sound onset and peaking immediately after sound onset, with the strongest activity observed at the posterior STG. Consistent with the spatial analyses, we observed significant differences between audiovisual and auditory-only conditions, with audiovisual trials demonstrating reduced auditory-related theta power (Figure 6). This condition difference was clearest at the posterior

STG, which was significant from -93 to 500 ms (min p = 2.2e-06, peak time = 47 ms), yet also present at the middle STG, which was significant from 108 to 274 ms (min p = 0.030, peak time = 193 ms). No significant differences were observed at the anterior STG after correcting for multiple comparisons. To examine whether visual speech information differentially affected the three STG regions, we conducted a group-level linear mixed-effects model with additional factors of Time and ROI (see Methods for additional information). As expected, the effect of visual information varied as a function of time (Condition x Time interaction: [F(1, 2.1e+06) =

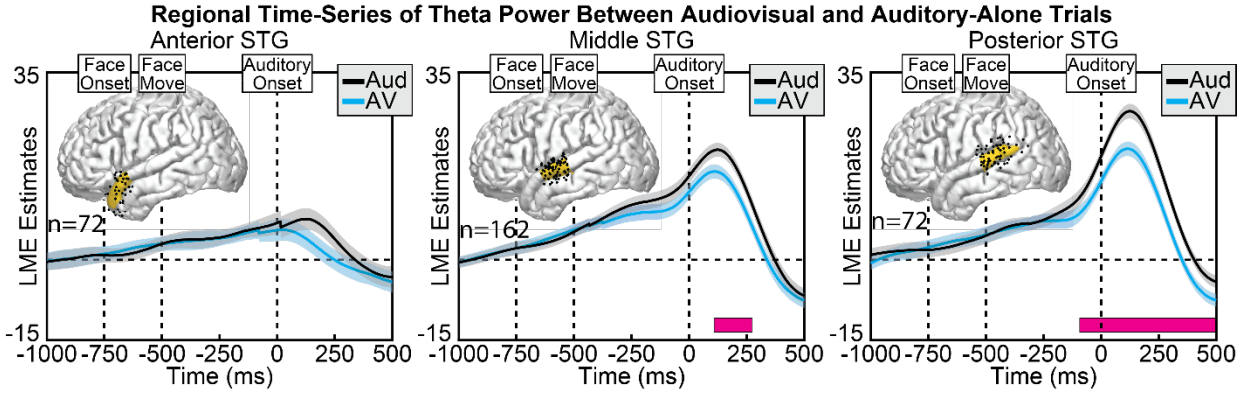


Figure 6. ***Group level LME plots in the theta band****. Group linear mixed-effect model (LME) estimates for each time-point of theta power in auditory-only (black) and audiovisual (blue) trials, calculated separately at anterior (left), middle (middle), and posterior (right) regions of the STG. Shaded areas reflect 95% confidence intervals. Pink boxes reflect significant differences after correcting for multiple comparisons. Corresponding regions are highlighted on the cortical surfaces in yellow with the electrodes that contributed to the analysis shown as black dots (some depth electrodes are located beneath the surface and are not visible). Significant differences in theta power emerged largely after auditory, concentrated along the posterior STG.*

851.6, p = 3.6e-187]), STG region (Condition x ROI interaction: [F(2, 2.1e+06) = 28.7, p = 3.5e-13]) as well as the combination of the two (Condition x Time x ROI interaction: [F(2, 2.1e+06) = 147.5, p = 8.8e-65]). The model with interaction terms additionally demonstrated better fit compared to the same model without interaction terms with a difference in AIC = 0.1e+02. Taken together, these results indicate that visual speech information modulates auditory theta activity predominantly along the posterior STG, following sound onset.

2.3.7 Group-Level Regional Time-Series Analyses: Beta Power

Beta power in the STG showed a combination of power increases and power decreases (beta suppression), with the majority of activity focused on the mid- to posterior-STG. Across conditions, we observed significantly greater beta suppression during the audiovisual condition compared to the auditory-only condition, peaking before sound onset at mid- to anterior STG regions (Figure 7). This condition difference was significant at both the anterior STG, significant from -311 to -195 ms (min p = 0.002, peak time = -247 ms), and the middle STG, significant from -195 to 235 ms (min p = 0.003, peak time = -116 ms), with no significant differences observed at the posterior STG after correcting for multiple comparisons.

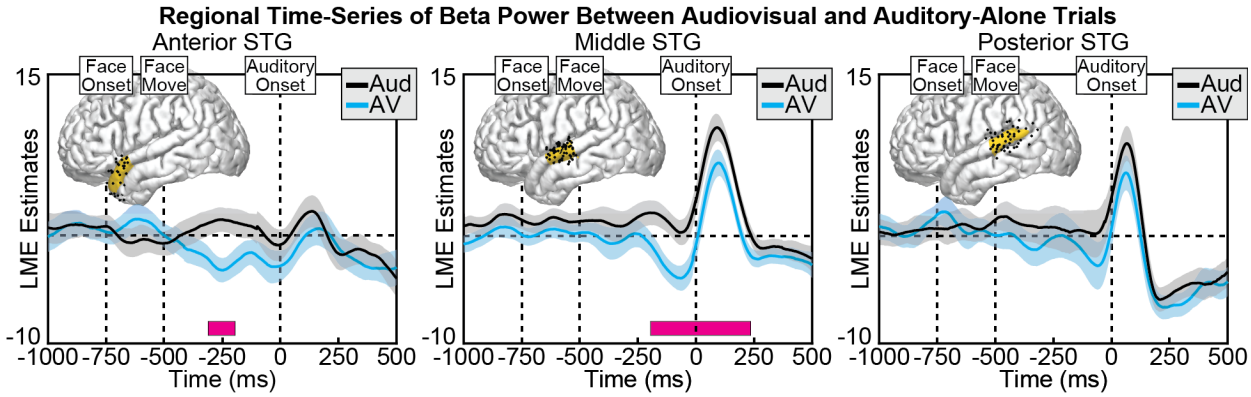


Figure 7. ***Group level LME plots in the beta band****. Group LME model estimates for each time-point of beta power in auditory-only (black) and audiovisual (blue) trials, calculated separately at anterior (left), middle (middle), and posterior (right) regions of the STG. Pink boxes reflect significant differences after correcting for multiple comparisons. Corresponding regions are highlighted on the cortical surfaces in yellow with the electrodes that contributed to the analysis shown as black dots. Significant differences in beta power peaked before sound onset, concentrated in the middle to posterior STG.*

To examine whether visual speech information differentially affected the three STG regions, we conducted a group-level linear mixed-effects model with additional factors of Time and ROI. As expected, the effect of visual information varied as a function of time (Condition x Time interaction: [F(1, 2.0e+06) = 48.5, p = 3.3e-12]), STG region (Condition x ROI interaction: [F(2, 2.0e+06) = 44.2, p = 6.3e-20]) but a non-significant combination of the two (Condition x Time x ROI interaction: [F(2, 2.0e+06) = 2.01, p = 0.134]). The model with interaction terms

nevertheless demonstrated better fit compared to the same model without interaction terms , with difference in AIC = 0.01e+02.

2.3.8 Group-Level Regional Time-Series Analyses: High-Gamma Power

In general, HGp in the STG showed auditory-related power increases that were biased towards the posterior STG. Across conditions, we observed significantly greater HGp in the audiovisual condition compared to the auditory-only condition, occurring before sound onset and localized to the posterior STG (Figure 8). This condition difference was significant only at the posterior STG, from -45 to 24 ms (min $p = 0.028$, peak time = -9 ms). No other significant differences were observed. To examine whether visual speech information differentially affected the three STG regions we conducted a group-level linear mixed-effects model with additional factors of Time and ROI. As expected, the effect of visual information varied as a function of time (Condition x Time interaction: [$F(1, 1.8e+06) = 86.7$, $p = 1.3e-20$]), STG region (Condition x ROI interaction: [$F(2, 1.8e+06) = 29.6$, $p = 1.3e-13$]) as well as the combination of the two (Condition x Time x ROI interaction: [$F(2, 1.8e+06) = 20.0$, $p = 2.1e-09$]). The model with interaction terms additionally demonstrated better fit compared to the same model without interaction terms, with difference in AIC = 0.03e+02.

2.3.9 Group-Level Regional Time-Series Analyses: Interactions Across Frequencies

Analyses conducted separately at each of the frequency bands demonstrated audiovisual effects in putatively distinct time ranges and spatial distributions. However, to test the claim that the spatial and temporal patterns observed across the frequency bands are indeed distinct, it is necessary to model frequency band and time-points in relation to task conditions.

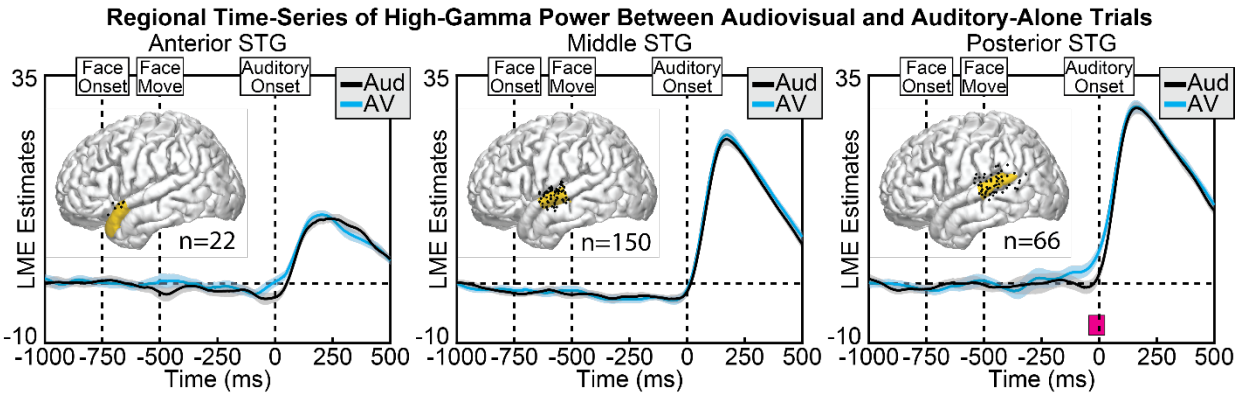


Figure 8. ***Group level LME plots in high gamma power****. Group LME model estimates for each time-point of HGp in auditory-only (black) and audiovisual (blue) trials, calculated separately at anterior (left), middle (middle), and posterior (right) regions of the STG. Pink boxes reflect significant differences after correcting for multiple comparisons. Corresponding regions are highlighted on the cortical surfaces in yellow with the electrodes that contributed to the analysis shown as black dots. Significant differences in HGp peaked before sound onset in the posterior STG.*

To this end, we constructed a group-level linear mixed-effects model that included fixed effects of task condition, frequency band, region of interest along the STG, and time, modeling both random intercepts and random slopes for trial condition. Including all frequency bands in the model yielded significant interactions of Condition x Frequency Band [$F(2, 2.9e+07) = 277.4$, $p = 2.3e-64$], Condition x Frequency Band x ROI [$F(4, 2.9e+07) = 72.7$, $p = 4.8e-43$], Condition x Frequency Band x Time [$F(2, 2.9e+07) = 2254.0$, $p = 1.2e-294$], and Condition x Frequency Band x ROI x Time [$F(4, 2.9e+07) = 397.8$, $p = 3.7e-163$]. Consistent with these significant interactions, the addition of each parameter improved model fit based on AIC. Repeating this analysis with only low-frequency signals associated with neural oscillations (theta and beta) yielded the same pattern, with significant interactions of Condition x Frequency Band [$F(2, 2.0e+07) = 346.7$, $p = 2.2e-40$], Condition x Frequency Band x ROI [$F(2, 2.0e+07) = 43.2$, $p = 1.9e-15$], Condition x Frequency Band x Time [$F(1, 2.0e+07) = 1645.2$, $p = 2.5e-121$], and Condition x Frequency Band x ROI x Time [$F(2, 2.0e+07) = 357.6$, $p = 9.6e-78$]. Again, the addition of each parameter improved model fit based on AIC. Taken together, these data

demonstrate that visual speech information evokes distinct temporal and spatial patterns through theta, beta, and HGp.

2.3.10 Individual Differences in Neural Activity

While the linear mixed-effects models demonstrate effects that are present at the group level, it is important to note that highly significant condition differences that deviated from these group patterns were observed at individual electrodes in individual participants. In particular, HGp results showed greater variability across electrodes and participants than did theta and beta bands. For example, while the most consistently observed response was increased activity before sound onset in posterior regions of the STG, this was not present in all participants or all electrodes. Figure 9 shows pairs of individual electrode responses from 5 participants, with the top row highlighting one STG electrode from that participant that matches the pattern observed at the group level, and the bottom row highlighting a second STG electrode demonstrating a different (sometimes opposite) pattern. Indeed, Participant 9 (first column) showed the opposite pattern across two electrodes, with the lower row demonstrating more HGp for auditory trials before sound onset. Of note, many of the electrodes showed significantly reduced HGp to audiovisual versus auditory-only stimuli during sound processing (100 - 200 ms), as reported previously (Karas et al., 2019). While this pattern was demonstrated in many electrodes and participants, the anatomical region varied throughout the STG and the overall pattern did not reach significance at the group level.

2.3.11 Predictability of distinct time-ranges across frequency bands.

While the group-level spatial and regional time series analyses demonstrated the audiovisual effects at individual frequency bands, the effects contributed by individual electrodes

are not interpretable in these results. To study the variability across electrodes and the predictability of distinct time-ranges, we constructed pair-wise linear mixed effect models for every frequency band between different stimuli onsets in the anterior, middle and posterior STG regions.

The mixed-effect models were constructed in three separate pair-wise analyses across distinct time-ranges; 1) predictability of post-auditory onset from pre-auditory onset 2) predictability of pre-auditory onset in any given frequency band on pre-auditory onset in other frequency bands 3) predictability of post-auditory onset in any given frequency band on post-auditory onset in other frequency bands.

To construct the linear-mixed effect models, activity in each frequency band was averaged across the time ranges of interest to capture the observed audiovisual effects.

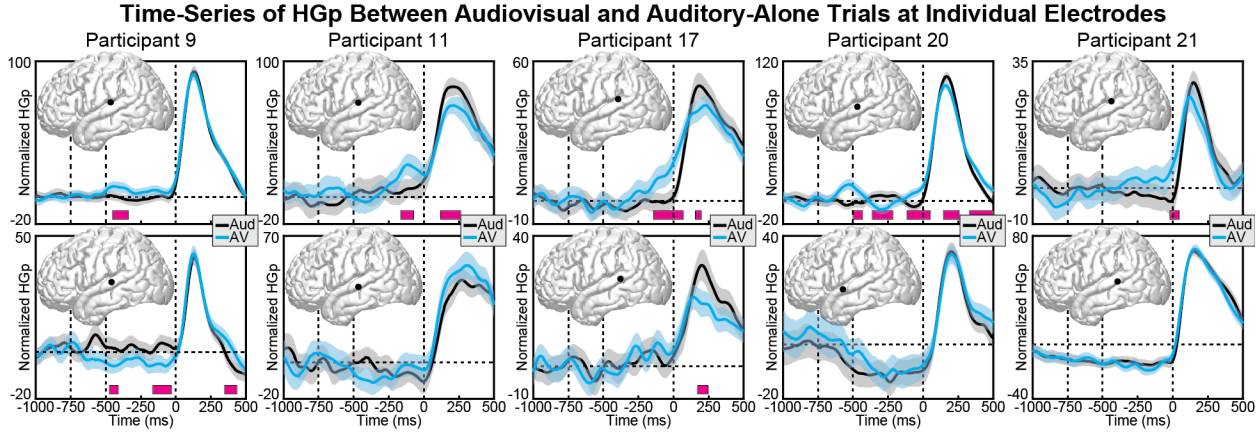


Figure 9. ***Individual participant high gamma power activity****. Individual participant HGp activity at audiovisual (blue) and auditory-only (black) conditions. Each column displays data from a different participant (two electrodes per participant). Top row displays electrodes that showed the same pattern of HGp results observed at the group-level, with increased activity in the audiovisual condition starting before sound onset. Bottom row shows a proximal electrode that demonstrated a different (sometimes conflicting) pattern. Shaded areas reflect 95% confidence intervals. Pink boxes reflect significant differences after correcting for multiple comparisons.*

2.3.12 Predictability of distinct time-ranges across frequency bands: Theta power

From the group-level spatial analyses and regional time-series analysis, we had demonstrated that the effects of individual frequency bands varied distinctly over the time-period

of our stimuli. This effect can be seen in the predictability of pre-auditory onset theta band on post-auditory onset theta band across the anterior, middle and posterior STG, confirming the findings from our group-level spatial analysis and the regional time series analysis. From supplemental figure 3, we see that this effect is highly significant at $p < 0.001$ in all three regions. Supplemental figure 4 shows that the theta power in pre-auditory onset time period is significantly predictive of beta power in the post-auditory onset time period at $p < 0.005$. Post-auditory onset theta power is not shown to be significantly predictive of any other frequency bands.

2.3.13 Predictability of distinct time-ranges across frequency bands: Beta power

Pre-auditory time period in the beta power was shown to be highly predictive of beta power in the post-auditory time period at $p < 0.001$ across the anterior, middle and posterior STG. Beta power in the pre and post-auditory time periods was not seen to be predictive of any other frequency bands.

2.3.14 Predictability of distinct time-ranges across frequency bands: High-Gamma Power

The predictability of pre-auditory onset on post-auditory onset across individual electrodes in the high-gamma power band was seen in the anterior, middle and posterior STG, at $p < 0.001$ in all three regions (supplemental figure 3).We also see that high-gamma power in pre-auditory onset time period positively predicts post-auditory onset theta band in the anterior STG and negatively predicts post-auditory onset theta band in the middle STG. Supplemental figure 5A, shows that high-gamma power in the pre-auditory onset time period negatively predicts pre-auditory onset theta band in the middle STG. Finally, from supplemental figure 5B, we see that

high-gamma power in the post-auditory onset time period negatively predicts post-auditory onset theta band.

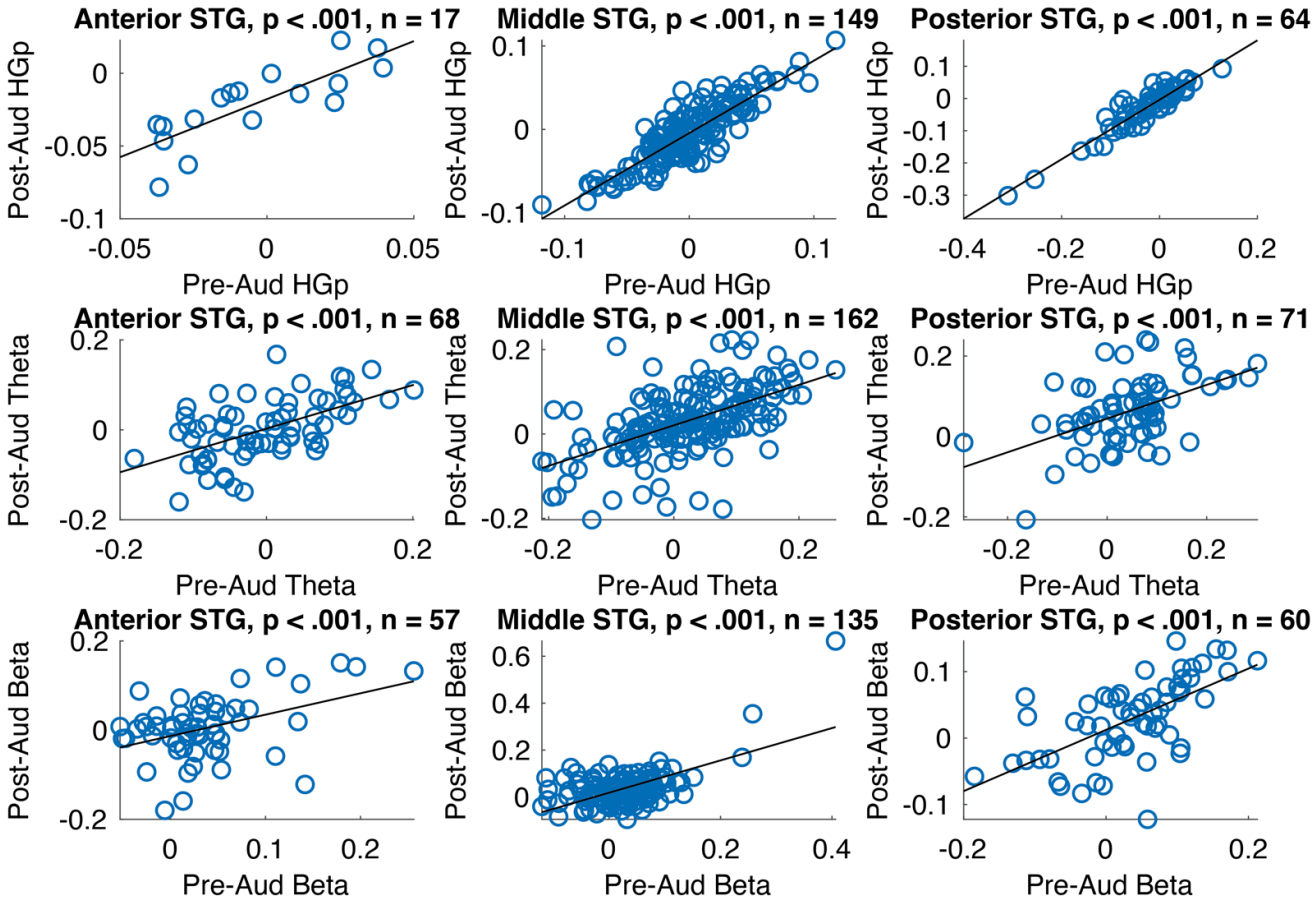


Figure 10. ***Predictability of post-stimuli activity from pre-stimuli activity****. Scatterplots showing the magnitude of audiovisual effects (auditory-only minus audiovisual) for the same frequency band, before (x-axis) or after (y-axis) auditory-onset. Columns reflect anatomical electrodes localized to the anterior STG (left), middle STG (middle), and posterior STG (right). Rows reflect separate frequency bands. Activity in each frequency band was averaged across time ranges to capture observed audiovisual effects (see methods). Subplot titles show linear mixed-effect modeled p-values (uncorrected) and the number of electrodes included. All 9 subplots showed significant positive relationships between audiovisual effects before and after auditory-onset.*

2.4 Discussion

Visual signals are known to affect auditory speech processes in multiple ways. For example, lipreading signals provide high-level phonemic representations (Bourguignon et al., 2020), visual motion information can relay timing information (McGrath et al., 1985), lip closure facilitates the parsing of word-boundaries and speech rate (Chandrasekaran et al., 2009), lip-shape provides spectral information (Plass et al., 2020), and speaker identity can further enhance spatial localization and multisensory binding (Vatakis and Spence, 2007; Brang 2019). Indeed, a persistent challenge in identifying the various effects of audiovisual speech information has been largely methodological in nature. fMRI studies lack the temporal resolution to identify whether visual speech modulates auditory regions before, simultaneously with, or after the onset of auditory speech. On the other hand, iEEG studies face two critical shortcomings: (1) Past studies investigating audiovisual speech integration have analyzed data using single-participant designs or traditional parametric statistics making it hard to generalize the findings to the group-level and thus to the general population (Micheli et al., 2020; Besle et al., 2008; Plass et al., 2020). (2) Even while using variants of group-level analysis such as linear mixed-effects modeling, previous studies (Ozker et al., 2017; Ozker et al., 2018) have focused on HGp, which indexes local population firing rates, ignoring low-frequency oscillations which potentially reflect distinct audiovisual information.

To test for the presence of separate but concurrent visual processes in auditory areas, we measured neural activity using intracranially implanted electrodes in a large number of human participants (n = 21) during an audiovisual speech perception paradigm. These data demonstrated that at least three distinct patterns of activity occur in the STG during audiovisual speech perception relative to unimodal auditory speech perception. (1) For the theta band, visual speech

suppressed the auditory response predominantly in the posterior STG from before auditory speech onset to well after auditory speech onset (-93 ms to 500 ms, peak time = 47 ms). (2) For the beta band, suppression was seen in the anterior STG from -311 to -195 ms before auditory speech onset (peak time = -247 ms) and in the middle STG from -195 ms to 235 ms after speech onset (peak time = -116 ms). (3) For high gamma, suppression was seen from -45 ms to 24 ms only in the posterior STG (peak time = -9 ms). We interpret these distinct patterns to reflect distinct neural processing in auditory regions, potentially responsible for encoding different types of visual information to aid in auditory speech perception. Of note, filtered spectral power produces temporal smoothing of the data (e.g., one cycle of theta band activity is ~200 ms in duration) which reduces the precision of the reported time ranges.

