Visual Modulation of Speech Processing in the Auditory Cortex, Insights From iEEG, fMRI and Computational Modeling

by

Cody Zhewei Cao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Psychology)
in the University of Michigan
2025

Doctoral Committee:

Associate Professor David Brang, Chair
Associate Professor Jonathan Brennan
Associate Professor Ross Maddox
Professor Thad Polk

Cody Z. Cao

zhewei@umich.edu

ORCID ID: 0000-0003-0786-711X

© Cody Z. Cao 2025

Dedication

This dissertation is dedicated to my maternal grandparents, 徐惠民 (Huiming Xu) and 芮萍 (Rui Ping), who planted the passion for science in me since I was a child, also, who showed me first and foremost, how to be a good person.

Acknowledgements

I am deeply grateful to my doctoral advisor, Dr. David Brang. I've had the extraordinary fortune to work with David, whose patience, discipline, intelligence, and extraordinary kindness have guided me through the challenging journey of doctoral studies. Beyond teaching me how to conduct rigorous multisensory neuroscience, David has shown me how to be a professional scientist, an effective mentor, and a better person overall. David, I owe much of my intellectual and professional growth at the University of Michigan to your guidance.

I extend my sincere appreciation to my colleagues at the Multisensory Perception Lab who supported me throughout graduate school: Karthik Ganesen, EunSeon Ahn, Yike Li, Emily Cunningham, Iain Dewitt, Carli Fine, Areti Majumbar, Zoe Hajratwala, and the many undergraduate research assistants who helped with my studies. Thank you for creating a warm, collaborative environment. I hope all my future workplaces reflect the collegiality you've shown me.

I'm indebted to the faculty and staff at the University of Michigan who have invested in my development: Jon Brennan, Taraz Lee, Cindy Lustig, Andy Jahn, James Brissenden, Mara Bollard, Rick Lewis, and Abby Barrett. You challenged me to become a better scientist, academic, and educator.

To my graduate school companions – Colleen Frank, Madi Fansher, Natasha Vernooij, Maddie Quirk, Logan Walls, Quỳnh Nguyen, Tanner Nichols, Soo Ryu, Jahla Osborne, and Selena Tran – thank you for your camaraderie. This journey would have been insurmountable without your support.

I am grateful to my mentors at the University of Rochester who saw potential in me and continue to guide me with respect and encouragement: John Foxe, Ed Freedman, Edmund Lalor, Brad Mahon, Kamy Wakim, Evan Myers, Emily Knight, Tiziana Vercillio, Eric Nichols, Simone Montaya, Nicole Vieyto, Orion Haunstrap, and David Richardson. Special thanks to Aaron Nidiffer for welcoming me when I first arrived in Rochester, and to Aisling O'Sullivan for your

continued mentorship. You collectively shaped my research trajectory, and I am profoundly grateful.

I wish to thank my internship mentors – Christi Miller, Owen Brimijoin, Calvin Murdock, Greg Gage, and Christopher Harris – along with dedicated colleagues Ellie Bean Abrams, Davide Salvi, and Safa Jabri, who generously devoted their time over two summers to foster my scientific growth.

Beyond academia, I treasure the friends who have become my American family: Junjie Chen, Yanyi Chen, Juming Shen, Romy Summers, Max Pfingsten, Richard Blanco, Antu Quayum, Daniel Calleri, Jesse Rodkin, Jerry Okler, Kristen & Brandon Konschuh, Dan Hunter, Susie Stiles, John Keane, Aaron Decker, Julie Akin, Jared Haas, Peter Walcyck, Christopher Clark, David Menendez, Mitchell Fehrle, Jared Katzman, Valerie Wang, Tianyi Tang, Craig Seijas-Rodriguez, James Yoon, Michael Barbour, Chris Nameish, Andrew Harding, Safa Jabri, Jeremy Lapham, and James Archuleta. When I arrived in America nine years ago, I was simultaneously eager and apprehensive, curious yet terrified. Through all of life's uncertainties, you have been my foundation. I am honored to call you friends.

To my mother, father, and grandmother: I regret not being more present to care for you. Despite our differences, my love for you remains unwavering, and I miss you deeply. I hope my accomplishments have made you proud.

Table of Contents

Dedication ii

Acknowledgements ii

List of Tables viii

List of Figures ix

List of Appendices xi

Abstract xii

Chapter 1 Introduction 1

1.1 Components of Visual Speech 1

1.2 Contributions of Lipreading to Speech Perception 3

1.3 Spectrotemporal Information in Audiovisual Speech 3

1.4 Overview of the Dissertation 4

Chapter 2 Visual Speech Enhances Auditory Onset Timing and Envelope Tracking Through Distinct Mechanisms 6

2.1 Introduction 6

2.2 Methods 8

2.2.1 Participants, Implants and Recordings 8

2.2.2 iEEG Electrode Registration 9

2.2.3 iEEG Preprocessing 9

2.2.4 Electrode Selection 10

2.2.5 Auditory-Visual Speech Paradigm 10

2.3 Analysis 12

2.3.1 Group-Level Time-Series Analyses 12
2.3.2 Analysis of Inter-Trial Phase-Coherence (ITPC) 13
2.3.3 LME Interaction at the Group Level 14
2.3.4 Correlations at the Single Subject Level 14
2.4 Results 15
2.4.1 Spectral Responses Across Conditions 15
2.4.2 Comparing Theta ITPC and Theta Power in Auditory-Alone and Congruent-Audiovisual Trials 15
2.4.3 Interactions between Measures and Conditions 18
2.4.4 Correlations at the Individual Subject Level 18
2.4.5 Correlations at the Group Level 20
2.4.6 Comparison of Theta ITPC and Theta Power in Congruent-Audiovisual and Incongruent-Audiovisual Trials 20
2.5 Discussion 22
2.5.1 Correlated and Complementary Modes of Audiovisual Speech Processing 22
2.5.2 Preparatory Mouth Movements Modulate Theta ITPC across the STG 23
2.5.3 Sustained Theta ITPC Differences During Speech Envelope Encoding 23
2.5.4 Theta Power and Attention Modulation in Posterior STG 24
2.5.5 Limitations 25
2.6 Conclusion 26
Chapter 3 Vision Non-selectively Restores Speech Information in the Auditory Cortex 27
3.1 Introduction 27
3.2 Methods 29
3.2.1 Subjects 29
3.2.2 Stimuli: 29
3.2.3 Task 31

3.2.4 fMRI Data Collection 31
3.2.5 fMRI Data Preprocessing 32
3.2.6 Analysis 33
3.3 Results 36
3.3.1 Behavioral 36
3.3.2 Univariate Contrast