Converging behavioral and neurophysiological evidence suggests that audiovisual enhancements from audiovisual speech (e.g., better detection and faster reaction times) and visual recovery of phoneme information are subserved by two distinct mechanisms (Eskelund et al., 2011; Plass et al., 2014). This distinction may reflect a neural dissociation between predictive multisensory interactions that optimize feedforward encoding of auditory information and later feedback processes that alter auditory representations generated in the pSTS (Arnal et al., 2009; Arnal et al., 2011) and the posterior STG (Reale et al., 2007). In support of this view, both visual speech (Besle et al., 2004; Arnal et al., 2009; Van Wassenhove et al., 2005) and other anticipatory visual cues (Vroomen and Stekelenburg, 2010) can speed-up and reduce the magnitude of early physiological responses associated with auditory feedforward processing, potentially reflecting optimization of auditory encoding in accordance with temporal or acoustic constraints imposed by visual information. These early feedforward effects, which are insensitive to audiovisual congruity in speech, are temporally, spatially, and spectrally distinct from later

(>300 ms) responses that are specific to crossmodally incongruent speech (Arnal et al., 2011; Van Wassenhove et al., 2005). These later incongruity-specific interactions point to a hierarchical feedback regime in which unisensory speech processing is altered in accordance with integrated audiovisual information from the pSTS (Olasagasti et al., 2015; Kayser and Logothetis, 2009) and general speech perception areas in the STG (Mesgarani et al., 2014). These data are consistent with this dissociation, with several temporally and spatially discrete neural responses in the STG. It should also be noted that some of these activation patterns may be due to non-specific effects (e.g., elevated attention or physiological arousal to viewing a face).

Our observation of a dissociation among theta and beta frequency ranges is consistent with prior EEG and physiology research suggesting these mechanisms encode different information about a visual signal (e.g., Kumar et al., 2016; Wang et al., 2017). Theta activity effectively captures ongoing auditory timing information, including rhythmic events (e.g., Schroeder et al., 2009). Conversely, beta band activity has been more strongly associated with feedback signals that may predictively encode visual information in the auditory system prior to sound onset (e.g., Engel et al., 2010). The dissociation between theta and HGp observed is particularly interesting as HGp signals have also been implicated in a predictive coding framework, such that ensembles of neurons in the posterior STG initially activate neuronal ensembles before sound onset, leading to refined population tuning and thus less HGp following sound onset (Karas et al., 2019). While this reduction in HGp during audiovisual trials was observed in many participants (see Figure 9), it was not observed at the group level, potentially due to anatomical variability in the location of the response or due to heterogeneity across participants.

Research on the neural source of visual signals relayed to the auditory system have largely focused on the left posterior temporal sulcus (pSTS). This region demonstrates strong differences between auditory-only and audiovisual stimuli in both fMRI and iEEG research (Beauchamp et al., 2004b; Ozer et al., 2017; Ozker et al., 2018; Okada et al., 2013), and has potential causal roles in audiovisual speech integration as revealed by lesion mapping (Hickok et al., 2018) and inhibitory transcranial magnetic stimulation (Beauchamp et al., 2010). While these data indicate that some of the information observed in the present study was likely projected through feedback pathways originating in the pSTS, particularly given its role as a center for bottom-up prediction errors in language comprehension (Lewis and Bastiaansen, 2015), it is possible that each distinct temporal/spatial pattern has a unique corresponding source. While the present study does not provide evidence as to what information is encoded within each spatial/temporal pattern, we suggest that future research using causal measures or neural decoding identify the specific visual dimensions represented.

An important point of note here is that all of the analysis were performed individually for each subject, with the results then transformed to the MNI space. We performed our analysis in this manner since there is evidence to show that anatomical boundaries for individual regions of the brain including the location of the auditory cortex, and more specifically the boundaries of Heschl's gyrus varies across individuals (Leonard et al., 1998; Keuken et al., 2014). Hence, to generalize our findings across the entire sample, analysis were performed at an individual level followed by projection to the MNI space.

In summary, this study demonstrates that audiovisual speech integration elicits multiple distinct patterns of neural activity within the STG and adjacent cortex, occurring across separate frequencies and temporal/spatial distributions. These data suggest that visual modulation of

auditory speech perception utilizes multiple mechanisms, potentially reflecting independent sources of information. Our results are also consistent a hybrid family of integration models as proposed by Peelle and Sommers (2015). Finally, this study additionally shows the advantage of group-level analyses of iEEG data using linear mixed-effect models, which can improve statistical validity and power, and importantly, improve generalization of results across patients and to the population at large.

2.5 Supplementary Figures

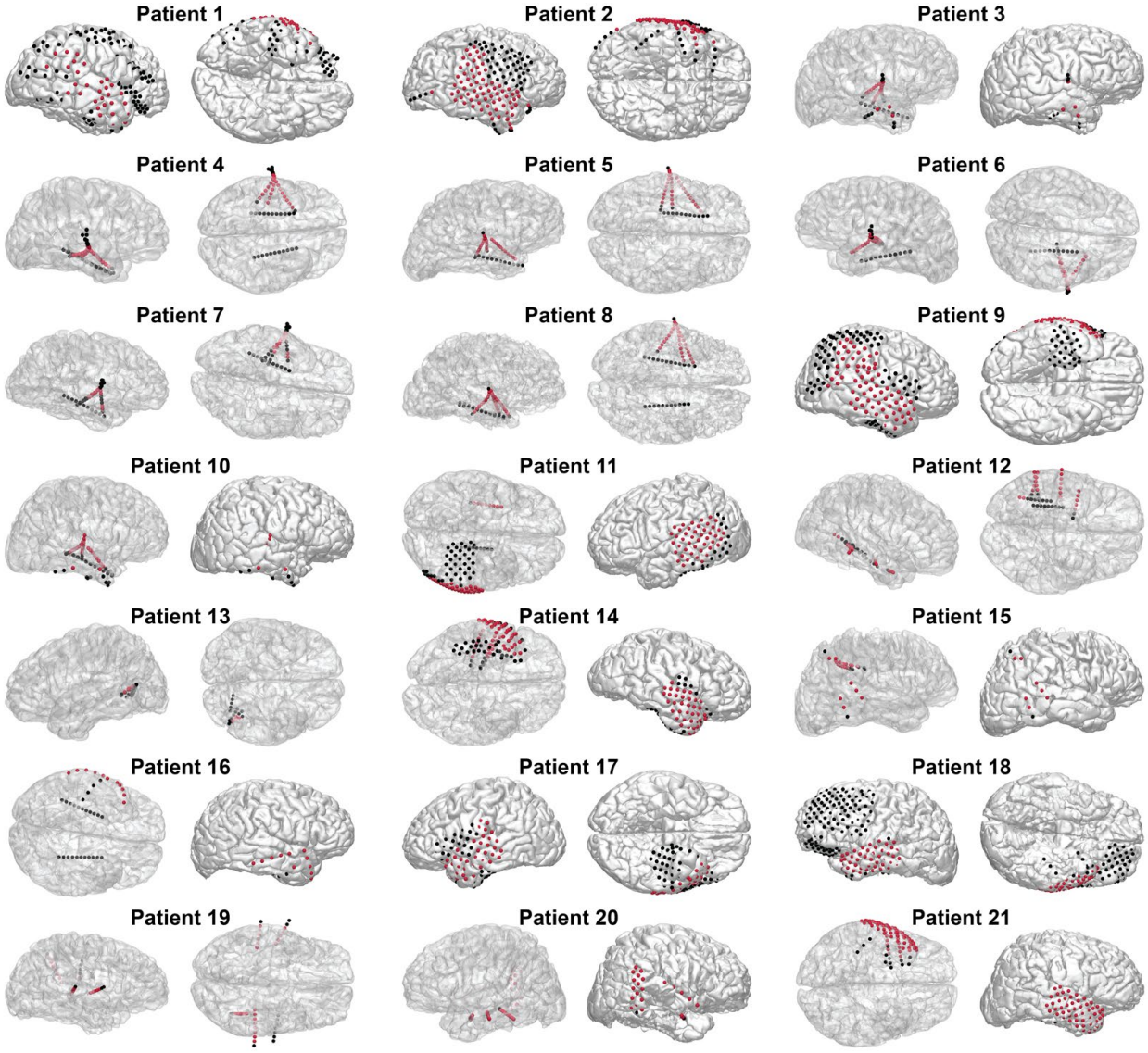


Figure 11. ***Individual electrode maps for each patient.*** *Red spheres reflect auditory electrodes that met anatomical criteria and that were not rejected during pre-processing for having excessive noise. Black spheres show remaining implanted electrodes that were not included in analyses.*

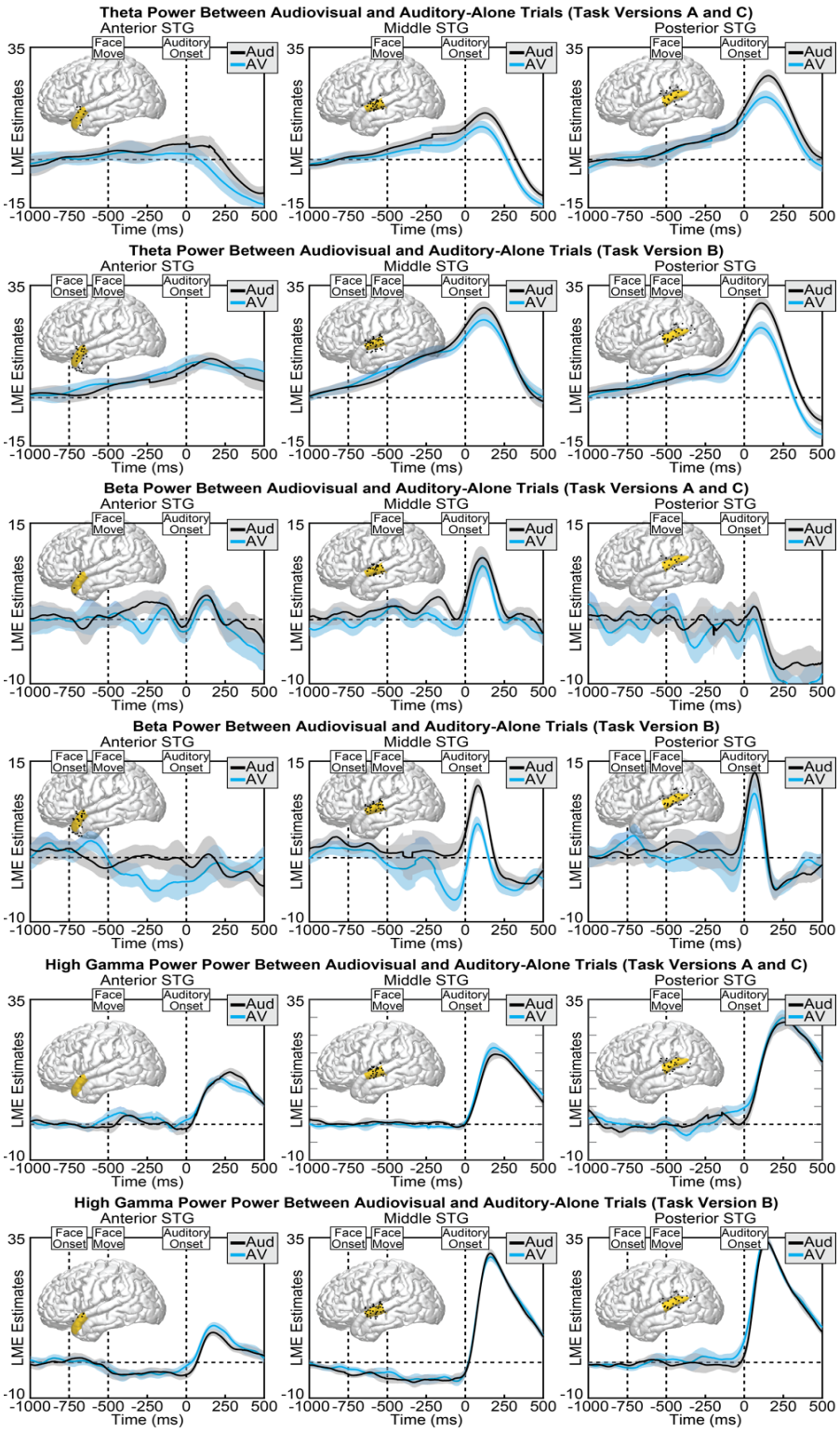


Figure 12. ***Individual participant activity in the AV and auditory conditions****. Individual participant activity at audiovisual (blue) and auditory-only (black) conditions separated by task variants and frequency bands. Task Variants A and C presented participants the moving face stimulus at 500 ms before auditory onset, whereas Task Variant B showed a static face beginning 750 ms before auditory onset (with motion starting at 500 ms before auditory onset). No clear differences across the task differences are present in the time series data.*

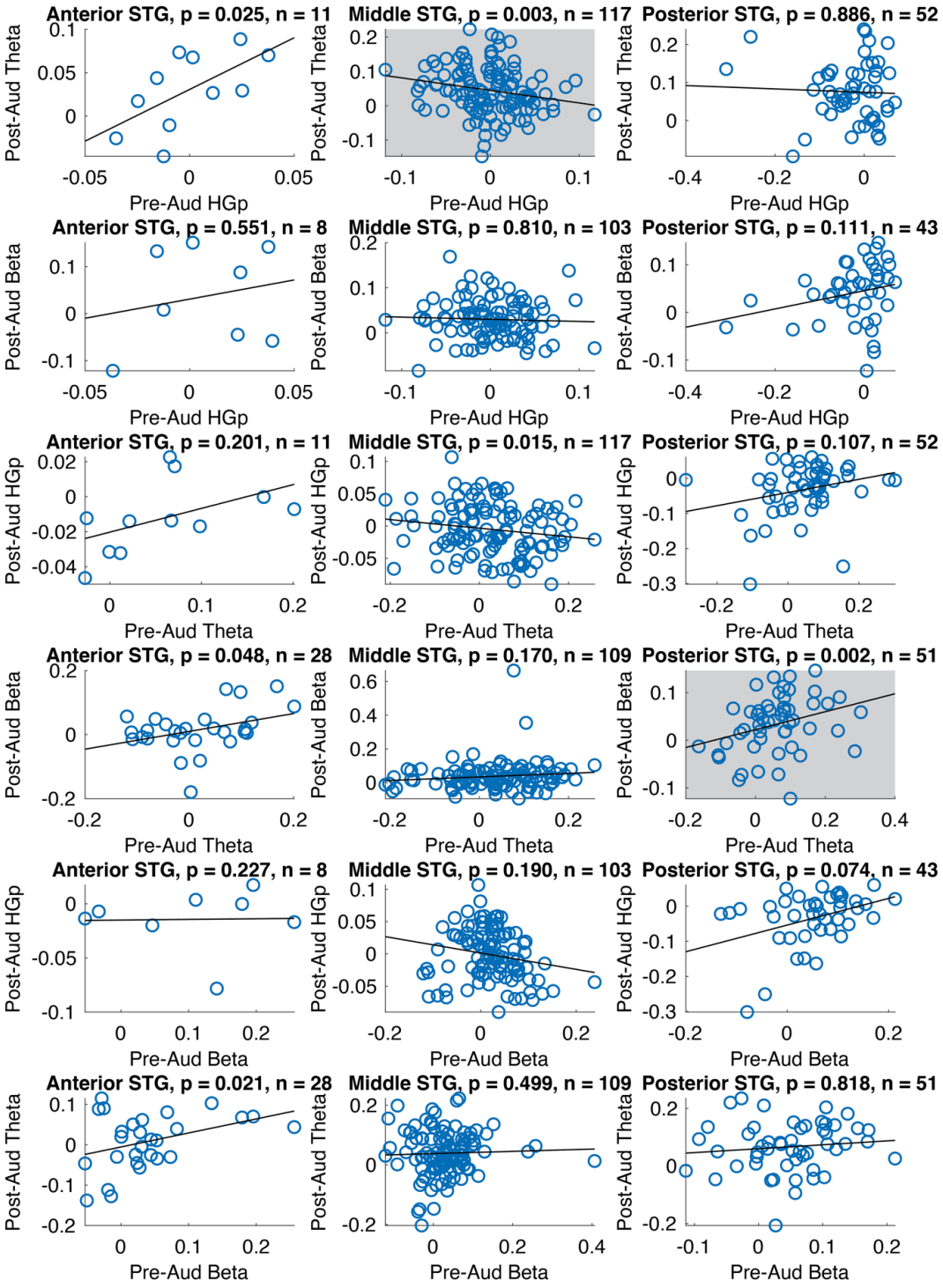


Figure 13. ***Inter-frequency post and pre-stimuli predictability****. Scatterplots showing the magnitude of audiovisual effects (auditory-only minus audiovisual) across pairs of frequency bands, before (x-axis) or after (y-axis) auditory-onset. Columns reflect anatomical electrodes localized to the anterior STG (left), middle STG (middle), and posterior STG (right). Rows reflect separate frequency band pairs. Activity in each frequency band was averaged across time ranges to capture observed audiovisual effects (see methods). Subplot titles show linear mixed-effect modeled p-values (uncorrected) and the number of electrodes included. Gray-shaded scatterplots highlight p<.01 significant relationships.*

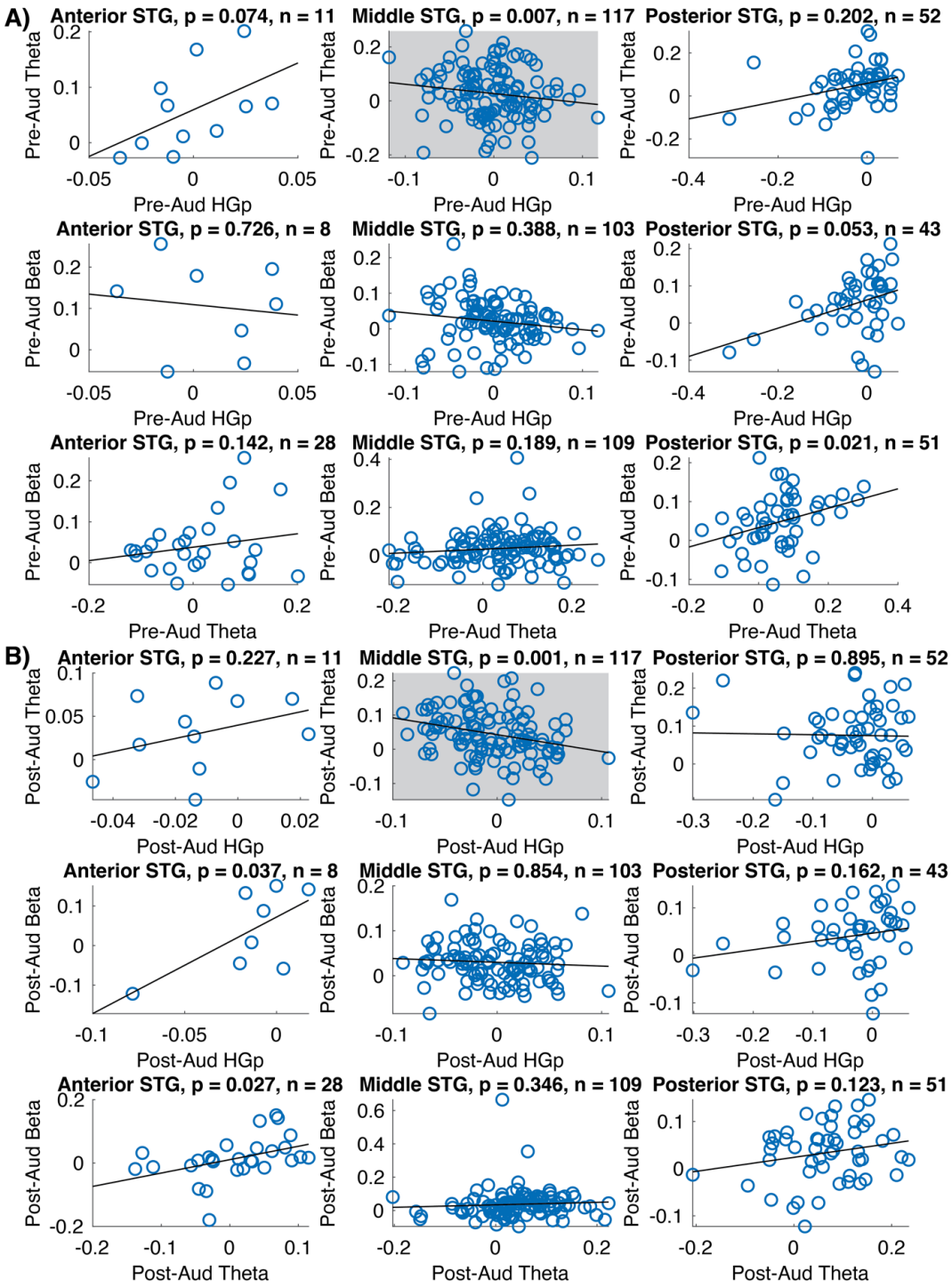


Figure 14. ***Inter-frequency post stimuli predictability****. Scatterplots showing the magnitude of audiovisual effects (auditory-only minus audiovisual) across pairs of frequency bands in the same time windows, either before (A) or after (B) auditory onset. Columns reflect anatomical electrodes localized to the anterior STG (left), middle STG (middle), and posterior STG (right). Rows reflect separate frequency band pairs. Activity in each frequency band was averaged across time ranges to capture observed audiovisual effects (see methods). Subplot titles show linear mixed-effect modeled p-values (uncorrected) and the number of electrodes included. Gray-shaded scatterplots highlight p<.01 significant relationships.*

Chapter 3 Phonemic Representations Encoded in Auditory Cortex During Visual Speech

The previous chapter discussed the various ways in which visual speech modulates auditory speech perception in auditory cortex. However, it is unclear if these modulations reflect meaningful phonemic information extracted from visual speech. In this chapter, I provide evidence that supports the hypothesis that visual speech encodes the identities of visemes in primary auditory areas during auditory-visual speech.

3.1 Introduction

Visual speech improves auditory speech perception (Plass et al., 2020, Micheli et al., 2020) during face-to-face conversations. Behavioral research has demonstrated robust benefits of visual speech across various special populations such as older individuals with age-related hearing decline (Rosemann & Thiel., 2018), hearing impairments (Lee et al., 2007), and cochlear implants (Blackburn et al., 2019). Visual speech also increases speech intelligibility for normative individuals in noisy environments (Sumby & Pollack., 1954, Lusk & Mitchel, 2016). However, there remains limited understanding of how the brain enables these benefits. One brain area of interest in examining the modulatory effects of visual speech on auditory speech processing includes auditory areas such as the superior temporal gyrus and sulcus (STG/STS) because of their role in processing auditory speech (Beauchamp et al., 2004a, 2010, Karthik et al., 2021, Arnal et al., 2009, 2011, Reale et al., 2007).

A recent study by our group has shown that activity in the STG is modulated by visual speech, suggesting that vision provides multiple types of information to the auditory system (Karthik et al., 2021). Specifically, we showed visual influences in multiple temporal, spectral and spatial configurations, indicating that vision has multiple distinct effects on auditory processing. However, it is largely unknown what information is encoded in these visual-to-auditory activations and which oscillatory frequencies and regions of the STG encode different visual information. To date, research has generally focused on the ability for visual motion information to bias auditory timing, such that lip closure is associated with the boundaries between words and that pre-articulatory speech can predict speech onset (Schroeder et al., 2008). However, behavioral research has additionally shown that vision can enhance auditory processing through general increases in attention and arousal, and by extracting useful statistics from the visual signals, including lip shapes that predict relative pitch, lipreading, and speaker identity (Chandrasekaran et al., 2009; Chen & Rao, 1998; Erber, 1975; Van Wassenhove et al., 2005).

The notion that lipreading can bias what is heard is perhaps the best studied aspect of auditory-visual speech perception, providing a clear candidate for one type of information that would be expected to be relayed to the auditory system. But to date, no direct evidence exists that lipreading information is represented in the auditory system. However, recent research in the auditory domain may shed light on how lipreading signals are transformed into auditory speech units. Specifically, Mesgarani and colleagues (Mesgarani et al., 2014) used human intracranial electroencephalography (iEEG) recorded from high-density electrodes to demonstrate that phonemes (basic units of speech sounds) are represented by distributed populations of neurons in the STG. Combined with past research, these data support a model in which the STG contains a

patchy distribution of neurons that are tuned to specific phonemes via their spectro-temporal profiles (Formisano et al., 2008; Mesgarani et al., 2014). For example, research has demonstrated spatially distinct responses in these regions to spectrally similar phonemes such as /ba/ and /da/ (Chang et al., 2010; Formisano et al., 2008; Raizada et al., 2010), and clustered activities across a large phoneme-space (e.g., the distributed pattern of activity to /ma/ is more similar to /na/ than is it to the spectro-temporally distinct phoneme /ba/ (Mesgarani et al., 2014)). Indeed, the identity of a heard phoneme can be decoded by the distribution of activity in the auditory cortex. (Leonard et al., 2016), even when the physical auditory stimulus remains the same.

In contrast, we do not have the same detailed understanding of audio-visual speech processes, or how they integrate with phoneme-tuned neuronal populations. Using iEEG we recently demonstrated that visual speech elicits multiple independent changes throughout the STG relative to auditory-only speech (Karthik et al., 2021). Varying across theta, beta, and high-gamma bands (HGp), we observed changes before sound onset due to the presence of a moving face, and separate modulations only during speech processing. Relatedly, functional magnetic resonance imaging (fMRI) studies have demonstrated broad activation of the STG during silent lipreading (Calvert et al., 1997). While these findings highlight the broad effect of visual information on auditory speech processing, differences in activity do not provide a mechanistic account for how visual speech signals are integrated with auditory neuronal populations.

The extraction high-level phoneme information from “visemes” (categorically encoded visual articulations analogous to auditory phonemes) during speechreading (lipreading) is one key component of crossmodal speech perception (Beauchamp et al., 2010; A. Nath & M. S. Beauchamp, 2011; Nath et al., 2011). Importantly, however, the computational mechanisms through which viseme information is transformed into phoneme representations have remained

theoretical. One candidate model, suggested by us and others (Beauchamp et al., 2010; Karthik et al., 2021), proposes that visemic signals modulate responses in neurons that are maximally sensitive to specific phonemes. Beauchamp and colleagues (Beauchamp et al., 2010) have suggested that this process occurs through a 'winner-take-all' mechanism in the pSTS, in which auditory and visual signals each cast votes for the 'heard' phoneme, and that the neural population with the highest activation profile produces the phoneme that is perceived. We agree with the computational properties of this model but believe it may also occur within the STG, consistent with recent evidence in the auditory domain, in which illusory shifts during the phoneme restoration effect were successfully modeled by a categorical shift in phoneme-tuned responses (Leonard et al., 2016).

Here we test this theory by using iEEG recordings in patients with epilepsy during a word perception task, in which participants either saw the lip movements or heard the speech sounds for the same groups of words, each beginning with one of four consonants (*'b', 'd', 'g', 'f'*). iEEG signals were then spatially and temporally classified as belonging to one of these four consonant groups using SVMs separately for auditory-only and visual-alone conditions. Results showed that phoneme and viseme identities were successfully classified from activity in the STG, supporting the proposed model of auditory-visual speech integration. Additionally, representational similarity analysis revealed that matching phonemes and visemes encode similar representations in the STG.

3.2 Materials and methods

3.2.1 Participants, implants and recordings

4 participants undergoing clinical evaluation using iEEG for intractable epilepsy consented to participate in this study under an institutional review board (IRB) approved

protocol at the University of Michigan or Henry Ford hospital. Clinically implanted depth electrodes (5 mm center-to-center spacing) and/or subdural electrodes (10 mm center-to-center spacing) were used to acquire iEEG data from participants. Data from a total of 210 electrodes were recorded from 4 participants. The type of electrodes implanted and locations were based on the clinical needs of the participants. iEEG recordings were acquired at either 4096 Hz (n = 3 participants) or 1000 Hz (n = 1 participant) due to differences in clinical amplifiers.

3.2.2 MRI and CT acquisition and processing

Preoperative T1-weighted magnetic resonance imaging (MRI) and postoperative computer tomography (CT) scans were acquired for all participants. The preoperative T1 MRI was registered to the postoperative CT using SPM12 using the 'mutual information' method (Viola & Wells 1997). The CT was not resliced or resampled. The localization of each electrode was performed using custom software (Brang et al., 2016). The algorithm works by identifying and segmenting electrodes from the CT image based on gray scale intensity, and projects subdural electrodes to the dura surface using the shape of the electrode disk to counteract postoperative compression. For all subsequent analyses including reconstruction of cortical surfaces, volume segmentation and anatomical labelling, the Freesurfer image analysis suite was utilized (http://surfer.nmr.mgh.harvard.edu/; Dale et al., 1999; Fischl et al., 1999).

3.2.3 Task and stimuli

4 participants were tested in the hospital (University of Michigan, n = 3 or Henry Ford, n = 1) using a laptop running Psychtoolbox (Kleiner et al., 2007) at their bedside. The task paradigm was adapted from a prior study (Ross et al., 2007) which was designed to behaviorally study multiple aspects of auditory-visual speech integration. The stimuli consisted of a female

speaker who produced 40 commonly used 1-2 syllable words that each started with one of the four consonants: *'b', 'f', 'g', 'd'* (10 of each). The phoneme in the second position of each of these words was generally balanced across each of the four groups. Each stimulus was recorded at a frame rate of 29.97 frames per second, and trimmed to 1100 ms in length. Further adjustments were made such that the first consonantal burst of each word occurred at 500 ms during the video playback by removing leading video frames.

Each participant underwent two task variants using the same stimuli and task design to increase trial numbers, to reduce classifier overfitting. Figure 15 shows the task schematic for both variants of the task. In variant one, participants were presented with words one at a time, in one of two main conditions: auditory-only or visual-alone. Participants then identified the initial speech sound of the presented stimulus using a button press to select one of four options shown on the computer screen. For example, on a trial with the word *"bag"*, the options presented to the participant were *'b', 'g', 'd', 'th'*. The paradigm included 40 trials per consonant in each main condition, such that each of the 40 words were presented 4 times in the visual-alone condition and another 4 times in the auditory-only condition. This resulted in a total of 320 trials for each participant using task variant 1. The words used in our task are presented in Table 2.

In task variant 2, participants were presented with trials in one of four main conditions: auditory-only, visual-alone, congruent audiovisual, or incongruent audiovisual. Task stimuli and instructions were the same as in variant 1. Variant 2 included 20 trials per consonant in each main condition. A second factor that was manipulated in this variant was the background noise level of the stimuli such that half of the words used in each condition were presented in either a low noise or a high noise context. In the low noise context, the auditory stimuli were presented as they were recorded (SNR = 32.2 dB SPL). In the high noise context, pink noise was added to

reduce the signal-to-noise (SNR) ratio of the signals to -6 db SPL. In this task variant, only data from the auditory-only and visual-only conditions were included in analyses because they matched the main conditions obtained from Task variant 1. This resulted in a total of 80 auditory-only and 80 visual-alone trials for each participant using task variant 2.

A total of 480 trials (Task variant 1: 320 trials, task variant 2: 160 trials) with 60 trials for each consonant (*'b', 'g', 'd', 'f'*) per condition was obtained from the combined data of both task variants.

Table 2 ***List of words used in the task. 10 words were used for each consonant:*** *'b', 'd', 'g', 'f'*

bag	base	bill	bible	bid	bank	beer	bias	beard	bye
fad	fat	fill	fine	fist	fast	file	fish	fit	five
gag	gas	guild	guile	gig	gang	gear	guise	gill	guy
dad	dash	dill	dine	disk	dance	dial	dish	digs	dive

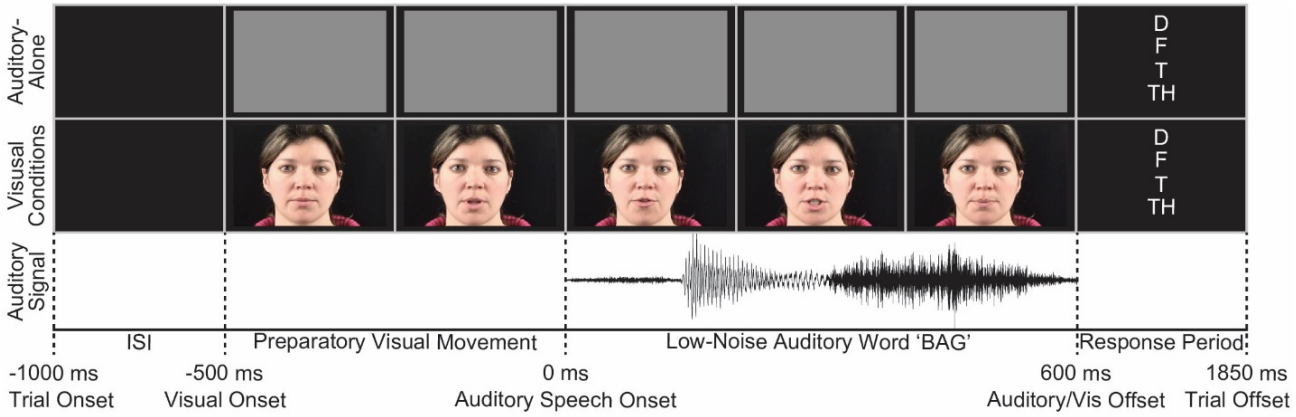


Figure 15. ***Schematic of the task****. Task schematic for the stimuli used. All trials had an initial fixation period with a blank screen. In the auditory-only condition, a gray rectangle appeared on screen at 500 ms before onset of phonemes, which occurred at 0 ms. In the visual-only condition, following an initial fixation period, a face appeared on screen at 500 ms before the time when speech sounds would naturally begin. Stimuli offsets for both the auditory-only and visual-only conditions occurred at 600 ms after phoneme onset times. This was followed by a response window of 1250 ms in both conditions.*

Each participant received a randomized trial order. For the auditory-only condition, a grey rectangle was presented 500 ms before sound onset. Stimuli offset occurred 600 ms after sound onset time. In the visual-only condition, face onset occurred 500 ms before the time when

phoneme onset would naturally occur. A wait time of 1.25 seconds was provided for the participants to respond to each of the stimuli.

3.2.4 iEEG data preprocessing

Data were preprocessed using bipolar referencing, such that signals from adjacent electrodes were subtracted in a pairwise manner. This ensured that the final signals of interest were obtained from neuronal populations that provided maximal localized responses (Yao et al., 2019). The analyses were restricted to electrodes anatomically located within the STG for the main analyses. This restriction required that every electrode that was registered in MNI space be proximal to the STG as defined within 10 mm of the Freesurfer anatomical label 'superiotemporal'; subsequent analyses examining the spatial distribution of activity included the 'middletemporal' and 'supramarginal' labels. This anatomical restriction resulted in a total of 210 bipolar-referenced 'virtual' electrodes in the STG. Excessively noisy electrodes were removed either manually or statistically by identifying electrodes with raw signals that were 5 SD greater in comparison to all other electrodes. This resulted in a total of 197 electrodes across participants. For complementary analyses in visual regions, electrode locations were anatomically restricted to the 'inferiortemporal' and 'fusiform' labels. This resulted in a total of 65 electrodes in these two regions, across patients.

Drift was removed from each channel (using residuals from fits to a 3rd order polynomial and high-pass filtering at .1 Hz). Power-line interference was removed by notch-filtering at 60 Hz and its harmonics. Continuous time-series were then filtered into three frequency ranges using wavelet convolution and then power transformed. This resulted in signals at three distinct frequencies reflecting the theta band (3-7 Hz wavelet cycles linearly varying from 3 to 5), beta band (13-30 Hz, wavelet cycles varied linearly from 5-10), and high gamma power (70-150 Hz

in 5 Hz intervals, wavelet cycles = 20 at 70 Hz, and increased linearly to maintain the same wavelet duration across frequencies). Data were then segmented into 2 second epochs centered around speech onset time for a specific stimulus: trial onset was defined as the point when the initial consonant burst occurred. All data were then resampled to 1000 Hz. The three frequency bands were chosen based on evidence of modulatory activity of visual speech on audiovisual speech integration at these frequencies (Arnal et al., 2009; Kaiser et al., 2005, 2006; Micheli et al., 2020; Peelle & Sommers, 2015, Karthik et al., 2021). Apart from extracting spectral components of the signals, single-trial event related potentials (ERPs) were also extracted as the raw voltage from electrode responses.

Electrodes from both the left and right hemispheres were projected into the left hemisphere for analyses and visualization. This projection was performed by registering each participant's skull stripped brain to the Freesurfer cvs_avg35_inMN152 template image through affine registration using the Freesurfer function mri_robust register (Reuter et al., 2010). Right hemisphere electrode coordinates were then reflected onto the left hemisphere across the sagittal axis.

3.2.5 Classifiers for calculating decoding accuracy

A support vector machine (Boser et al., 1992) classifier was utilized for calculating decoding accuracy. Classifiers for stimulus trials were built for individual subjects and group-level analyses were performed by combining results from individual subjects (subject as a random effect). Decoding accuracies were calculated across three different levels for each of the four subjects and two stimulus conditions analyzed. These analyses were performed across three different frequency bands (theta, beta and HGp) along with the single trial ERPs across the conditions. The three levels of analysis performed were 1) Group-level omnibus classification of

auditory-only signals and visual-only signals from all STG electrodes across the time window of interest from 0 ms phoneme onset time until 500 ms after onset. 2) Classification of individual phonemes and visemes across all electrodes together at discrete time points (10 ms bins starting at 2000 ms before phoneme onset time until 2000 ms after onset. 3) Classification of phonemes and visemes at each electrode individually in an omnibus fashion from 0 ms (phoneme onset time) until 500 ms after onset.

For each of the three levels of classification, accuracy rates were calculated using a four-fold cross-validation approach with each fold consisting of 75% of the data for training and the remaining 25% for testing. Classifiers were built for 4-class classification problems investigating the differentiability of the four phonemes or four visemes, evaluated separately.

Table 3. ***Representation of the temporospatial configurations for the various classifiers built****. 0 ms reflects phoneme onset time. Face onset occurred at -500 ms.*

Analysis Type	Spatial range for input	Temporal window of interest for input	Temporal window length	Classes
Group-level omnibus analysis	All electrodes	0-500 ms	500 ms	Phoneme/viseme identity
Time Series Analysis	All electrodes	-2000 ms-2000 ms	Moving 100 ms temporal window	Phoneme/viseme identity
Individual electrode analysis	Individual electrodes	0-500 ms	500 ms	Phoneme/viseme identity

3.2.6 Group-level omnibus analysis

For the group-level omnibus analysis, signals were downsampled to 10 Hz such that every 100 ms time window represented discrete interpolated activity. Starting at 0 ms (phoneme onset time), until 500 ms after onset, downsampled data points (6 in total) were considered at all STG electrodes in each of the subjects across all the trials for each stimulus in the auditory-only and visual-only modalities separately. These data were then used to train and test a classifier for

each subject individually. Following omnibus classification for each individual subject, classification accuracies across subjects were combined to calculate the group-level omnibus classification accuracy and statistics for the auditory-only and visual-only modalities separately.

3.2.7 Time-series analysis

For the time series analyses, signals were downsampled to 100 Hz. Then, using a sliding window of 500 ms in duration, data were examined at each time point starting at 2000 ms before phoneme onset time until 2000 ms after onset. Data from all electrodes were included in the analyses, for all the trials in the auditory-only and visual-only modalities separately. With these data, individual classifiers were built at every time point for individual subjects, using four-fold cross validation. Once classification accuracies were calculated for individual subjects, group-level accuracies at each time-point were calculated as the average of classification accuracies at each of the time points. This analysis was performed separately for auditory-only and visual-only conditions.

3.2.8 Individual electrode analysis

For the classification at individual electrodes, data from each electrode for every subject was examined from 0 ms (phoneme onset time) until 500 ms after onset at 100 Hz (6 points per trial). Data for this time duration was prepared in the same manner as to the group-level omnibus analysis. Then, a classifier was built at every electrode using four-fold cross-validation for the auditory-only and visual-only conditions separately, to identify individual trial labels.

3.2.9 Representational similarity analysis

Apart from the three levels of classification applied in the individual modalities, a representational similarity analysis was performed to investigate a combined omnibus

classification. The difference between the omnibus analysis performed in section 3.2.6 and the representational similarity analysis was that instead of using a four-class classifier to examine the differentiability between phonemes in either the auditory-only or visual-only modalities, we used an eight-class classifier to study the differentiability between the four phonemes and four visemes. Following this, a correlational analysis using Pearson correlation was used to measure the association between misclassification between phonemes and visemes. Specifically, we compared the off-diagonal classification frequencies across matching stimuli in auditory-only and visual-alone conditions. This method enabled a test of whether incorrectly classified phonemes and visemes were related, which would suggest overlapping representations.

3.2.10 Calculating individual subject classification significance

Small sample datasets, especially in neural signals can be faced with the challenge of exceeding chance-level by chance (Combrisson et al., 2015). This would mean that in a four-class classification problem for discriminating between phonemes or visemes with initial consonants '*b', 'f', 'g'* and *'d'*, a chance level of 25% will not accurately capture the underlying characteristics of the data; e.g., randomly sampled labels will be classified at more than 25% with relatively high frequency. Hence, a binomial chance threshold needs to be calculated at each instance of classification using the binomial probability distribution function (Demandt et al., 2012, Combrisson et al., 2015). We utilized the 'binocdf' function in MATLAB for this, by considering two parameters: the number of trials, and probability of success at each instance (25%). This gives rise to a binomial chance-level probability that varies depending on the number of data points used for classification in each of the models that were built. This resulted in a chance probability of 29.58% ($p = 0.05$) for a 4-class classifier with 240 trials and 15.00% ($p = 0.05$) for a 8-class classifier with 480 trials.

3.3 Results

3.3.1 Behavioral results

Participants' mean behavioral accuracy across the two conditions was significantly above chance at the group level: Auditory-only (M = 96.15%, SD = 1.65%, $t(3) = 74.7$, $p < .001$), visual-only (M = 76.35%, SD = 7.54%, $t(3) = 11.8$, $p = .001$). As expected, auditory trials were correctly identified significantly more often than visual trials, $t(3) = 5.76$, $p = .01$.

3.3.2 Decoding analyses at individual frequency bands

A previous iEEG study from our group (Karthik et al., 2021) reported spectral power changes during an auditory-visual speech task observed across multiple frequency bands. To test whether visual speech information encodes visemic information in the STG, we first classified single-trial event-related potentials (ERPs) due to the high temporal resolution of ERPs and because they reflect a combination of spectral phase and power information. To provide a stronger mechanistic interpretation, we additionally classified single-trial responses in theta, beta, and high-gamma power (HGp) filtered data.

3.3.3 Single-Trial ERP Classification

Figure 16 shows the group-level average confusion matrices using SVM classifiers on single-trial ERPs produced in response to phonemes and visemes. SVMs were run separately for the classification of four phonemes and four visemes at the individual subject level: single trial labels (60 for each phoneme and viseme for a total of 480 trials) were classified using time points and electrodes as dimensions. Across all four subjects we observed significant classification (evaluated using binomial statistics) of both auditory-only (phoneme) and visual-alone (viseme) trials using activity recorded from STG electrodes (single-subject values shown

in Table 4). Averaging classification matrices across subjects additionally revealed significant group-level classification for both phonemes ($t(3) = 14.45$, $p<0.001$) and visemes ($t(3) = 21.13$, $p<.0001$). However, we did not observe a significant difference between group-level phoneme and viseme classifications ($t(3)= 1.73$, $p = .18$), likely due to poor group-level statistical power.

The successful classification of phonemes and visemes in the STG indicated that auditory areas were able to reliably discriminate between the consonant initial words for both auditory-only and visual-only speech stimuli. The diagonal of the confusion matrices shows that this classification was robust and significantly above chance for each of the four phonemes and visemes considered (binomial chance = 29.58%, $p = 0.05$).

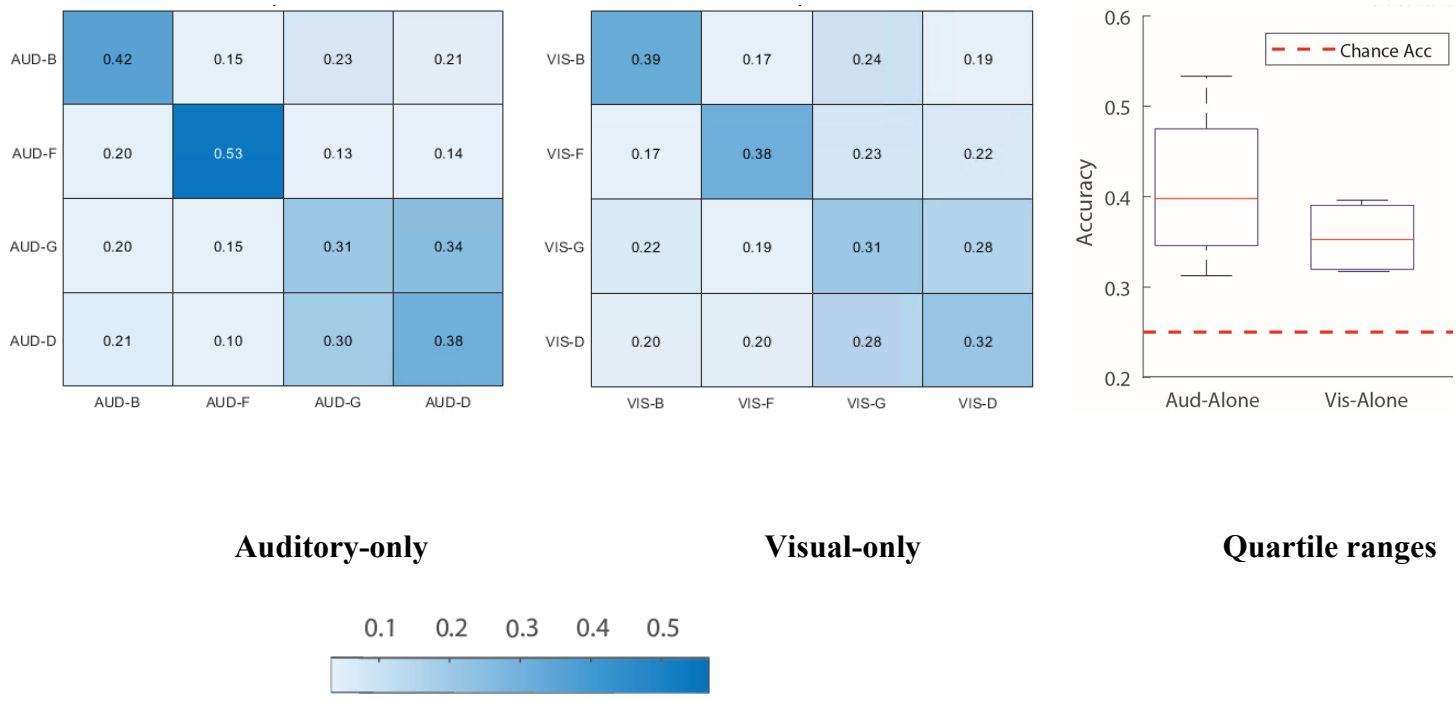


Figure 16. ***Confusion matrices for ERPs****. (Left, Middle) Group-level confusion matrices accuracy using ERPs for discriminating between 4 phonemes and 4 visemes in the auditory-only and visual-only conditions, respectively. Cell values denote the frequency at which each consonant was predicted (x-axis) relative to the true labels (y-axis) Both the auditory-only and visual-only modalities demonstrated reliable above-chance classification performance. No significant differences were observed between the group-level classification performances of the auditory-only and visual-only modalities.*

3.3.4 Time-series classification performance

Classification accuracy for the auditory-only and visual-only stimuli were also calculated across the entire epoch time-course. Each of the classifiers were built for each subject separately at each of the time points considered across all the electrodes.

Table 4. ***Single-trial ERP classification accuracies for individual subjects*** *in the auditory-only and visual-only conditions.*

Subject	Auditory-only	Visual-only
1	34.16 ($p < 0.001$)	36.66 ($p < 0.001$)
2	38.75 ($p < 0.001$)	32.91 ($p = 0.002$)
3	45.00 ($p < 0.001$)	38.75 ($p < 0.001$)
4	46.25 ($p < 0.001$)	31.66 ($p = 0.008$)
	M = 41.04%, SD = 4.88%	M = 34.99%, SD = 2.84%

Figure 17 shows the average time series classification accuracy using an SVM classifier on single-trial ERPs across all four subjects considered over the entire duration of the stimuli in both the auditory-only and visual-only modalities. From the time-series classification accuracy pattern, we see that the auditory-only modality stayed at chance level until 100 ms before phoneme onset time and reached peak classification accuracy (43.54%) at 300 ms following phoneme onset time. This above-chance performance continued until 900 ms after phoneme onset time, followed by a drop back to chance-level.

The visual-only modality stayed at chance level until 0 ms after phoneme onset time, 100 ms after significant classification was observed in the auditory-only condition. Peak classification accuracy (34.27%) was observed at 200 ms following phoneme onset time, and

100 ms before classification peak in the auditory-only condition. Above-chance performance lasted until 1300 ms after phoneme onset time.

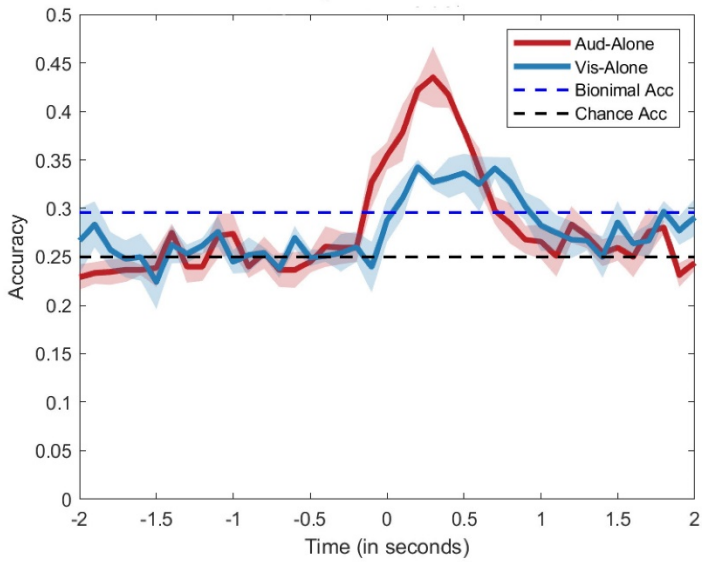
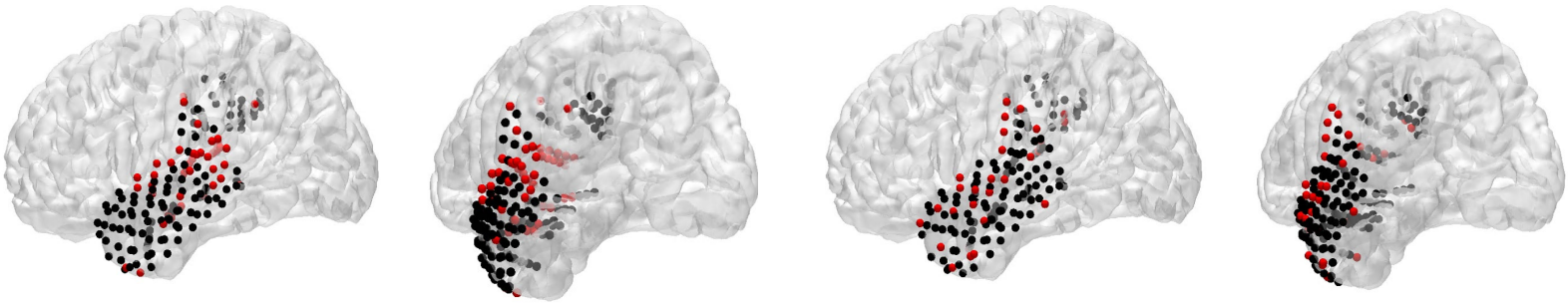


Figure 17. ***Time series classification accuracies using ERPs*** *for the entire duration of the stimuli starting at 2000ms before phoneme onset time until 2000ms after onset.*

3.3.5 Classification performance at individual electrodes

Classifiers were also built for individual electrodes for each subject to understand the spatial distribution of activity that helped capture variance across trials to provide an above-chance classification performance. Figure 18 shows the spatial distribution of STG, MTG, and SMG electrodes present across the four subjects. This figure also indicates the location of electrodes that reliably classified between identities of phonemes or visemes using ERPs. Table 5 shows the total number of electrodes in each subject across different analysis conditions. Figure 19 shows a conjunction analyses that shows the location of electrodes that 1) classified only phonemes 2) classified only visemes 3) classified both phonemes and visemes.



Auditory only **Visual only**

Figure 18. ***Spatial distribution of the electrodes*** *that reliably decoded (indicated in red) and electrodes that did not reliably decode (indicated in black) between identities of phonemes and visemes using ERP responses. All classification were done as a group-level analysis at individual electrodes in the time-range between 0 ms and 500 ms after onset of stimuli.*

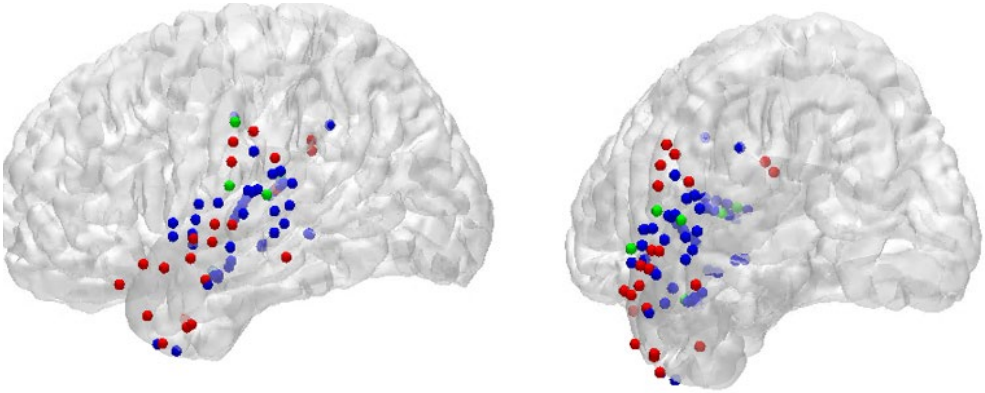


Figure 19. ***Across condition spatial distribution of electrodes****. Spatial distribution of the electrodes that reliably decoded phonemes in the auditory modality (indicated in blue) and electrodes that did reliably decoded phonemes in the visual modality (indicated in red) and electrodes that reliably decoded in both the auditory and visual modalities (indicated in green).*

From these results, we observed that out of a total of 53 electrodes that had significant above chance classification in the auditory-only condition, 44 (83.01%) were present in the STG. Out of a total of 33 electrodes that had significant above-chance classification in the visual-only condition, 25 (75.75%) were present in the STG. Out of these two samples (44 and 25 electrodes), 7 had significant above-chance classification in both the conditions.

Table 5. ***Total number of electrodes for the four subjects****. 'Sig' indicates the number of electrodes in each of the subjects that were able to reliably classify phonemes above chance in the time range of 0 ms after onset of stimuli to 500 ms after onset of stimu*

Subject			Auditory-only		Visual-only		
	Total Elecs	Elecs in STG	Sig. in STG	Sig. in aud but not in vis	Sig. in STG	Sig. in vis but not in aud	Sig. in both aud and vis
1	36	11	5	4	2	1	1
2	45	25	10	6	10	6	3
3	66	42	11	7	11	7	2
4	48	38	18	18	2	0	1
Total	195	116	44	35	25	14	7

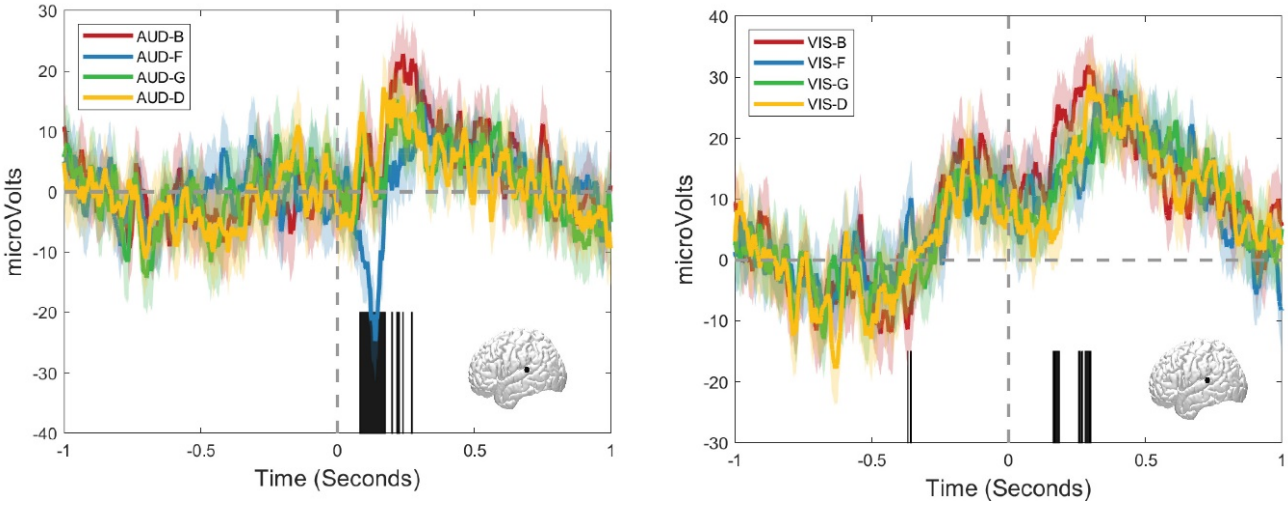


Figure 20. ***Example ERPs from an STG electrode*** *in one of the patients. The left plot shows auditory ERPs to the four phonemes and the right plot the ERPs to the four visemes. Black bars denote significant time-points from univariate analyses (one-way ANOVAs) examined separately at each time-point and corrected for multiple comparisons using FDR. Viseme-related activity begins shortly after face-onset (at -.5 seconds) whereas phoneme-related activity begins after sound onset (at 0 seconds). F-initial words evoked more negative going ERPs in the first 300 ms of both phoneme and viseme trials, whereas B-initial words evoked more positive going ERPs in this same time range.*

3.3.6 Representational similarity analysis

From analyses performed in the previous sections, we saw that the ERPs were able to reliably encode phonemic information from visual speech in the auditory areas. But these analyses do not indicate if phoneme and viseme information is represented in a similar manner across stimulus types. To understand the representation of phonemes and visemes in the auditory areas, we performed representational similarity analysis using 8-class confusion matrices (4 visemes and 4 phonemes).

Figure 21 (left) shows the confusion matrix of an omnibus classification between the individual stimuli across auditory-only and visual-only modalities. The upper left quadrant of the matrix shows the confusion matrix for the auditory-only conditions while the lower right quadrant shows the confusion matrix for the visual-only conditions. From this matrix, we can see that both the auditory-only and visual-only modalities performed significantly above-chance even in the 8-class classification framework as they did in their respective 4-class classification frameworks. To investigate if phonemes in the auditory-only conditions had similar representations in the auditory areas as the visemes in the visual-only condition, we compared the levels of misclassification for each of the phoneme/viseme pairs. For example, is the confusability between auditory *'f'* and auditory *'b'* similar to that between visual *'f'* and visual *'b'?*

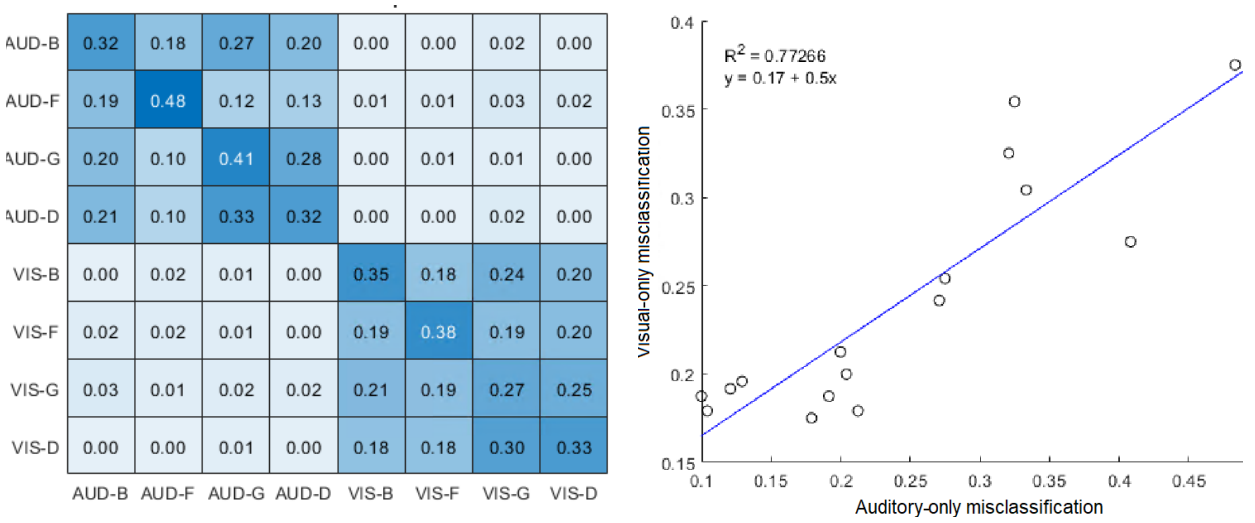


Figure 21. ***Confusion matrix across auditory and visual modalities****. (Left) Confusion matrix showing the classification accuracies across a 8-class classification framework. (Right) To test for a similar pattern of misclassification between auditory and visual trials, we compared off-diagonal responses in the upper left and lower right quadrants in the confusion matrix, shown as a scatter plot. Mis-classification rates were highly correlated (r = 0.88 , p = 0.001) consistent with the hypothesis that visemes evoke activity in matching auditory phoneme populations in the STG.*

Because phonemes are represented through population coded responses, misclassification can reveal information about related neural processes. To test for related representations, we calculated a correlation between each of the phoneme-pairs across both the modalities on group-averaged confusion matrices. Figure 21 (right) shows the scatterplot for this analysis, where the x-axis represents the off-diagonal values of the auditory-only confusion matrix and y-axis represents the off-diagonal values of the visual-only confusion matrix (12 values from each). This analysis reveals that there is very high correlation ($r = 0.88$) between the confusability of phonemes in the auditory-only modality and visemes in the visual-only modality. We calculated the significance of this test by randomly permuting the stimulus labels of each trial and repeating the full classification analysis $n = 1000$ times. Results showed the observed correlation value to be highly significant, $p = .001$. This is consistent with our hypothesis that the spatiotemporal neural representation of viseme identities in the auditory areas is similar to that of phonemes.

3.3.7 Classification of Spectral Power

Average auditory and visual classification rates for theta, beta, and HGp are shown in table 6. In auditory trials, significant group-level classification was observed for theta band, $t(3) = 4.14$, $p = 0.01$; and HGp filtered data $t(3) = 4.67$, $p = 0.009$, but not beta, $t(3) = -0.45$, $p = 0.65$. In visual trials, only a marginal group-level classification was observed for HGp: $t(3) = 3.16$, $p = 0.06$, with no above-chance differences observed for theta band, $t(3) = -0.81$, $p = 0.48$, or beta band, $t(3) = -2.37$, $p = 0.1$, activity.

Table 6. ***Average group-level classification accuracies across all four subjects*** *considered for the theta band, beta band and HGp. The numbers in the parentheses indicate the number of significant subjects out of the total number of subjects analyzed.*

Modality	ERP	Theta band	Beta band	High Gamma power
Aud	41.04% (4/4)	35.93% (3/4)	27.91% (2/4)	36.14% (4/4)
Vis	34.99% (4/4)	23.43% (0/4)	24.27% (0/4)	27.26% (2/4)

3.3.8 Classification in Visual Regions

To understand the complementary encoding of viseme information from visual regions, we investigated the ability to classify single-trial responses from the inferior temporal lobe (including the fusiform gyrus) during auditory-only and visual-only speech. Two out of the four subjects had electrode coverage in these regions and classifiers were constructed separately for each of these two subjects across 47 electrodes (subject 1, 27 electrodes; subject 2, 20 electrodes).

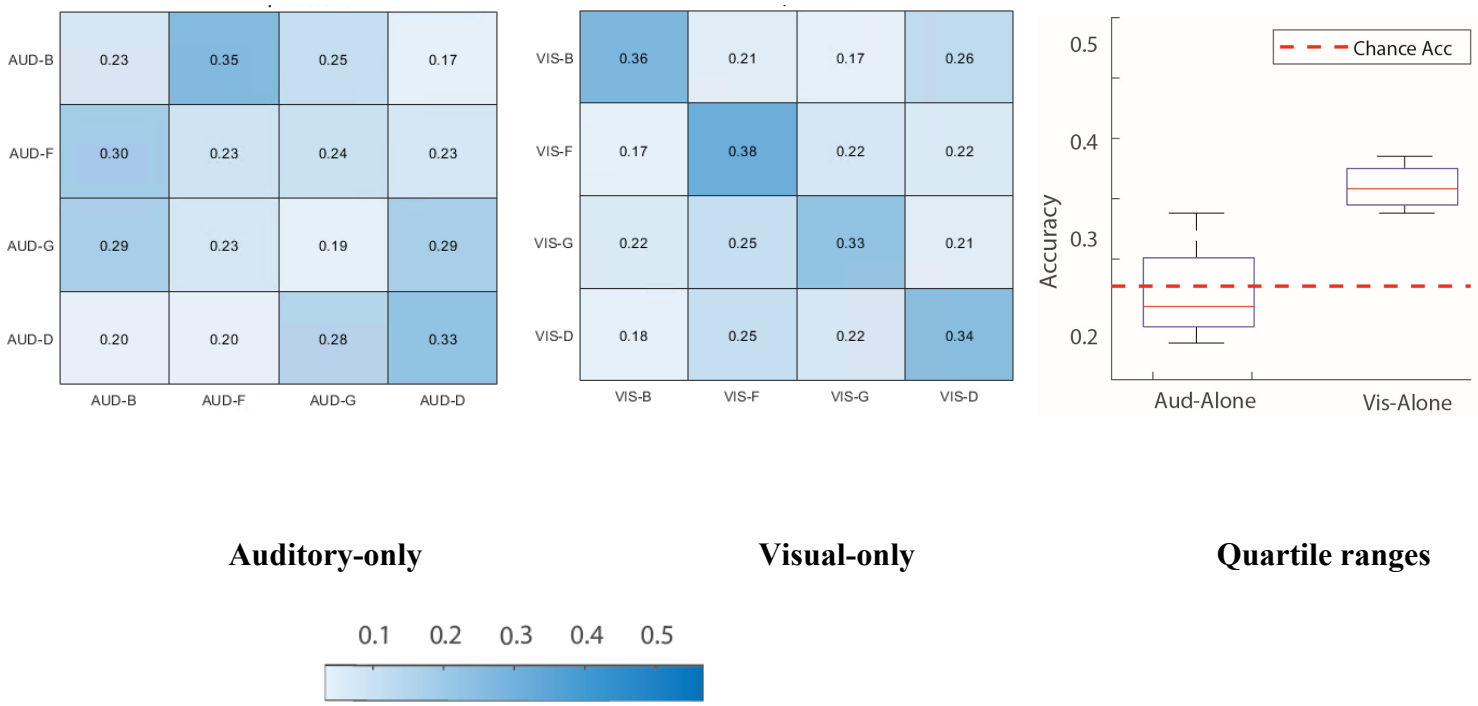


Figure 22. ***Confusion matrix in the fusiform region****. Group-level classification accuracies using ERPs for discriminating between 4 phoneme and 4 visemes conditions in the auditory-only and visual-only modalities respectively in the fusiform region.*

Figure 22 shows the group-level average confusion matrices across 2 subjects using SVM classifiers on single-trial ERPs recorded from visual regions in response to phonemes and visemes. Classifiers were built in the same manner as for analyses in auditory regions. Confusion matrices revealed above-chance classification accuracies for visemes in both subjects: Subject 1 = 37.08%, $p < 0.001$, Subject 2 = 34.58%, $p = 0.001$. However, consistent with the unidirectional transfer of speech information, we observed chance-level classification accuracies for phonemes in both subjects: Subject 1 = 24.17%, $p = 0.53$, Subject 2 = 25.25%, $p = 0.45$..

Decoding at individual time-points can be seen in Figure 23, and demonstrated that phonemic classification stayed at chance levels throughout the time-course of the stimuli. Conversely, the visual condition stayed at chance-level until phoneme onset time around 0 ms,

lasting until 1500 ms after phoneme onset time. A peak classification of 37.50% was seen at 900 ms after phoneme onset time.

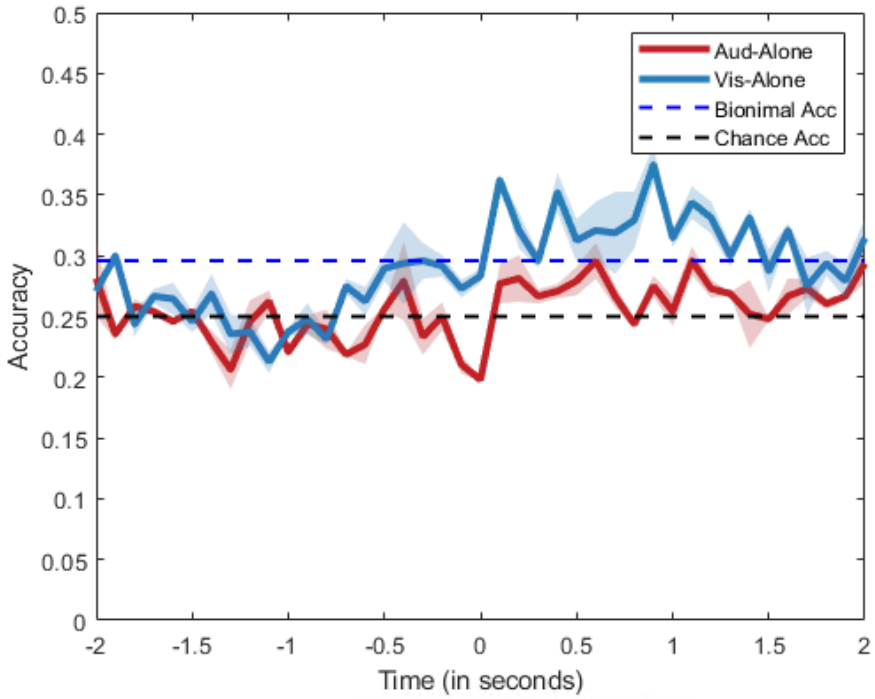
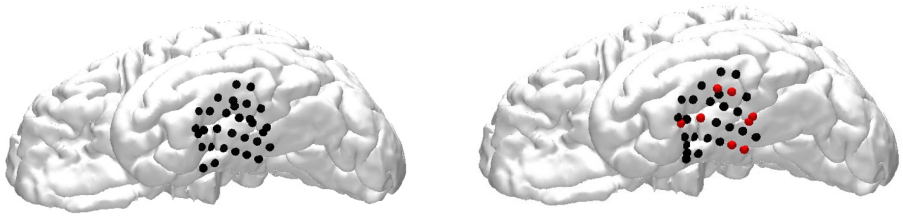


Figure 23. ***Time series classification accuracies using ERPs at the fusiform region,*** *for the entire duration of the stimuli starting at 2000 ms before onset of stimuli until 2000 ms after onset of stimuli.*

Decoding at individual electrodes revealed that no electrode from the fusiform region in either of the subjects were able to reliably classify phoneme identities in the auditory-only condition. On the other hand, in the visual-only condition a total of 11 electrodes (subject 1 = 6 electrodes, subject 2 = 5 electrodes) were able to reliably classify phonemic identities above-chance. The spatial distribution of electrodes in each of the condition can be seen in Figure 24.



Auditory-only **Visual-only**

Figure 24. ***Spatial distribution of the electrodes in the fusiform region*** *that reliably decoded (indicated in red) and electrodes that did not reliably decode (indicated in black) between identities of phonemes and visemes using ERP responses. All classification were done as a group-level analysis at individual electrodes in the time-range between 0 ms and 500 ms after onset of stimuli.*

3.4 Discussion

Prior research has demonstrated that visual speech can influence auditory speech perception in multiple distinct ways. Visual speech has been most studied for the visemic influence it provides, as seen in the McGurk effect (McGurk & McDonald, 1976). Visual speech also provides complimentary information about what is heard via phonemic representations (Bourguignon et al., 2020), timing information (McGrath & Summerfield., 1985), speech rate (Chandrasekaran et al., 2009), and spectral information (Plass et al., 2020). Moreover, it has also been shown that these visual influences are represented spectrally, temporally and spatially with multiple distinct processes occurring simultaneously (Karthik et al., 2021). One possibility is that some of this visual activity in auditory areas reflects the transformation and representation of viseme information to bias and improve auditory speech processing. To test this hypothesis, we examined whether viseme identities during visual speech could be decoded from iEEG signals in the human auditory cortex. We analyzed data from 4 patients with epilepsy who had intracranial electrodes intracranially implanted in the auditory cortex including the STG, MTG and

supramarginal regions. We constructed a classification pipeline to decode the identities of phonemes and visemes in the primary auditory areas during auditory and visual speech.

Classification results from our group-level omnibus analysis on electrodes in auditory areas (STG) demonstrated significant above-chance decoding in both the auditory-only and visual-only conditions using ERP responses. Hence, we performed all our further analysis using ERP responses in both the auditory-only and visual-only stimuli conditions. These results demonstrate that visual speech information is indeed encoded in the auditory cortex. We also noticed above chance-level decoding in the visual-only condition in the fusiform region, while the auditory-only condition provided chance-level decoding in the fusiform. The fact that this information are maximally available in ERPs as opposed to spectral components could mean that visual speech by itself is not sufficient to generate large field potentials in the auditory areas, indicating subthreshold activity. Moreover, given the small sample size we utilized in our study, the high variance in beta and theta band could further hamper our ability to interpret results in those frequency bands.

With ERPs, the classifiers showed above-chance decoding accuracies for individual phonemes, indicating that the classifiers captured the variances reliably across all the phonemes analyzed. Notably, in both the auditory-only and visual-only conditions, *'b'* and *'f'* had lower confusability than *'g'* and *'d'*. More specifically, *'f'* had a higher decoding accuracy in comparison to all the other phonemes in both the auditory-only and visual-only modalities. In conjunction with the representation similarity analysis, this difference in decoding accuracies across different phonemes sheds more light on the nature of phonemic representations in the neuronal populations of the auditory cortex. From the individual phonemic decoding accuracies in each of our stimuli and their representational similarity analysis, we see that there is very high

correlation between the decoding accuracies of auditory-only and visual-only stimuli. This similarity extends to the higher decoding accuracies of /f/ in both these modalities and similarity in patterns of decoding accuracies across the other phonemes.

The similarity in pair-wise representations of these individual phonemes align with the known representation of phonemes in the auditory cortex, which follow a highly selective, spatially distributed pattern (Gyol Yi et al., 2020). This representation allows for the identities of individual families of phonemes to be selectively encoded in specific regions of the auditory areas. Accordingly, individual phonemes are grouped into specific families depending on their similarities and it was shown that *'b', 'd'*, and *'g'* belonged to the same family, while *'f'* belong to a second, but highly related family of phonemes. Our results complement the results from Gyol Yi et al by replicating a similar pattern of phoneme representation in the auditory areas, and extend their findings by showing that viseme distributions similarly overlap with phoneme encoding.

A time-series analysis of decoding accuracies in the auditory-only and visual-only modalities across the entire time course of the stimuli showed that it is possible to decode identities of individual phonemes/visemes over the time course of our stimuli. Interestingly, it was observed that the above-chance classification performance in the visual-only condition had a temporal lag in comparison to the auditory-only condition. Interestingly, the visual-only condition has a 100ms head start (100ms after stimuli onset) compared to the auditory condition (200ms after stimuli onset) when comparing when the electrodes start encoding stimuli identity information. Seen together with previous results from our group (Karthik et al., 2021) which showed modulatory effects of visual speech on auditory speech perception, this visual head start of 100ms lines up neatly with established behavioral evidence (Karas 2019), and hints that the

visual speech's neural modulatory effects in fact encode meaningful information about the phonemic representation of the speech content. Critically, this time-lag could help argue against the idea that neural activity in the auditory cortex in response to visual speech might be generated by internally vocalized speech (Bourginon 2020). Previously, this idea has been argued by research that showed low frequency entrainment of the unheard envelope during silent speech in the auditory cortex (Bourginon 2020, Hauswald 2018), but such theories fail to explain word and phonemic level visual speech benefits (Ross 2007) and would be anticipated to have a much longer lag time.

The topographical distribution of phoneme and viseme information across auditory regions was quantified in the individual electrode analysis. This analysis showed that the spatial distribution of electrodes that reliably decoded visual-only information is distributed over the STG and parts of the MTG. Additionally, individual electrodes that reliably decoded information from both auditory-only and visual-only modalities were concentrated in the pSTG and pSTS. The dual encoding of visemes and phonemes in pSTS is well in line with a previous body of work that demonstrated the role of pSTS in audiovisual speech integration (Nath, 2010). Here, we extended such findings by providing the first evidence that some of these audiovisual responses in the pSTG and STG reflect population-coded information about viseme identities. Moreover, some electrodes in the STG and MTG showed significant classification for visual-only by not auditory-only stimuli. We think two possible explanations account for this pattern. First, some neuronal populations in the STG are targeted purely by visemes and not phonemes (Falchier et al., 2002, Barraclough et al., 2005). Second, and perhaps more likely, the viseme-targeted regions included auditory neurons that are responsible for processing phonemes in noise, in more challenging situations (e.g., under noise). This is ecologically valid because

visemes are most useful when phonemes are hard to hear, and this is consistent with the lack of phonemes classification in these regions because the auditory-only stimuli were largely clear (50 out of 60 of the trials used for each phoneme).

While we obtained a larger data size for each subject by combining two different task sets, there might be concerns about the validity of combining task variant 1 (where participants were presented with only auditory-only or visual-only stimuli) and task variant 2 (where participants were presented with four conditions: auditory-only, visual-only, incongruent auditory-visual and congruent auditory-visual). However, these concerns were addressed by the results that showed comparable decoding performance across both the task variants.

In summary, the current study demonstrates that phonemic and visemic level stimuli of spoken auditory and visual speech are both encoded in the auditory cortex during speech perception. Phonemic and visemic information aren't just encoded in the auditory cortex, we also provided evidence that categorical visemic information is encoded in the fusiform area, echoing previous findings that visual cortex performs visemic level processing of visual speech stimuli (Nidiffer, 2021). This provides a foundation for future work to investigate the role of the visual regions in audiovisual integration and information transfer between the auditory and visual regions during audiovisual speech processing. Furthermore, results in literature have shown that it is possible to decode identities of spoken words in a noisy environment from the auditory cortex (Chan et al., 2014). Given that visual components of spoken speech provide enhancements in audiovisual speech perception (Bourguignon et al., 2020, Plass et al., 2020), the methodology and results from the current study could help build towards understanding what phonemic information from visual speech aids in enhancement of perception during audiovisual speech in a naturalistic environment.

3.5 Supplemental Material

3.5.1 Electrode preselection

In order to circumvent the problem of curse of dimensionality (Aggarwal et al., 2005, Verleysen et al., 2005 Bach et al., 2017), we present a novel technique to preselect electrodes that are not only functionally significant but have similar event related potentials associated with the stimuli of interest. This would lead to potential improvements in classification performance by acting as a feature selection procedure (Remeseiro et al., 2019) by reducing the dimensionality of data and ignoring noisy electrodes, and hence leading to better signal to noise ratio for classifier discriminability of phonemes. For the preselection procedure, we used a clustering-based approach utilizing a k-medoids algorithm with dynamic time-warping (DTW) as a distance measure.

3.5.2 K-medoids clustering with dynamic time warping

To preselect electrodes, we chose to cluster electrodes that responded similarly to the stimuli of interest. This could be achieved through a clustering algorithm. K-means clustering (Ahmed et al., 2020) which is a popular algorithm for clustering has been widely utilized for grouping together observations in data that are similar to each other. However, a drawback with this algorithm is that the centroids of each of the chosen clusters are virtual (i.e., the centroids do not exist in the actual data and are created virtually in an iterative fashion). Moreover, the algorithm does not perform optimally for time series data of high dimensions that have similar shapes in individual clusters (Ahn et al., 2018). To handle these issues we propose using k-medoids clustering (Kaufman et al., 1990), which is a modified version of the more popular k-means algorithm.

3.5.3 K-medoids

The advantage of using k-medoids clustering over k-means clustering is that the algorithm allows for using real observations within the data set as its centroids. This leads to partitioning of the data with observational data present within the dataset as a reference point. This leads to better interpretability of the clustering results, where an electrode present within each of the clusters can be identified as the cluster's reference point and verified for robustness. Additionally, the k-means algorithm conventionally utilizes Euclidean distances as a distance-measure for calculating the distances between individual data points and the cluster's centroids. This reduces the versatility of this algorithm since in some cases (e.g., clustering time series data) Euclidean distance measures might not be an accurate representation of distances (Kalpakis et al., 2001). In k-medoids, the distance measures are versatile and can be chosen according to the data in question. Hence, for clustering of data generated by individual electrodes we utilized dynamic time warping (Berndt et al., 1994), which is a distance measure utilized for calculating the similarity between pairs of temporal sequences

3.5.4 Dynamic Time Warping (DTW)

The algorithm functions by calculating the warped distances between two temporal sequences of data, hence finding how similar/dissimilar two signals are (Berndt et al., 1994). Thus, the technique allows for calculating similarities between signals of different shapes. By using DTW as a distance metric for calculating the similarity between electrodes, we calculate the similarities between the data generating processes underlying these electrodes. Once the similarities between pairs of time series signals generated by the electrodes are calculated, it is used as a proxy for the distance measure required for the k-medoids clustering algorithm. Electrodes with similar data generating processes as evidenced by similarities in the time-series

signals generated by these electrodes are then clustered together using k-medoids algorithm with DTW as a distance metric.

3.5.5 Choosing the optimal value of k

One heuristic that needs to be chosen for the optimal functioning of any clustering algorithm is selection of the required number of clusters. While there are no standard ways to choose the optimal number of clusters, there are acceptable ways to measure the stability of the chosen number of clusters (Ben-David et al., 2007). The most commonly used technique is the elbow method, where the intra-cluster variance (calculated as the inertia of each cluster) is calculated at each k, indicating the k at which the variance begins to become unstable. While this technique works well for low-dimensional data, it is unclear about empirical methods that can be used for high dimensional temporal sequences. Hence, we visually inspected the cluster robustness by checking the patterns of data for varying values of k and chose 3 as the optimal number of clusters for our data of interest.

3.5.6 Partitioning the data to avoid overfitting

Selecting electrodes as a preselection procedure before utilizing them for classification could lead to overfitting since there could be a double bias in terms of which electrodes get selected, and hence which data is used for the final training and testing of the model. To avoid this, we split the data from each electrode into two halves (e.g., if an electrode contains data from a total of 80 trials, it was split into 40 trials in each half). One half of this data was used to cluster the electrodes into different clusters. Once the electrodes were clustered using one half of the data, an omnibus classifier was built (section 3.2.6) across each of these clusters individually

using the same data that was used to cluster the electrodes. The cluster with the highest test accuracy was chosen as the one which contained the most amount of decodable information.

3.5.7 Representational responses of electrodes clustered with k-medoids and DTW

The electrode preselection performed using k-medoids and DTW acted as a feature selection technique where only electrodes of interest in specific clusters were chosen. Representational plots of these clusters are shown in Figure 25 shows results from the visual only condition for high gamma power where we can see the individual electrode responses to each phoneme in each of the three clusters. Figure 26 shows the mean responses of these clusters to each of the four phonemes across all electrodes.

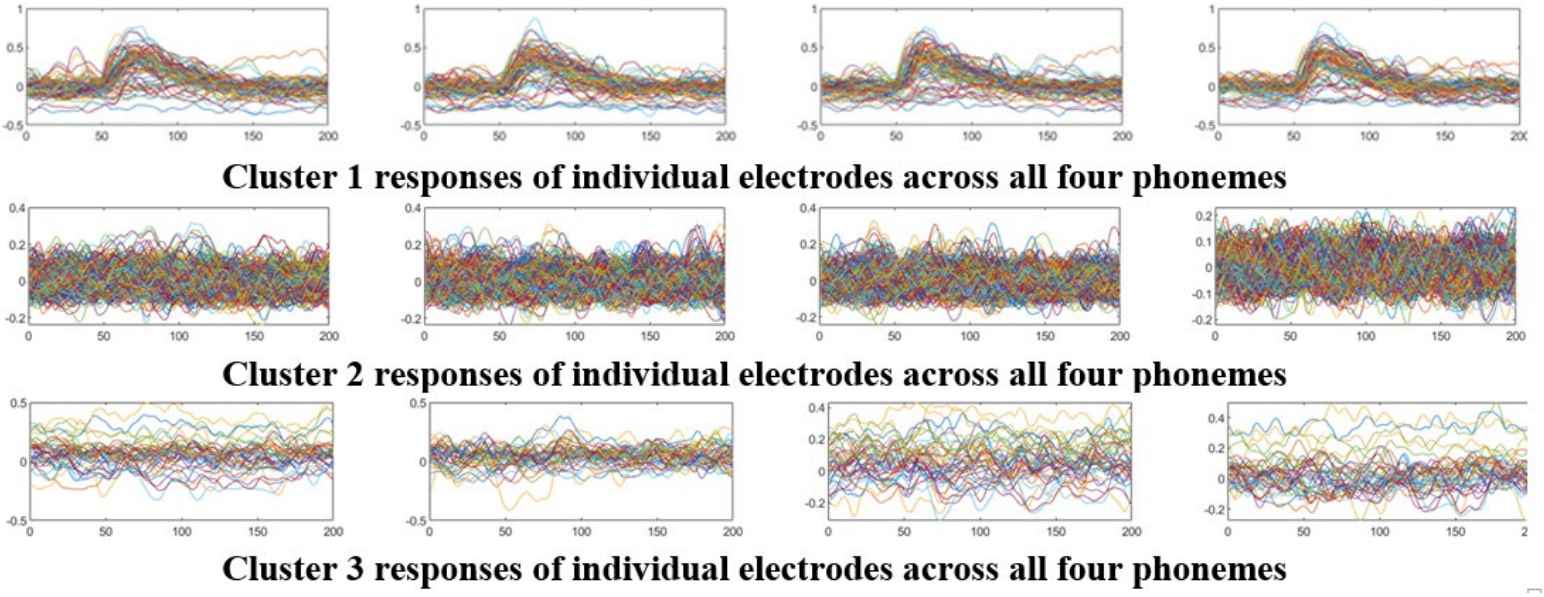


Figure 25. ***Individual cluster responses****. Responses of the individual electrodes to each of the four phonemes in the three clusters considered in the visual only condition of high gamma power.*

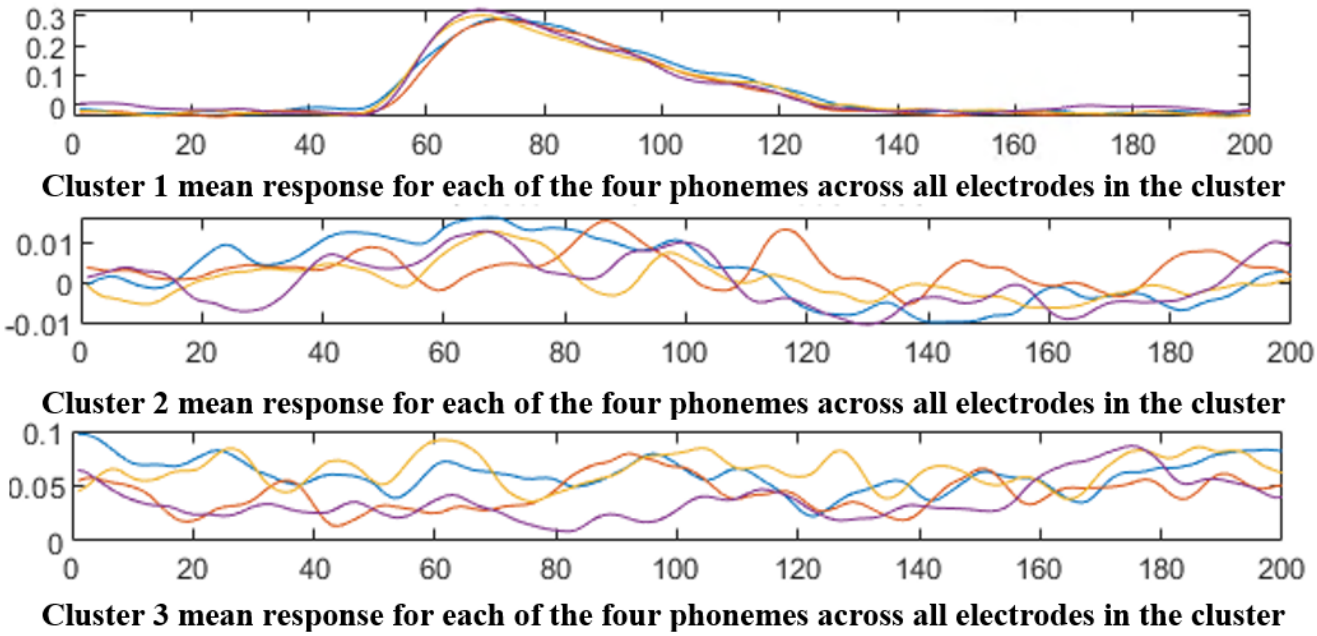


Figure 26. ***Average responses of the individual phonemes*** *across all electrodes present in each of the three clusters in the visual only condition of high gamma power.*

Chapter 4 Phonemic Representations Encoded in Auditory Cortex During Visual Speech: A Study Using fMRI

While speech perception is largely an auditory process, visual information from the speaker's face perceptually enhances relevant information for the listener. This information includes the speaker's voice as well as articulatory movements from the speaker's face and lips. While several studies have shown significant activation of auditory regions in response to visual speech, the information encoded by these activations remains poorly understood. Indeed, our results in the previous chapters demonstrated that visual speech modulates auditory speech processing at multiple temporal, spatial, and spectral scales, indicating multiple discrete influences of vision on speech. We also showed evidence in support of the hypothesis that it is indeed possible to decode information about phonemic representations encoded in the auditory cortex during visual speech. But this information was decoded using a single-subject analysis, on spatial data whose locations were constrained with respect to where the electrodes were implanted. Specifically, the location of the electrodes was dictated by clinical needs. This leads to the question of generalizability across a normative population, with replicable group-level effects across a larger set of subjects. Moreover, the sparse coverage in patients prevents a comparison of decoding accuracy across regions, preventing a hierarchical view of how viseme information is represented across auditory, auditory-visual, and visual regions.

The goal of this chapter is to conceptually replicate the results from the previous chapter in a normative population with comparable spatial data across multiple subjects, and to further

understand how visemes influence auditory population responses . To this end we used functional magnetic resonance imaging (fMRI) to test whether visemes (visual units of speech) can be classified from auditory cortex. This chapter presents data collected from a large sample of individuals (*n*=64, pre-registered at osf.io) who were presented with three phoneme and viseme exemplars (consonant-vowel pairs /FAFA/, /MAMA/, /KAKA/) in a randomized event-related design. We trained and tested classifiers to discriminate between the three specific phonemes and separately the three specific visemes. Analyses were performed using a whole-brain searchlight classifier as well as ROI-based analyses focusing on auditory, auditory-visual, and visual regions. Results showed significant above-chance decoding accuracy in both the phoneme and viseme conditions in auditory regions, including the posterior superior temporal sulcus (pSTS) and superior temporal gyrus (STG). Multivariate analyses showed similar spatial patterns between phoneme and viseme pairs, suggesting that visemes target matching phoneme populations. These results demonstrate that visual speech crossmodally activates and encodes phonemic information in auditory regions.

4.1 Introduction

Though sounds play a major role in understanding spoken speech, meaningful information provided by visual cues to speech content, timing, and speaker identity aid in perceptual enhancements for the listener (Chandrasekaran et al., 2009; Van Wassenhove et al., 2005). In the presence of audio-visual stimuli, listeners naturally tend to integrate information from both modalities to create a unified percept. This unified percept is thought to be created as a result of using complementary information, such as spatiotemporal and statistical correspondences obtained from the two modalities (Spence., 2007, Reale et al., 2007). This

information includes the speaker's voice as well as articulatory movements from the speaker's face (Chandrasekaran et al., 2009; Erber, 1975; Chen and Rao, 1998; Van Wassenhove et al., 2005) and lips (Besle et al, 2008).

Consistent with behavioral evidence that visual information can alter what is heard, such as in the McGurk effect, visual signals strongly modulate the response of auditory neurons to sounds (Ghanzafar et al., 2010, Ghanzafar et al., 2008, Zhu & Beauchamp, 2017). It has also been observed that silent lip-reading activates the auditory cortices and entrains cortical oscillations in the same manner and anatomical regions as auditory-only speech (Bourguignon et al., 2020). Further, apart from activating the auditory cortex, visual speech also modulates auditory speech in multiple ways (Karthik et al., 2021), indicating that audiovisual speech integration is not a unitary phenomenon. The anatomical regions implicated in these modulations involve the primary auditory cortex and the superior temporal gyrus (STG) (Beauchamp et al., 2004). These audiovisual modulations likely occur through feedback connections from the multisensory posterior superior temporal sulcus (pSTS) (Beauchamp et al., 2010).

Though the auditory cortex has been shown to be responsive to silent lipreading (Yi et al., 2019, Beauchamp et al., 2004, Ye et al., 2017), the functional bases of these responses are still unclear. It is yet to be understood whether the neural responses in STG during visual speech are caused purely due to visual cues, such as speaker identity, lip movements and articulatory gestures, or whether they also encode information about phonemic representations in the speech. The neural responses in STG during visual speech could also be driven by arousal or attentional increases that are general perceptive processes with no phonemic information being represented in them. While prior studies with fMRI (DeWitt & Rauschecker, 2011) and intracranial electrode recordings (Chan et al., 2014) have implicated the STG and pSTS in multiple aspects of

phonological processing and visual representations of speech (Ye et al., 2017), they also lead to other related questions about the very nature of neural representations of multisensory speech signals. These questions include fundamental expositions about the identity of features encoded in the auditory cortex and the acoustic, visual, and phonemic descriptions of speech in general. These questions can be understood by differentiating between task-based regional activations and the type of information encoded in those regions. This is motivated by studies that have shown visemes to be responsible for activating phonemic population of neurons during visual speech (Karthik et al., 2021, Beauchamp et al., 2010). Past literature also point to evidence which indicates that visemes can be transformed into categorical phonemic units in the pSTS (Beauchamp et al., 2010).

One way to differentiate between activation magnitude and informational content encoded in activations is to perform a multivariate analyses using a classifier-based decoding approach such as fMRI-based multivariate pattern analysis (MVPA) (Haxby et al., 2014). Previous decoding-based approaches using electrocorticography demonstrated that speech patterns could be reconstructed with signals from the auditory cortex (Mesgarani & Chang, 2012, Makin et al., 2020). These findings also provide evidence to show that phonemic identities are represented in a distributed fashion in the neuronal populations of the STG. Moreover, they indicated the validity of using informational-decoding based analysis to understand the nature of information encoded during speech comprehension. In this study, we used both searchlight (Kriegeskorte et al., 2006) and ROI-based MVPA analyses to investigate the nature of visual activations that occur in the auditory cortex during speech perception. To this end, we test the hypothesis that part of the visually-evoked activity in the STG reflects the activation of phoneme-specific populations by visual speech. In line with results from previous studies, we

focus on the STG and the pSTS, including the supramarginal gyrus (Beauchamp et al., 2010, Claire & Chang, 2019, Yi et al., 2019, Bourguignon et al., 2020).

4.2 Materials and methods

This study was pre-registered at OSF (https://osf.io/6fzwd/). Minor deviations from the pre-registered protocol are noted throughout the methods section. The study was approved by the Institutional Review Board (IRB) of the University of Michigan. A power analysis was conducted to estimate the number of subjects required to test the proposed hypothesis (see section 4.2.2 below). Based on this power analysis, data was acquired from 64 subjects. Subjects were recruited by emailing individuals who had voluntarily registered to be participants in the University of Michigan's Psychology paid-subject pool and through word of mouth, including individuals who had previously expressed interest in studies at the University of Michigan. We recruited 64 participants (F = 47, M = 17) in the age range of 18-32 (Mean: 22.87, SD = 3.29) irrespective of their gender, race or handedness. Participants were paid USD $20 per hour for their time. All subjects were provided with details about the study and IRB approved consenting procedures were completed before proceeding with data collection. Data was collected from each participant in a single session lasting approximately 1 hour and 15 minutes.

4.2.1 Tasks, Stimuli and Experimental Design

We used an auditory and visual speech paradigm optimized for an event-related fMRI design. On each trial, participants were presented with a three-alternative forced-choice task that consisted of either an auditory-only stimulus or a visual-alone stimulus. Three types of phonemes; /fafa/, /kaka/ and /mama/ and three types of visemes; /fafa/, /kaka/ and /mama/ were used for this task. These specific phonemes were chosen to maximize the differentiability

between the individual phonemic representations in the neuronal populations of the STG (Mesgarani & Chang, 2012, Gyol Yi et al., 2019). Figure 27 shows the timing and structure of the task. Each trial for both the auditory-only and visual-alone conditions lasted for 2 seconds. The auditory-only trials began with a fixation cross against a black screen, with the phonemes presented 250 ms after the appearance of the fixation cross. The visual-alone trial began with the appearance of a female actor's face on the screen, with lip movements beginning 250 ms after face onset. After the presentation of each auditory-only or visual-alone trial, subjects were presented with 3 options (fa, ka, and ma) and were instructed to press one of three associated buttons on an MRI-safe button response box.

The first 24 participants were shown response choices that always appeared in the same order (fa, ka, or ma) with a stable mapping between response choice and button (the index finger was always used to make the response for /fa/, the middle finger for /ka/ and the ring finger for /ma/). While performing the sample size estimates for our power analysis, we saw that the stable mapping between response choices and button presses resulted in response type differentiability in the motor cortex consistent with prior evidence for motor regions encoding information about finger movements (Shen et al., 2014). Hence, to counteract this effect and to negate the confounds of motor region responses during speech perception (Wilson et al., 2004), we altered the pre-registered protocol for the remaining 40 participants, who were shown response choices that were randomized after each trial.

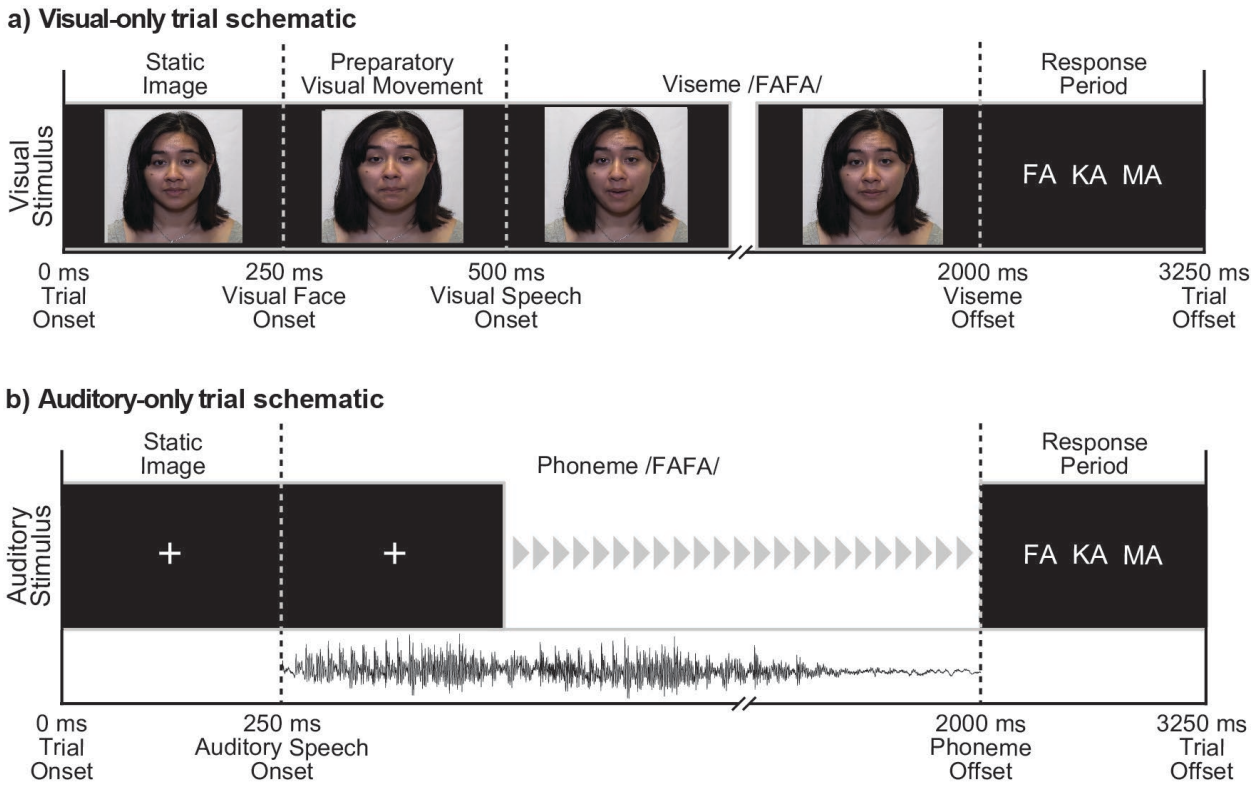


Figure 27. ***Schematic of the task****. Trial schematic for the Visual and Auditory conditions. All visual stimuli begin with a female actor's static face appearing on the screen. 250 ms later the actor makes preparatory lip movements, following which at 500 ms visual-only speech onset occurs. This lasts for 1500 ms. This is followed by a 1250 ms response period. All auditory stimuli begin with the appearance of a fixation cross. 250 ms following the appearance of a fixation cross, auditory-only speech onset occurs that lasts for 1750 ms. This is followed by a 1250 ms response period. Both the visual-only and auditory-only conditions use three types of double-visemes or phonemes respectively: /fafa/, /kaka/, and /mama/*

Participants had 1.25 seconds to respond to the answer choices. If the participant failed to register a response within 1.25 seconds, the trial was recorded as a missed response. Every trial was followed by a 5-6 second jitter period (sampled from a uniform random distribution) which acted as the intertrial interval (ITI). In each run, participants completed 60 trials that were split between 30 auditory-only and 30 visual-alone trials, with 10 trials each for every phoneme and viseme; trial types and stimuli were randomly intermixed in each run.

In total, participants completed five runs, resulting in 300 trials in total (150 phonemes, 150 visemes) during the task, with each run lasting 8 minutes and 30 seconds. Psychtoolbox was used for stimulus delivery and recording timing information and participant responses. Auditory

stimuli were presented using fMRI compatible Avotec headphones that had integrated earmuffs in order to achieve maximum reduction of scanner noise. The sound level of stimuli was held constant for all participants. While presenting auditory speech stimuli in an MRI scanner can be challenging, the undegraded nature of the auditory stimuli enabled near perfect accuracy throughout the task. A mirror system reflected the visual stimuli from an LCD projector onto a mirror (width of the mirror: 12cm, approximate viewing distance between eye and mirror: 15cm; width and height of the face on screen: 9cm x 12cm) located inside the magnet bore of the scanner.

4.2.2 Sample Size, Stopping Rule and Pre-registration

The main comparison of interest in our study was whether viseme identity can be decoded from a region in the brain, particularly within auditory and auditory-visual regions. To estimate a required sample size, we conducted a power analysis on a separate condition (phoneme perception trials) in the first 24 subjects as an orthogonal contrast to our main question of interest. Since our study intends to perform multivariate pattern analysis (MVPA) based decoding, there is no direct way to perform a power analysis in order to obtain a suitable sample size. Hence, we performed a series of paired t-tests across phonemes, based on the assumption that this approach would yield a more conservative estimate to identify group differences in comparison to MVPA based decoding, and used these results for power analyses. In this analysis, we examined the group differences between the three phonemes in the auditory-only condition. Whole-brain auditory phoneme power analyses were calculated using NeuroPower (http://neuropowertools.org/neuropower/neuropowerstart/).

Results demonstrated 80% power estimates at the following sample sizes for each of the three comparisons (using random field theory, cluster threshold $p = .05$, alpha = .05, $n = 24$):

/fafa vs /mama n = 64, /fafa vs /kaka n = 62, /kaka vs /mama n = 63. On the assumption that equivalent visual stimuli would yield similar magnitude effect sizes, we used the maximum sample estimated by these comparisons: n = 64. Using this estimated sample size, we pre-registered the number of subjects to be scanned and analyzed (https://osf.io/6fzwd/). As part of the pre-registration process, we also specified the ROIs to be analyzed and chose four specific ROIs: pSTG, pSTS, fusiform and hMT+

4.2.3 Measured Behavioral Variable and data exclusion criteria

To ensure that the participants paid attention during the task, we set exclusion criteria to remove participants with behavioral accuracy rates less than 75% for either auditory or visual conditions: no participants were excluded based on this cutoff.

4.2.4 fMRI data collection

Subjects were scanned in a GE Discovery MR750 3.0 Tesla scanner with a Nova 32 channel standard adult-sized coil (Milwaukee, WI). One high-resolution T1-weighted structural image was obtained for each participant that was used in preprocessing, flip angle = 8, FOV = 25.6 mm, slice thickness = 1 mm, 256 slices. Then, for each of the five runs, functional T2*-weighted BOLD images were obtained using a multiband gradient-echo, echo planar imaging sequence with a resolution of 2.4 x 2.4 x 2.4 mm^3, TR of 800 ms and, TE of 30 ms, Flip Angle of 52, for a total of 644 3D volumes of the whole brain with a FOV of 216 mm. To account for signal saturation, the task did not start until the first 10 TRs were acquired and discarded by the scanner in each run.

4.2.5 Data Processing

fMRI data was reconstructed with realignment and fieldmap correction applied using SPM12 to each of the five T2* runs for inhomogeneity recovery of signal in the B0 field. Physiological noise was removed using RETROICOR (Glover at al., 2000). For both the univariate and multivariate analysis, preprocessing steps were completed using SPM12 (Wellcome Department of Cognitive Neurology, London, UK). For all of the post-processed multivariate analysis including searchlight and ROI based decoding, we utilized The Decoding Toolbox (https://sites.google.com/site/tdtdecodingtoolbox/) version 3.997.

4.2.6 Preprocessing

Before preprocessing the functional images, SPM's display tool was used to set the origin of the anatomical volumes for each subject manually by picking the location of the anterior commissure. After this, functional volumes were reconstructed and realigned, physiological noise was removed, and field map correction was applied. This was followed by slice time correction to account for acquisition time differences between slices for each of the whole brain functional volumes. This data was then co-registered to the subject's anatomical space using a 4th degree B-spline, followed by segmentation of the tissues from the anatomical image with a forward deformation field. Information generated during the segmentation process was then used to transform the co-registered functional volumes into the standard MNI anatomical space with isotropic voxel volume dimensions of 2mm. The normalized data was then spatially smoothed using a full-width half maximum (FWHM) kernel of 5mm.

4.2.7 Univariate Analysis

We performed a univariate, contrast-based analysis of auditory-only phonemes (averaged across the 3 phonemes) and visual-alone visemes (averaged across the 3 visemes) in order to identify the regions that demonstrate significantly different activation patterns across stimulus types. We utilized a canonical hemodynamic response function with event duration set to 2 seconds for each of the phonemes (AuditoryFA + AuditoryKA + AuditoryMA) and visemes VisualFA + VisualKA + VisualMA) and 5.5 seconds for the fixation periods (Fixation). Event onsets times were defined as the moment when the fixation cross (for auditory trials) or face (visual trials) appeared on the screen.

In the first level analysis, whole brain beta maps were generated individually for all seven conditions for each of the 64 subjects. These maps also included information from regressors for motion correction (six head movement parameters). In the second level, contrasts were defined in the model estimation stage for each of the conditions of interest ([AuditoryFA + AuditoryKA + AuditoryMA > Fixation] and [VisualFA + VisualKA + VisualMA > Fixation]) to calculate the averaged main effects of the phonemes and visemes across all subjects. These averaged contrast estimates were then z-scored to compute the t-statistic at each voxel. Significant voxels were then reported at a threshold of $p < 0.0001$ family-wise error corrected (FWE) for multiple comparisons at the voxel level, with a cluster-correction size of 100 voxels.

4.2.8 Multivariate analysis

To identify regions that reliably differentiate classes of phonemes and classes of visemes, we performed searchlight based MVPA analyses. Preprocessing steps for univariate and multivariate analyses were matched except for the normalization and smoothing, such that for the multivariate analysis, these two steps were performed after the first level analysis was

completed. For the decoding analysis, we utilized The Decoding Toolbox (Hebart, Görgen, & Haynes, 2015) with a LIBSVM (Chang & Lin, 2011) based support vector machine (SVM) implementation. For each of the individual subjects, we built a SVM classifier with a cross-validation scheme for the five runs. We used these classifiers to build two separate models: one to classify between the three phonemes and the other to classify between the three visemes. The phoneme models were constructed to identify voxels that reliably decoded the identity of each of the three phonemes while the viseme models were built to identify voxels that reliably decoded the identity for each of the three visemes. These models were implemented as independent whole-brain searchlight analyses in the first level of the MVPA model. For each of the models, beta estimates were calculated and extracted from a 3-voxel radius sphere. 80% of the data from each run was used for training while the remaining 20% was used as the testing set. The searchlight center was shifted through voxel-wise patterns throughout the brain to extract whole-brain accuracy maps for auditory-only and visual-alone conditions.

Auditory-only and visual-alone mean accuracy maps were then transformed into an map of differences between accuracy (VolAcc) and chance (33%) at each voxel. These volumes (VolAcc-Chance) were then normalized to the MNI space through the individual subject anatomical volumes, similar to the first-level preprocessing step described in the univariate analysis. Following normalization, volumes were smoothed using a FWHM kernel of 5mm. These smoothed volumes were then used to estimate average decoding accuracy across all the subjects. Average decoding volumes were then z-scored to compute the t-statistic at each voxel and significant voxels were reported at a cluster corrected threshold of $p < 0.0001$, with a cluster-correction size of 100 voxels.

4.2.9 ROI based decoding analysis

Following the whole-brain searchlight analysis, we selected four regions of interests (ROI) based on results from literature (Beauchamp et al., 2004, Beauchamp et al., 2010, Yi et al., 2017). Three out of the four ROIs (pSTG, pSTS, and hMT+) were chosen based on our pre-registration at osf.io. The fourth preregistered ROI (fusiform region) did not provide sufficient evidence for a difference in decoding between phonemes and visemes. Hence, we chose to exclude this region from the final analyses and replace it with an early visual cortex ROI (V1 and V2). The four regions were chosen based on previous works (Mesgarani & Chang, 2012, Gyol Yi et al., 2019, Beauchamp et al., 2004a, Beauchamp et al., 2010, Karthik et al., 2021) that implicated them in encoding phonemic representation of phonemes during visual speech. The ROI decoding procedure was similar to the process described in Section 4.2.8 with the only difference being that analyses were restricted to the individual ROI under consideration. Each of the ROIs were analyzed in both the left and right hemispheres. The masks for the regions were created for each individual subject using Freesurfer anatomical labels generated during the recon-all procedure. The pSTG ROI was generated by dividing the "superiortemporal" label of the Freesurfer parcellation into three equal parts and choosing the posterior most label (Karthik et al., 2021). The pSTS ROI was generated using the "bankssts" label, while the medial temporal ROI was created using the "MT_exvivo.thresh" label. The visual ROI was created by merging the "V1_exvivo.thresh" and "V2_exvivo.thresh" labels.

Once individual ROI decoding accuracies were obtained, accuracy rates were calculated for each subject using their confusion matrices. Using these individual accuracy rates, average decoding accuracy rates across all subjects for each ROI was calculated and submitted to one-sample t-tests.

4.2.10 Conjunction Analysis

To investigate if viseme activity targets phoneme specific neurons in the auditory cortex, we performed a conjunction analysis by overlaying regions from the MVPA decoding analysis that showed significant decoding accuracy for auditory-only stimuli with regions that showed significant decoding accuracy for visual-only stimuli. As an initial step of this analysis, we performed a volumetric voxel count to quantify the proportion of voxels that were significant in the decoding vs. univariate analyses. This would provide a coarse measure of how extensive phoneme/viseme information extends within active regions of the STG. To further understand the spatial decoding patterns, we performed a slice-wise analysis of decoding accuracies in the STG. For this analysis, we created a mask using voxels that had significant decoding accuracy in the auditory-only and visual-only conditions. Since a visual analysis of the slices showed significant overlap between the two conditions only in slices 172 to 189 (MNI x-coordinate), we utilized only these 9 slices in our analysis. In each of these slices, we calculate the average decoding accuracy from posterior to anterior (MNI y-coordinate) along the MNI z-axis.

4.2.11 Multivariate Similarity Analysis

In addition to presenting univariate analysis activations and decoding decoding accuracies for each of the stimulus types, we also report similarity measures from the contrast estimates across each pair of stimuli. For this analysis, we calculated the similarities of the beta estimates obtained from the contrast maps for each pairwise stimulus across all six stimuli pairs (AuditoryFA-Fixation, AuditoryKA-Fixation, AuditoryMA-Fixation, VisualFA-Fixation, VisualKA-Fixation, VisualMA-Fixation). For this, we begin by extracting the beta estimates from all the voxels in the left pSTG from the contrast maps generated through the procedure described in section 3.2.7 separately for each subject. We then calculated the pairwise

similarities for each of the six pairs of stimuli as the Spearman correlation estimate (r) between the beta estimates of their contrast maps in the left pSTG. To investigate the differences in representations between like-phoneme (AuditoryFA: VisualFA, AuditoryKA: VisualKA, AuditoryMA: VisualMA), and unlike-phoneme pairs of the auditory and visual stimuli, we performed a paired t-test between the diagonal elements of the lower left quadrant of the similarity matrix shown in Figure 34 representing like-phoneme pairs and its off-diagonal elements representing unlike-phoneme pairs (random effect = participant). The final similarity matrix is obtained as the mean of *r* across all subjects in each cell of the pairwise similarity matrix.

4.3 Results

4.3.1 Behavioral results

The overall accuracy of behavioral responses in participants was 93.99% (n = 64, SD = 3.15%), with a mean accuracy of 95.67% (n = 64, SD = 3.01%) in the auditory-only condition and a mean accuracy of 92.31% (n = 64, SD = 3.72%) in the visual-alone condition. As expected, the mean accuracy in the auditory-only condition was significantly higher than the visual alone condition; $t(63) = 6.57$, $p < 0.0001$, Cohen's $d = 0.96$. None of the 64 participants performed below the pre-registered exclusion threshold (accuracy in either condition below 75%).

4.3.2 Imaging results overview

The primary goal of this study was to test the hypothesis that visual speech provides phonemic information to auditory regions (the pSTG and pSTS). We first replicated prior observations that silent visual speech activates auditory areas using a whole brain univariate

analysis. Next, to examine whether viseme information is spatially encoded in auditory regions, we used whole brain MVPA decoding to identify regions at which individual phonemes and visemes could be reliably classified. To further measure the relative strength of phoneme and viseme representations, we used MVPA decoding within *a priori* set ROIs. Finally, to test whether visemes and phonemes produced shared representations within the STG, we used multivariate similarity analysis.

4.3.3 Univariate contrast analysis

For the univariate analysis, we utilized contrast estimates calculated through parametric t-tests comparing our stimuli conditions of interest against a fixation period. Figures 28 and 29 show fMRI-BOLD activations in the auditory-only and visual-alone conditions respectively. All the values are reported at a cluster correction threshold of 100 voxels with the individual p-values corrected for multiple comparisons using family-wise error rate values of $p < 0.001$. From Figure 28, we see that in the auditory-only condition, contrast estimates between BOLD signals from phonemes and fixation revealed peak activations in the right pSTG (MNI: x=52, y=-12, z=6, $t(63) = 15.25$, $p < 0.001$) and left pSTG (MNI: x=-54, y=-18, z=4, $t(63) = 15.03$, $p < 0.001$). From Figure 29, we see that in the visual-alone condition, contrast estimate revealed peak activations in the right pSTG (MNI: x=52, y=-12, z=6, $t(63) = 15.25$, $p < 0.001$), right fusiform (MNI: x=42, y=-46, z=-18, $t(63) = 12.43$, $p < 0.001$), right occipital/v1 (MNI: x=20, y=-94, z=0, $t(63) = 12.16$, $p < 0.001$) , left occipital/v1 (MNI: x=-20, y=-98, z=4, $t(63) = 12.32$, $p < 0.001$), and the right pSTS (MNI: x=54, y=-38, z=10, $t(63) = 11.38$, $p < 0.001$). Table 7 shows the peak activations in each of the regions and their corresponding cluster-sizes along with the MNI coordinates with the maximum t-statistic.

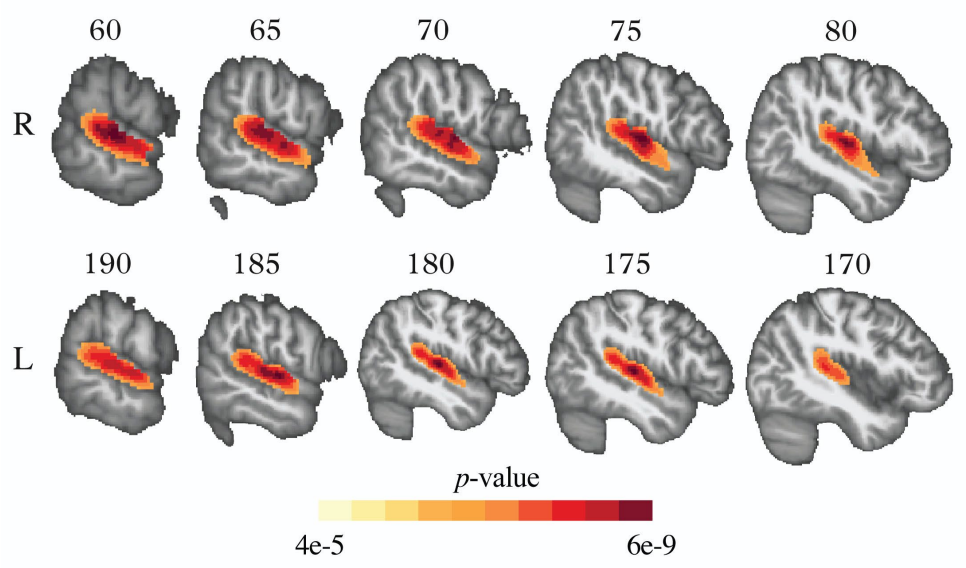


Figure 28. ***Univariate analysis in the auditory condition****. Univariate analysis comparing contrast estimates between phonemes in the auditory-only stimuli condition and the fixation period. All results are reported at a cluster correction threshold of 100 voxels with the individual p-values corrected for multiple comparisons using family-wise error rate values of $p < 0.001$. Peak t-values were seen in both right and left pSTG, apart from broad activity noticed bilaterally in the STG.*

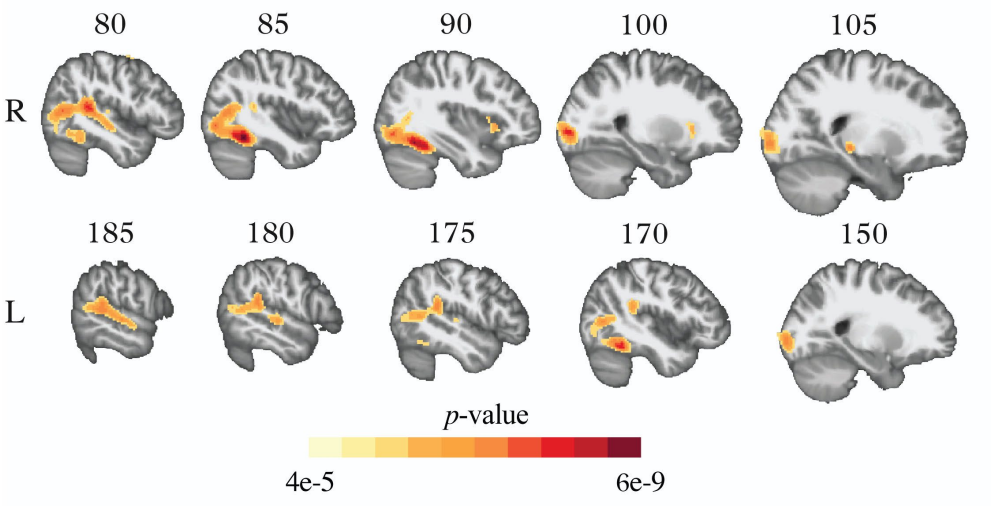


Figure 29. ***Univariate analysis in the visual condition****. Univariate analysis comparing contrast estimates between phonemes in the visual-only stimuli condition and the fixation period. All results are reported at a cluster correction threshold of 100 voxels with the individual p-values corrected for multiple comparisons using family-wise error rate values of $p < 0.001$. Peak t-values were seen in the right pSTG, right fusiform, right occipital/V1, left occipital/V1 and the right pSTG and right pSTS.*

Table 7. ***MNI coordinates and cluster sizes in the univariate analysis****. MNI coordinates and cluster sizes of peak t-values calculated in the univariate analysis for the auditory-only and visual-alone stimuli conditions.*

Condition	Brain region	Peak MNI coordinates X,Y,Z	$P_{FWE\text{-}corr}$	$t(63)$	K_E
Auditory	Right pSTG	52,-12,6	<0.001	15.25	1962
		66,-24,8	<0.001	14.67	
		64,-32,10	<0.001	13.34	
	Left pSTG	-54,-18,4	<0.001	15.03	1816
		-64,-30,8	<0.001	12.69	
		-50,-10,-2	<0.001	12.63	
Visual	Right pSTG	52,-12,6	<0.001	15.25	1962
		66,-24,8	<0.001	14.67	
		64,-32,10	<0.001	13.34	
	Right Fusiform	42,-46,-18	<0.001	12.43	1520
		38,-56,-14	<0.001	11.50	
	Right Occipital/V1	20,-94,0	<0.001	12.16	1520
	Left Occipital/V1	-20,-98,4	<0.001	12.32	237
	Right pSTG/s	54,-38,10	<0.001	11.38	355

4.3.4 MVPA decoding analysis

Decoding was conducted separately for each of the two conditions of interest (phonemes and visemes). Initially, whole-brain exploratory decoding was performed using an SVM-based decoder. The results of this analysis can be seen in Figures 30 and 31 for the auditory-only and visual-alone conditions respectively. For the auditory-only condition, peak decoding accuracy was seen in the right pSTG (MNI: x=-54, y=-16, z=2, $t(63) = 6.44$, $p < 0.001$) and the left pSTG (MNI: x=-60, y=-34, z=10, $t(63) = 7.25$, $p < 0.001$). For the visual-alone condition, peak decoding accuracy was seen in the right occipital/v1 (MNI: x=28, y=-94, z=4, $t(63) = 6.56$, $p < 0.001$), right hMT+ (MNI: x=50, y=-70, z=8, $t(63) = 4.95$, $p < 0.001$), left occipital/v1 (MNI: x=-16, y=-98, z=2, $t(63) = 5.30$, $p < 0.001$), left pSTG/s (MNI: x=-54, y=-36, z=12, $t(63) = 5.10$, $p <$

0.001) and left MTG (MNI: x=-60, y=-16, z=-2, $t(63) = 4.69$, $p < 0.001$). All p-values are reported after thresholding at $p < 0.001$ with a cluster correction threshold of 100 voxels. Table 8 shows the p-values for peak decoding accuracy in each of the regions and their corresponding cluster-sizes along with the MNI coordinates and their maximum t-statistic.

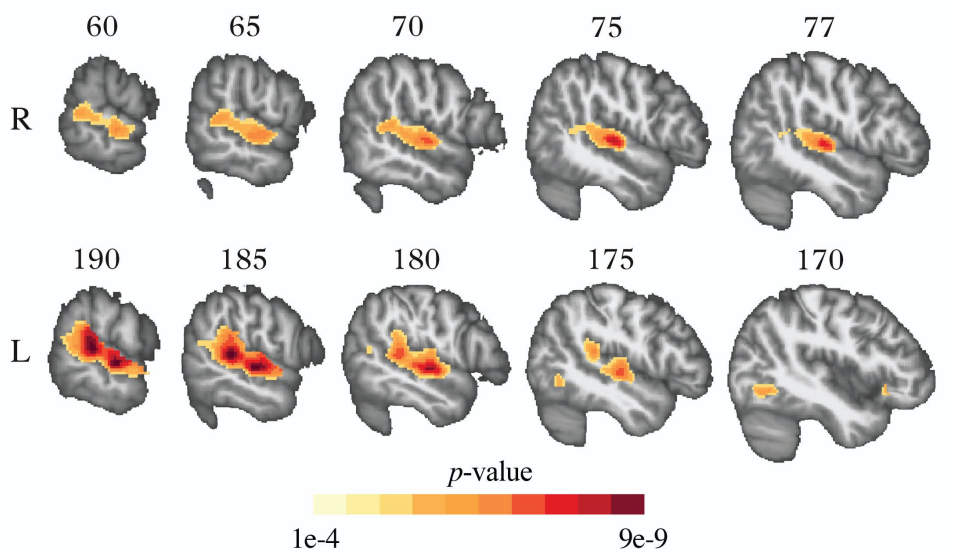


Figure 30. ***MVPA analysis in the auditory condition****. MVPA analysis representing decoding p-values for differentiability between different phonemes in the auditory-only stimuli condition. All results are reported at a cluster correction threshold of 100 voxels. Peak t-values were seen in both right and left pSTG, apart from broad activity noticed bilaterally in the STG.*

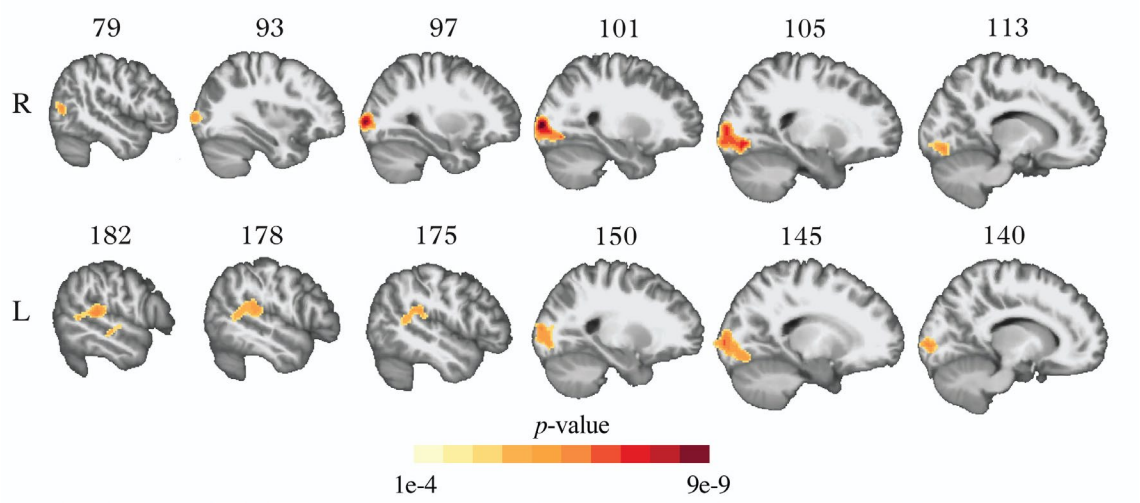


Figure 31. ***MVPA analysis in the visual condition.*** *MVPA analysis representing decoding p-values for differentiability between different phonemes in the visual-alone stimuli condition. All results are reported at a cluster correction threshold of 100 voxels. Peak t-values were seen in right occipital/V1, left occipital/V1, right hMT+, left pSTG, left pSTS, and left MTG.*

Table 8. ***MNI coordinates and cluster sizes in the MVPA analysis*** *MNI coordinates and cluster sizes of peak t-values calculated in the MVPA analysis for the auditory-only and visual-alone stimuli conditions*

Condition	Brain region	Peak MNI coordinates X,Y,Z	$P_{FWE\text{-}corr}$	$t(63)$	K_E
Auditory	Right pSTG	54,-16,2	<0.0001	6.44	1266
		64,-18,0	<0.0001	5.24	
		72,-20,-2	<0.0001	5.16	
	Left pSTG	-60,-34,10	<0.0001	7.25	2410
		-60,-18,0	<0.0001	7.13	
		-66,-36,18	<0.0001	6.86	
Visual	Right Occipital/V1	28,-94,4	<0.0001	6.56	1112
		20,-80,-10	<0.0001	5.52	
	Right hMT+	50,-70,8	<0.0001	4.95	182
	Left Occipital/V1	-16,-98,2	<0.0001	5.30	795
		-26,-88,4	<0.0001	5.02	
		-16,-76,-12	<0.0001	4.86	
	Left pSTG/s	-54,-36,12	<0.0001	5.10	475
		-48,-36,22	<0.0001	4.71	
		-50,-46,12	<0.0001	4.52	
	Left MTG	-60,-16,-2	<0.0001	4.69	115
		-54,-22,-6	<0.0001	4.34	

4.3.5 MVPA ROI Decoding analysis

To validate our hypothesis about phonemic and visemic information being encoded in specific regions of interest, we performed a decoding analysis in the pre-registered ROIs. We observed that the auditory-only decoding accuracy was above chance in two of the four ROIs in the right hemisphere, including right pSTG (M = 41.32%, $t(63) = -4.51$, $p < 0.0001$) and right pSTS (M = 38.33%, $t(63) = -2.52$, $p = 0.007$). We also observed that the auditory-only condition showed significantly above chance decoding in three of the four ROIs in the left hemisphere; including left pSTG (M = 40.62%, $t(63) = -3.68$, $p < 0.0001$), left pSTS (M = 40.75%, $t(63) = -4.08$, $p < 0.0001$), and left hMT+ (M = 37.65%, $t(63) = -2.38$, $p = 0.01$)

The visual-alone condition showed significant decoding accuracy in two of the right hemisphere ROIs: hMT+ (M = 39.63%, $t(63) = -4.25$, $p < 0.0001$) and occipital/v1 (M = 37.86, $t(63) = -2.38$, $p = 0.01$). The visual alone condition showed significant decoding accuracy in all four left hemisphere ROIs: left pSTG (M = 40%, $t(63) = -3.26$, $p < 0.0001$), left pSTS (M = 39.68%, $t(63) = -3.23$, $p < 0.0001$), left visual (M = 40.36%, $t(63) = -4.16$, $p < 0.0001$), and left hMT+ (M = 36.77%, $t(63) = -2.11$, $p = 0.01$). The ROIs considered (overlaid on a MNI template) and their respective decoding accuracies are shown in Figure 32.

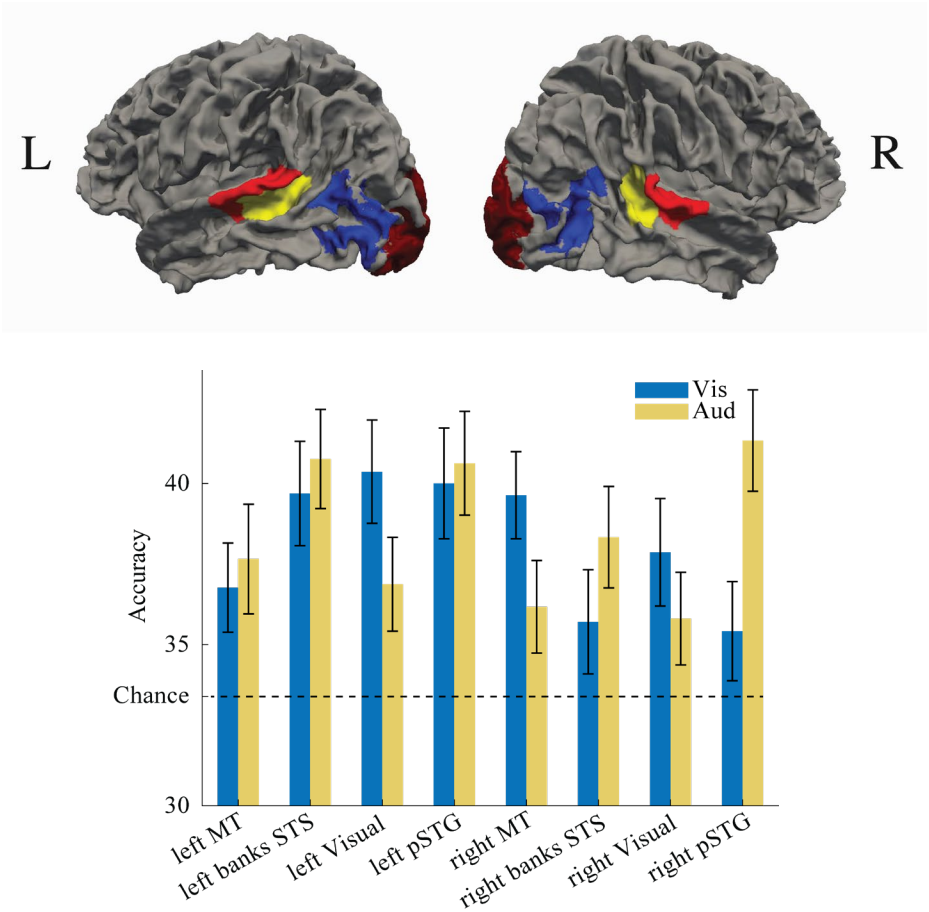


Figure 32. ***ROIs and their decoding accuracies****. The ROIs considered for investigating the MVPA decoding accuracies for the auditory-only and visual-alone stimuli conditions. Bilateral ROIs were considered for the pSTG (highlighted in red), banks of STS (highlighted in yellow), MT (highlighted in blue), and V1 (highlighted in maroon). All the ROIs showed statistically above chance decoding accuracies bilaterally. While the right pSTG and pSTS showed significantly higher decoding accuracies in the auditory-only stimuli condition compared to visual-only stimuli condition, the left V1, right V1 and right MT showed significantly higher decoding rates in the visual-only stimuli condition compared to the auditory-only stimuli condition.*

4.3.6 Conjunction analysis

The univariate analysis and decoding analysis individually revealed information about voxels that responded to phonemes and visemes, as well as voxels that had above-chance decoding accuracy for each of the stimuli conditions. To understand the distribution and overlap of the voxels that overlapped in each of the analysis, we performed a volumetric voxel count of the overlap between voxels in the two stimuli conditions. This analysis revealed that out of 22032 voxels in the STG that responded significantly to auditory only stimuli in the univariate analysis, 15905 voxels (72.19%) had above chance accuracy in the decoding analysis. Similarly, out of 6221 voxels in the STG that responded significantly to visual only stimuli in the univariate analysis, 612 voxels (9.83%) had above chance accuracy in the decoding analysis. These results indicate that a high percentage of voxels that respond to phonemes in the STG also encode phonemic representations in the STG. Conversely, only a minority of voxels that respond to visemes in the STG also encode visemic representations in the STG.

Table 9. ***Overlap between univariate and MVPA voxels****. Number of voxels in the STG that were significant in the univariate analysis and their overlap with the voxels in the decoding analysis.*

Stimuli condition	**Satistically significant Univariate voxels in STG**	**Decoding voxels that overlapped with univariate voxels in the STG**	**Overlap percentage**
Auditory-only	22032 voxels	15905	72.19%
Visual-only	6221 voxels	612	9.83%

In the decoding only analysis, 17070 voxels in the STG had above chance decoding accuracy in the auditory-only condition. 2576 voxels in the STG had above chance decoding accuracy in the visual only condition. 2479 voxels in the STG had above chance decoding accuracy in both the auditory only and visual only conditions.

Table 10. ***Number of voxels that had above-chance decoding accuracy in the STG*** *for phonemes and visemes.*

Decoding analysis in the STG	**Voxels with above-chance accuracy for phonemes**	**Voxels with above-chance accuracy for visemes**	**Voxels with above-chance accuracy for both phonemes and visemes**
	17070	2576	2479

The above conjunction analysis revealed that visemes targeted a unique population of neurons in the left pSTG/s. The results from this seen in Figure 33 show that viseme activity indeed targeted spatially unique neuronal populations in the STG that did not provide significant decoding accuracies to phoneme activity. These results support the visual evidence showing a unique population of neurons in the posterior regions of the left pSTG/s that encode viseme information and not phoneme information.

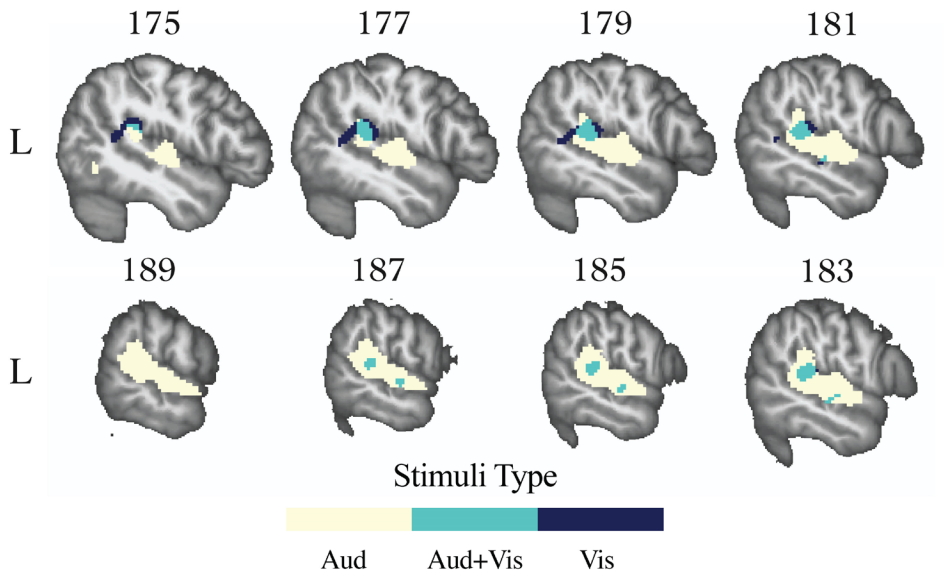


Figure 33. ***Conjunction analysis****. Slices showing voxels that had significant above chance decoding in the auditory-only stimuli condition (highlighted in yellow), visual-only stimuli condition (highlighted in blue), and voxels that had significant above chance decoding in both the auditory-only and visual-only stimuli conditions (highlighted in cyan). We observe that visemes target population of neurons that did not encode phonemic information in regions of the pSTG and pSTS.*

4.3.7 Multivariate similarity analysis

While the decoding analyses provide information about which regions of the brain encode the identities of individual phonemes and visemes, it is not possible to directly investigate similarities between how these phonemes and visemes are represented in these regions. For example, an examination of the spatial and temporal (dis)similarities for phonemes vs visemes would aid in the interpretation of how visemic identities are transformed and encoded in the auditory regions.

Figure 34. ***Multivariate similarity analysis*** *showing the similarity representations in contrast estimates between individual phonemes and visemes in the left pSTG.*

Hence, to understand the similarity in representation between each of the phonemes and visemes in the left pSTG, we performed a multivariate similarity analysis. This analysis revealed that the similarity in representation of phonemes (upper left-quadrant in Figure 34, mean r = 0.90) was significantly higher than the similarity in representation of visemes (lower right-quadrant in Figure 34, mean r = 0.87), at $t(63) = 3.26$, $p < 0.001$. A t-test between the similarity measures of like-phoneme/viseme pairs (VisualFa:AuditoryFa+ VisualKa:AuditoryKa+

VisualMa:AuditoryMa) and unlike-phoneme/viseme pairs (VisualFa:AuditoryKa + VisualFa:AuditoryMa + VisualKa:AuditoryFa + VisualKa:AuditoryMa + VisualMa:AuditoryFa + VisualMa:AuditoryMa) revealed a significant difference between representations of like-phoneme/viseme pairs (mean $r = 0.72$) and unlike-phoneme/viseme pairs (mean $r = 0.69$), at $t(63) = 3.18$, $p = 0.001$. This indicated that identities of phonemes and visemes have similar representations in the left pSTG.

4.4 Discussion

Several studies have shown that silent visual speech can activate neuronal populations in the primary auditory areas (Beauchamp et al., 2004, Beauchamp et al., 2010, Mesgarani & Chang., 2012, Gyol Yi et al., 2019, Karthik et al., 2021).

However, what information is represented by this activation has remained an open question. Here we tested the hypothesis that some of this activity reflects the transfer of phonemic information from visual speech into auditory areas, potentially targeting corresponding phoneme neurons in the STG.

This question is important because fMRI activity evoked by silent visual speech in auditory areas may reflect a variety of processes and information types, including motion timing information (McGrath & Summerfield, 1985), speech rate (Chandrasekaran et al., 2009), spectral information (Plass et al., 2020), general effects on attention or arousal (Schroeder et al., 2008), or as we sought to investigate, viseme-to-phoneme transformations (Karthik et al., 2021). It has also previously been observed that silent lip-reading activates the auditory cortices reflecting fast synthesis of the auditory stimulus (Bourguignon et al., 2020). In this study, we utilized an MVPA based information decoding approach and multivariate similarity analyses to study the informational content encoded in auditory areas during a silent lip-reading task.

Participants demonstrated high accuracy during the behavioral task, indicating that the auditory and visual stimuli used were suitable for testing in an fMRI environment. In a univariate analysis, we observed that visemes produced a distributed network of activation that included auditory, visual, and classical multisensory regions (pSTS). Followed by this, we performed whole-brain searchlight MVPA analyses to quantify where phonemic and visemeic information was encoded. Replicating past research in the auditory domain, we observed significant decoding of individual phonemes from the STG bilaterally. Consistent with our predictions, significant decoding of individual visemes was also observed from the STG, along with other canonical visual areas (V1/V2 and hMT+) and multisensory regions (pSTS). This result demonstrates that visual speech information is represented within the auditory system. ROI-based analyses revealed a similar pattern of results and highlighted that maximal viseme decoding was present in visual cortex at almost equal accuracy to that in the the STG and pSTS.

The results obtained here indicate that visual speech indeed encodes phonemic information in the primary auditory areas including left pSTG/s and left MTG apart from the motion sensitive right hMT+ region. Moreover, we also noticed that though there was strong overlap between neuronal populations that encoded phoneme information and populations that encoded viseme information, there was a unique population of neurons in the pSTG/s area that encoded purely viseme information. This indicates that phonemic information in visual speech targets an independent set of neurons in regions that have conventionally been associated with multisensory integration of auditory and visual speech (Beauchamp et al., 2004a, Beauchamp et al., 2010, Karthik et al., 2021).

The targeting of an independent set of neurons by visemes is interesting given that there are phonemes that could sound unique and can be differentiated easily in the auditory modality

(e.g., the words ‘pet’ and ‘bet’) with a corresponding viseme that cannot be differentiated without the underlying sound. Hence, it can be putatively hypothesized that audiovisual integration targets a unique population of neurons that have a probabilistic distribution with what kind of sensory modality they encode or respond to. This becomes more relevant when seen in relation to conditions where visual information can alter phonemic perception, such as in the McGurk effect (McGurk & McDonald, 1978).

We also studied this hypothesis through a multivariate similarity analysis to investigate whether viseme spatial patterns overlap with phoneme spatial patterns. If they do, they would target the same neurons with similar information encoded for identical phoneme/visemes pairs (e.g., auditory /ba/ and visual /ba/). This analysis revealed that there was similar information encoded in the left pSTG, with like phoneme/viseme pairs eliciting significantly different activation levels compared to unlike phoneme/viseme pairs. This indicates that along with the presence of a unique informational hierarchy in the multisensory regions about the type of modality the neurons respond to (Karthik et al., 2021), there also exists a hierarchy of information encoding with individual population of neurons selectively responding to and encoding phonemic information about the speech content.

These results provide strong support to complementary studies that show a strong entrainment of cortical activity in the auditory areas during visual speech (Bourguignon et al., 2020). Our results also provide opposing evidence to previous studies that hypothesized that activations in auditory areas during a silent lip-reading task might reflect imagery information that is unrelated to the spoken speech (Bernstein and Liebenthal, 2014). This evidence is also strengthened by the fact that there exist unique neuronal populations that encode phonemic information about visual speech in the multisensory regions. This would provide directions for

future investigation where this evidence can be analyzed considering previous studies which indicate that phonemic information in visual speech is obtained from the visual areas (Hauswald et al., 2018).

Taken together, these results suggest that during visual speech, phonemic information causes activation patterns in the audiovisual multisensory regions widely reported in literature. They also indicate that there exist visual-speech specific neuronal populations in these regions that exclusively encode information from visual speech, providing a framework for further investigation to implore into the nature of this dissociation. The absence of a statistically significant difference in decoding accuracy between phonemes and visemes in all the multisensory ROIs examined, indicate that these regions encode equally probable phonemic information from both the auditory and visual modalities. The validity of our results is strengthened by the fact that all the analysis were preregistered, and the sample size chosen through a power analysis.

Chapter 5 Summary, Limitations and Future Research

Face-to-face verbal communication is an important aspect of social behavior in humans. The visual component in audiovisual speech both facilitates and enhances auditory speech perception (Sumbly & Pollock., 1954, Grant & Seitz, 2000). However, it is unclear how visual cues contribute to these effects and what multisensory information they provide. In this dissertation, I investigated neural processes in auditory cortex that subserve the perceptual benefits of visual information during audiovisual speech perception. Across three studies, I used a multimodal fMRI-iEEG approach to investigate audiovisual speech processes and provided evidence for the multiple ways in which visual speech modulates and encodes information in the major auditory areas of the human brain.

In Study 1 (Karthik et al., 2021), I showed that audiovisual speech integration elicits multiple distinct patterns of neural activity within the STG and adjacent cortices. These processes were shown to occur in multiple ways across different oscillatory bands in which the brain functions, and at different temporal scales across multiple regions of the auditory cortex. This study demonstrated that visual modulation of auditory speech processing is not a unitary phenomenon, but rather consists of multiple functionally distinct processes. Past studies investigating audiovisual speech integration have analyzed iEEG data using single-participant designs with fixed-effect statistics, making it hard to generalize the findings to the group-level and thus to the general population (Micheli et al., 2020; Besle et al., 2008; Plass et al., 2020). Even while using variants of group-level analysis such as linear mixed-effects modeling,

previous studies (Ozker et al., 2017; Ozker et al., 2018) have focused on HGp, which indexes local population firing rates, ignoring low-frequency oscillations which potentially reflect distinct audiovisual information. I addressed both shortcomings in this study by using linear mixed effects models to analyze iEEG signals at multiple frequency bands (theta band, beta band and high gamma power). This helped uncover multiple distinct processes that occur in visual modulation of auditory speech processing. Additionally, as a novel contribution, I also showed the effectiveness of utilizing a linear mixed effects model in the group-level analysis of iEEG signals obtained from a large cohort of subjects (n = 21).

In Study 2, I built on results from Study 1 by investigating the underlying information encoded in the neural processes involved in visual modulation of audiovisual speech perception. Previous studies have shown strong entrainment of cortical activity in the auditory areas during visual speech (Bourguignon et al., 2020). Using machine-learning based approaches, I showed that visual speech not only modulates the neural processes in auditory cortex, but also encodes the identity of lipread visemes. This result is of particular interest since information encoding in auditory areas during visual speech is yet to be fully understood. Critically, this is the first study in literature to show that it is possible to identify phonemic information from neural activity in the auditory cortex during visual speech. I also showed that the phonemic representations of both auditory and visual speech in auditory cortex have a statistically significant correlation. This signifies that viseme information targets phoneme populations in the auditory cortex during audiovisual speech processing.

In Study 3, I acquired fMRI from a large cohort of participants (n = 64) to replicate and extend my findings from Studies 1 and 2 in a non-patient population. The benefits of this study are twofold. It allowed me to 1) extend the results obtained from an epileptic sample to a larger

cohort of normative population, and 2) confirm the robust iEEG results with fMRI data, which has better generalizability and provides superior spatial resolution. Results from this study also provided novel evidence that there are distinct areas of the auditory cortex that are targeted selectively by information from visual speech input. The results from my study complement past studies in literature where there is mounting evidence indicating that phonemic content in visual speech can be transformed into categorical phonemic units in the multisensory regions of auditory cortex (Beauchamp et. al., 2010). While past studies focused on localizing the cortical regions involved in these transformations, I extend these findings by demonstrating the information content encoded in the neuronal populations of these cortical regions during auditory-visual speech processing.

These results provide concrete evidence to the fact that viseme information is used to modulate or prime associated populations of phoneme-sensitive neurons in the auditory areas including the STG. This could in turn enable visemes to bias, modulate or influence auditory speech processing. As argued in literature before (Magnotti & Beauchamp., 2017), this would likely occur through a "winner-take-all" mechanism such that what is heard/perceived depends on the maximal representation of the specific population of neurons. This could also potentially explain the McGurk effect where the neuronal activity disagrees between the distribution of phonemes and visemes. This might lead to the creation of a spatial pattern that matches an unrelated third phoneme. Another interesting observation from these results is that visual information encoding was not observed in the high gamma power, reflecting potential subthreshold effects. This is important since large amounts of evoked action potentials in the auditory areas in response to visual speech could lead to auditory hallucinations (or synesthetic experiences).

Discussion

The results from these studies lead us to an understanding of the neural bases of audiovisual speech perception. More specifically, we understand that the auditory areas of the brain utilizes information from visual speech not only to alter the ways in which auditory speech processing is performed, but also encodes phonemic information about visual speech in these regions. We also showed that the auditory areas not only encode phonemic information about visual speech, but the representations of these information are highly similar to phonemic information from auditory speech. All of these evidences point to the fact that audiovisual speech processing actively utilizes information from both auditory and visual speech signals and these signals are transformed into similar type of representations while they are processed in the multisensory regions of the auditory cortex. This lends support to a converging idea in literature which posits that phonemic information about visual speech undergoes categorical transformations in the auditory areas (Mesgarani et al., 2014).

Indeed, there could be conflicting information provided by visual speech signals during audiovisual speech processing. For instance, while auditory speech has unique phonemic representations for every sound processed in the auditory areas, visual information need not have a unique representation for every visual input. One such example could be the visemes /ba/ and /ma/ that have unique auditory or phonemic representations, but similar visemic representations. This could lead to a many-to-one mapping in the auditory regions. These type of ambiguous visual information would have us argue that while auditory speech perception can be greatly enhanced with additional information from visual speech, the effect of these enhancements in the context of natural speech is yet to be fully investigated.

Arguments could also be made about the effect of internal vocalization on the neural responses seen in the auditory areas during visual speech (such as during a silent lip reading task). But results from all three studies provide evidences contrary to this argument due to two related observations from the data. 1) In studies 1 and 2, we see that the ERPs evoked during a visual-only task follows a pattern where the activation arises immediately after the onset of stimuli. This would mean that the processes enabling the activation patterns do not follow the patterns that would be expected if they were a result of internal vocalization. These patterns also closely follow the ERP patterns of auditory-only signals. 2) In study 3, we notice that while the auditory-only and visual-only stimuli evoke activations in the auditory areas of the left STG, only the auditory-only stimuli evoke activations in the right STG. If the activations patterns were caused due to internal vocalization, we would expect similar activations in both the auditory-only and visual-only stimuli conditions.

From these arguments we can safely provide support to indicate that the effect of visual speech in the auditory areas are in fact caused by the processing of these signals by neuronal populations in the auditory regions as opposed to internal vocalization.

These results indicate a preliminary support to the idea that visual speech influences audiovisual speech processing in the auditory areas. But it should be noted that the data do not lead to an understanding of the ways in which natural speech perception occur in everyday settings since the experiments conducted in this dissertation rely on the use of non-naturalistic stimuli. But, we provide concrete evidence that could be used as a basis for further investigation to understand the multiple processes that could subserve the utilization of visual information for perceiving audiovisual speech in everyday settings and naturalistic face-to-face communication.

The understanding of the ways in which the neural bases of audiovisual speech perception happens can assist in developing better hearing-aid technologies that utilize information not only from purely auditory signals, but also from visual speech inputs, thereby providing greater fidelity in hearing experiences for individuals utilizing these aids.

Limitations & Future research

One potential limitation of the presented work is the use of iEEG to investigate audiovisual speech processing. While iEEG has multiple benefits, including superior temporal and spectral properties, the placement of electrodes is an invasive procedure leading to localized inflammation and changes in the brain's physical shape and structure. This can make the generalization of results challenging. Brain inflammation makes finding physical correspondences in electrode location among multiple subjects a computationally complex problem. Furthermore, because electrode placement is dictated by clinical as opposed to research needs, placements vary widely among subjects. It is therefore imperative to find ways to address the issue of generalizability. The studies in this dissertation address this problem with the use of a novel group-level analysis technique and by registering locations of electrodes across multiple participants. My results and technique have also been successfully replicated internally within our group. However, the techniques proposed in my study have yet to be replicated by other groups. Hence, it will be imperative to encourage replication of my proposed technique by other research groups.

Another limitation of iEEG research is that it only involves clinical populations; due to the technique's invasive nature, participants are often those with epilepsy or tumors. The use of an atypical sample may limit the generalizability of results to a normative population.

A further limitation is the relatively small sample size of iEEG studies. While Study 1 had quite a large sample with 21 patients, Study 2's sample only included 4 patients. Though iEEG data is difficult to obtain, and small sample sizes (n<10) are standard in the literature, this presents further concerns surrounding generalizability that need to be addressed by finding ways to incorporate data from multiple tasks and studies. Study 2 does so by integrating data from two types of tasks to expand the size of data available for individual participants. Apart from the small size of data in Study 1, another reason to use a different set of data for Study 2 was that the neuronal representations (as evidenced by the evoked potentials) of the various phonemes utilized were extremely similar to each other in Study 1. This made it difficult to build a classifier that was able to distinguish between the identities of the phonemes used. Hence, for Study 2, I utilized a dataset that had good differentiability between individual phonemes. However, this tactic presents limitations of its own, for instance: how do we expand the size of data if there are no comparable task sets that a participant has performed? This could be a direction for future research. While it might be difficult to increase the number of participants in an iEEG research study, we could explore ways to increase the sample size of data available for each participant by combining comparable task sets that a participant performs.

Across all the studies, no analysis was performed that related neural underpinnings with behavioral performance of the subjects. While this could be a potential shortcoming, the main reason for not investigating this relationship was that the performance of subjects across all the studies consistently remained at ceiling. Any analysis that was performed to investigate neural underpinnings with behavior that was at ceiling would result in underpowered results. One way to address this shortcoming could be to design follow-up experiments that provide performance

levels that capture a wide variance across individual subjects. This can also help in investigating if lower accuracies result in reduced activation levels at the individual subject level.

An additional limitation of both the iEEG and fMRI results is the use of non-naturalistic stimuli. We utilized stimuli that were tailored for event-related neural responses. While these types of stimuli provide extremely robust neural responses that are easier to model, it can be hard to extend these results to a more naturalistic setting. One way to tackle this issue in future work is to utilize more naturalistic stimuli to replicate the results obtained from these studies. A naturalistic stimuli with auditory, visual and audiovisual stimuli would help map out the complete distribution of phoneme and viseme representations in the auditory areas. Since visemes have a many-to-one mapping with respect to phonemes (i.e., the same lip movements could produce different phonemes), a naturalistic stimuli with all three components of speech (audio, visual and audiovisual) could help understand where visual and auditory neuronal populations overlap and where differences in processing might arise.

Though the dissertation touched upon details about information encoded by visual speech in the visual areas, it largely focused on the effects of visual speech in the auditory areas. But, to what extent are these effects in the auditory areas a result of visual speech modulating auditory speech as opposed to the auditory areas internally synthesizing internal speech? While Study 2 provides a starting point by arguing that the modulations are not just internally synthesized speech, (as evidenced by the temporal decoding accuracy patterns in the auditory and visual speech conditions), the sample size (n=4) was small. Future research could include a broader investigation of these results using a larger sample.

Another future direction of this work could include developing multimodal data processing techniques to integrate data from iEEG and fMRI. While similar conclusions were derived from each of these two modalities, the rich set of data available from these complementary imaging techniques could help build a more concrete framework for understanding the neural processes involved in audiovisual speech perception.

Bibliography

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., and Van Der Sluis, S. (2014). A solution to dependency: using multilevel analysis to accommodate nested data. Nature neuroscience, 17(4), 491-496.

Aggarwal, C. C. (2005, August). On k-anonymity and the curse of dimensionality. In VLDB (Vol. 5, pp. 901-909).

Agrawal, Y., Platz, E. A., & Niparko, J. K. (2008). Prevalence of hearing loss and differences by demographic characteristics among US adults: data from the National Health and Nutrition Examination Survey, 1999-2004. Archives of internal medicine, 168(14), 1522-1530.

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295.

Ahn, J., & Lee, J. H. (2018). Clustering algorithm for time series with similar shapes. KSII Transactions on Internet and Information Systems (TIIS), 12(7), 3112-3127.

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. Journal of Neuroscience, 29(43), 13445-13453.

Arnal, L. H., Wyart, V., and Giraud, A. L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nature neuroscience, 14(6), 797.

Atella, V., Piano Mortari, A., Kopinska, J., Belotti, F., Lapi, F., Cricelli, C., & Fontana, L. (2019). Trends in age-related disease burden and healthcare utilization. Aging cell, 18(1), e12861.

Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. The Journal of Machine Learning Research, 18(1), 629-681.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language, 68(3), 255-278.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. Neuron, 76(4), 695-711.

Beauchamp, M. S. (2016). Audiovisual speech integration: Neural substrates and behavior. In Neurobiology of language (pp. 515-526). Academic Press.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., and Martin, A. (2004b). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nature neuroscience, 7(11), 1190.
Beauchamp, M. S., Lee, K. E., Argall, B. D., and Martin, A. (2004a). Integration of auditory and visual information about objects in superior temporal sulcus. Neuron, 41(5), 809-823.

Beauchamp, M. S., Nath, A. R., and Pasalar, S. (2010). fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. Journal of Neuroscience, 30(7), 2414-2417.

Ben-David, S., Pál, D., & Simon, H. U. (2007, June). Stability of k-means clustering. In International conference on computational learning theory (pp. 20-34). Springer, Berlin, Heidelberg.

Berndt, D. J., & Clifford, J. (1994, July). Using dynamic time warping to find patterns in time series. In KDD workshop (Vol. 10, No. 16, pp. 359-370).

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. Frontiers in neuroscience, 8, 386.

Besle, et al., (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. Journal of Neuroscience, 28(52), 14301-14310.

Besle, J., Fort, A., Delpuech, C., and Giard, M. H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. European journal of Neuroscience, 20(8), 2225-2234.

Bleichner, M.G., Jansma, J.M., Salari, E., Freudenburg, Z.V., Raemaekers, M.,Ramsey, N.F., 2015. Classification of mouth movements using 7 T fMRI. J. Neural Eng. 12, 66026. https://doi.org/10.1088/1741-2560/12/6/066026.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152).

Bourguignon, M., Baart, M., Kapnoula, E. C., and Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. Journal of Neuroscience, 40(5), 1053-1065.

Brain-based decoding of human voice and speech. Science 322, 970–973. https://doi.org/10.1126/science.1164318

Branco, M.P., Freudenburg, Z.V., Aarnoutse, E.J., Bleichner, M.G., Vansteensel, M.J., Ramsey, N.F., 2016. Decoding hand gestures from primary somatosensory cortex using high-density ECoG. NeuroImage 147, 130–142. https://doi.org/10.1016/ j.neuroimage.2016.12.004

Brang, D. (2019). The Stolen Voice Illusion. Perception, 48(8), 649-667.
Brang, D., Dai, Z., Zheng, W., and Towle, V. L. (2016). Registering imaged ECoG electrodes to human cortex: A geometry-based technique. Journal of neuroscience methods, 273, 64-73.

Buzsa´ki G, Anastassiou CA, Koch C (2012) The origin of extracellular fields and currents: EEG, ECoG, LFP and spikes. Nat Rev Neurosci 13:407–420

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. PLoS computational biology, 5(7), e1000436.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. Nature neuroscience, 13(11), 1428.

Chen, T., and Rao, R. R. (1998). Audio-visual integration in multimodal communication. Proceedings of the IEEE, 86(5), 837-852.

Christensen, K., Doblhammer, G., Rau, R., & Vaupel, J. W. (2009). Ageing populations: the challenges ahead. Lancet, 374(9696), 1196–1208. https://doi.org/10.1016/S0140-6736(09)61460-4

Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. Journal of neuroscience methods, 250, 126-136.

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. Neuroimage, 9(2), 179-194.

Demandt, E., Mehring, C., Vogt, K., Schulze-Bonhage, A., Aertsen, A., & Ball, T. (2012). Reaching movement onset-and end-related characteristics of EEG spectral power modulations. Frontiers in neuroscience, 6, 65.

Elliott, T. M., and Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. PLoS comput biol, 5(3), e1000302.

Engel, A. K., and Fries, P. (2010). Beta-band oscillations—signalling the status quo?. Current opinion in neurobiology, 20(2), 156-165.

Erber, N. P. (1975). Auditory-visual perception of speech. Journal of speech and hearing disorders, 40(4), 481-492.

Eskelund, K., Tuomainen, J., and Andersen, T. S. (2011). Multistage audiovisual integration of speech: Dissociating identification and detection. Experimental Brain Research, 208(3), 447-457.

Fischl, B., Sereno, M. I., and Dale, A. M. (1999). Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. Neuroimage, 9(2), 195-207.

Fontana, L., Kennedy, B. K., Longo, V. D., Seals, D., & Melov, S. (2014). Medical research: Treat ageing. Nature, 511(7510), 405–407. https://doi.org/10.1038/511405a

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "what"?
Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Ghazanfar, A.A., C. Chandrasekaran, and N.K. Logothetis, Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. Journal of Neuroscience, 2008. 28(17): p. 4457-4469.

Ghazanfar, A.A., et al., Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. Journal of Neuroscience, 2005. 25(20): p. 5004-5012.

Glover, G. H., Li, T. Q., & Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 44(1), 162-167.

Groppe, D. M., Urbach, T. P., and Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. Psychophysiology, 48(12), 1726-1737.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. Annual review of neuroscience, 37, 435-456.

Hickok, G., Rogalsky, C., Matchin, W., Basilakos, A., Cai, J., Pillay, S., and Binder, J. (2018). Neural networks supporting audiovisual integration for speech: A large-scale lesion study. Cortex, 103, 360-371.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., and Chang, E. F. (2016). Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. Journal of Neuroscience, 36(6), 2014-2026.

Kadipasaoglu, C. M., Baboyan, V. G., Conner, C. R., Chen, G., Saad, Z. S., and Tandon, N. (2014). Surface-based mixed effects multilevel analysis of grouped human electrocorticography. Neuroimage, 101, 215-224.

Kadipasaoglu, C. M., Forseth, K., Whaley, M., Conner, C. R., Rollo, M. J., Baboyan, V. G., and Tandon, N. (2015). Development of grouped icEEG for the study of cognitive processing. Frontiers in psychology, 6, 1008.

Kaiser, J., Hertrich, I., Ackermann, H., & Lutzenberger, W. (2006). Gamma-band activity over early sensory areas predicts detection of changes in audiovisual speech stimuli. Neuroimage, 30(4), 1376-1382.

Kaiser, J., Hertrich, I., Ackermann, H., Mathiak, K., & Lutzenberger, W. (2005). Hearing lips: gamma-band activity during audiovisual speech perception. Cerebral Cortex, 15(5), 646-653.
Kalpakis, K., Gada, D., & Puttagunta, V. (2001, November). Distance measures for effective clustering of ARIMA time-series. In Proceedings 2001 IEEE international conference on data mining (pp. 273-280). IEEE.

Karas, P. J., Magnotti, J. F., Metzger, B. A., Zhu, L. L., Smith, K. B., Yoshor, D., and Beauchamp, M. S. (2019). The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. Elife, 8, e48116.

Karthik, G., Plass, J., Beltz, A. M., Liu, Z., Grabowecky, M., Suzuki, S., ... & Brang, D. (2021). Visual speech differentially modulates beta, theta, and high gamma bands in auditory cortex. European Journal of Neuroscience, 54(9), 7301-7317.

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). Finding groups in data: an introduction to cluster analysis, 344, 68-125.

Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452, 352–355. https://doi.org/10.1038/nature06713

Kayser C, Petkov CI, Logothetis NK (2008) Visual modulation of neurons in auditory cortex. Cereb Cortex 18: 1560-1574. PubMed: 18180245

Kayser, C., and Logothetis, N. K. (2009). Directed interactions between auditory and superior temporal cortices and their role in sensory integration. Frontiers in integrative neuroscience, 3, 7.

Keuken, M. C., Bazin, P. L., Crown, L., Hootsmans, J., Laufer, A., Müller-Axt, C., ... & Forstmann, B. U. (2014). Quantifying inter-individual anatomical variability in the subcortex using 7 T structural MRI. NeuroImage, 94, 40-46.

Kleiner M, Brainard D, Pelli D, 2007, "What's new in Psychtoolbox-3?" Perception 36 ECVP Abstract Supplement

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. Proceedings of the National Academy of Sciences, 103(10), 3863-3868.

Kumar, G. V., Halder, T., Jaiswal, A. K., Mukherjee, A., Roy, D., and Banerjee, A. (2016). Large scale functional brain networks underlying temporal integration of audio-visual speech perception: An EEG study. Frontiers in psychology, 7, 1558.

Lazic, S. E. (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis?. BMC neuroscience, 11(1), 5.

Lazic, S. E., Clarke-Williams, C. J., and Munafò, M. R. (2018). What exactly is 'N'in cell culture and animal experiments?. PLoS Biology, 16(4), e2005282.

Lega, B., Germi, J., and Rugg, M. D. (2017). Modulation of oscillatory power and connectivity in the human posterior cingulate cortex supports the encoding and retrieval of episodic memories. Journal of Cognitive Neuroscience, 29(8), 1415-1432.

Leonard, C. M., Puranik, C., Kuldau, J. M., & Lombardino, L. J. (1998). Normal variation in the frequency and location of human auditory cortex landmarks. Heschl's gyrus: where is it?. Cerebral Cortex (New York, NY: 1991), 8(5), 397-406.

Lewis, A. G., and Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. Cortex, 68, 155-168.
Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. Behavior research methods, 49(4), 1494-1502.

Makin, J. G., Moses, D. A., & Chang, E. F. (2020). Machine translation of cortical activity to text with an encoder–decoder framework. Nature neuroscience, 23(4), 575-582.

Markram, H., 2008. Fixing the location and dimensions of functional neocortical columns. HFSP J. 2, 132–135. https://doi.org/10.2976/1.2919545

McGrath, M., and Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. The Journal of the Acoustical Society of America, 77(2), 678-685.

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

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science, 343(6174), 1006-1010.

Micheli, C., Schepers, I. M., Ozker, M., Yoshor, D., Beauchamp, M. S., and Rieger, J. W. (2020). Electrocorticography reveals continuous auditory and visual speech tracking in temporal and occipital cortex. European Journal of Neuroscience, 51(5), 1364-1376.

Okada, K., Venezia, J. H., Matchin, W., Saberi, K., and Hickok, G. (2013). An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. PloS one, 8(6), e68959.

Olasagasti, I., Bouton, S., and Giraud, A. L. (2015). Prediction across sensory modalities: A neurocomputational model of the McGurk effect. Cortex, 68, 61-75.

Olusanya, B. O., Neumann, K. J., & Saunders, J. E. (2014). The global burden of disabling hearing impairment: a call to action. Bulletin of the World Health Organization, 92, 367-373.

Ozker, M., Schepers, I. M., Magnotti, J. F., Yoshor, D., and Beauchamp, M. S. (2017). A double dissociation between anterior and posterior superior temporal gyrus for processing audiovisual speech demonstrated by electrocorticography. Journal of cognitive neuroscience, 29(6), 1044-1060.

Ozker, M., Yoshor, D., and Beauchamp, M. S. (2018). Converging evidence from electrocorticography and BOLD fMRI for a sharp functional boundary in superior temporal gyrus related to multisensory speech processing. Frontiers in human neuroscience, 12, 141.

Peelle, J. E., and Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. Cortex, 68, 169-181.
Plass, J., Brang, D., Suzuki, S., and Grabowecky, M. (2020). Vision perceptually restores auditory spectral dynamics in speech. Proceedings of the National Academy of Sciences, 117(29), 16920-16927.

Plass, J., Guzman-Martinez, E., Ortega, L., Grabowecky, M., and Suzuki, S. (2014). Lip reading without awareness. Psychological science, 25(9), 1835-1837.

Polimeni, J.R., Fischl, B., Greve, D.N., Wald, L.L., 2010. Laminar analysis of 7 T BOLD using an imposed spatial activation pattern in human V1. NeuroImage 52, 1334–1346. https://doi.org/10.1016/j.neuroimage.2010.05.005

Ramsey et al., Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids 2017

Reale, R. A., Calvert, G. A., Thesen, T., Jenison, R. L., Kawasaki, H., Oya, H., ... and Brugge, J. F. (2007). Auditory-visual processing represented in the human superior temporal gyrus. Neuroscience, 145(1), 162-184.

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in biology and medicine, 112, 103375.

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: a robust approach. Neuroimage, 53(4), 1181-1196.

Riha, C., Güntensperger, D., Kleinjung, T., & Meyer, M. (2020). Accounting for Heterogeneity: Mixed-Effects Models in Resting-State EEG Data in a Sample of Tinnitus Sufferers. Brain topography, 33(4), 413-424.

Schroeder, C. E., and Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. Trends in neurosciences, 32(1), 9-18.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. Trends in cognitive sciences, 12(3), 106-113.

Schroeder CE, Foxe J (2005) Multisensory contributions to low-level, 'unisensory' processing. Curr Opin Neurobiol 15: 454-458. doi:10.1016/ j.conb.2005.06.008. PubMed: 16019202.

Shen, G., Zhang, J., Wang, M., Lei, D., Yang, G., Zhang, S., & Du, X. (2014). Decoding the individual finger movements from single-trial functional magnetic resonance imaging recordings of human brain activity. European Journal of Neuroscience, 39(12), 2071-2082.

Smith, E., Duede, S., Hanrahan, S., Davis, T., House, P., and Greger, B. (2013). Seeing is believing: neural representations of visual stimuli in human auditory cortex correlate with illusory auditory perceptions. PLoS One, 8(9), e73148.

Spence, C. (2007). Audiovisual multisensory integration. Acoustical science and technology, 28(2), 61-70.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams, R. M., Jr. (1949). The American Soldier: Adjustment During Army Life (Vol. 1). Princeton, NJ: Princeton University Press.

Tang, C. (2019). Cortical representation of vocal pitch during speech perception in human superior temporal gyrus. University of California, San Francisco.

Van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences, 102(4), 1181-1186.

Vatakis, A., and Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. Perception and psychophysics, 69(5), 744-756.

Verleysen, M., & François, D. (2005, June). The curse of dimensionality in data mining and time series prediction. In International work-conference on artificial neural networks (pp. 758-770). Springer, Berlin, Heidelberg.

Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., ... & Boufous, S. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. The lancet, 388(10053), 1545-1602.

Vroomen, J., and Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. Journal of Cognitive Neuroscience, 22(7), 1583-1596.

Wang, L., Wang, W., Yan, T., Song, J., Yang, W., Wang, B., ... and Wu, J. (2017). Beta-band functional connectivity influences audiovisual integration in older age: an EEG study. Frontiers in aging neuroscience, 9, 239.

Wang, X. J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. Physiological reviews, 90(3), 1195-1268.

Werner-Reiss U, Kelly KA, Trause AS, Underhill AM, Groh JM (2003) Eye position affects activity in primary auditory cortex of primates. Curr Biol 13: 554-562. doi:10.1016/S0960-9822(03)00471-8. PubMed: 12676085

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. Nature neuroscience, 7(7), 701-702.

Yao, D., Qin, Y., Hu, S., Dong, L., Bringas Vega, M. L., & Valdés Sosa, P. A. (2019). Which reference should we use for EEG and ERP practice?. Brain topography, 32(4), 530-549.

Ye, Z., Rüsseler, J., Gerth, I., & Münte, T. F. (2017). Audiovisual speech integration in the superior temporal region is dysfunctional in dyslexia. Neuroscience, 356, 1-10.

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. Neuron, 102(6), 1096-1110.

Zhu, L.L. and M.S. Beauchamp, Mouth and Voice: A Relationship between Visual and Auditory Preference in the Human Superior Temporal Sulcus. Journal of Neuroscience, 2017. 37(10): p. 2697-2708.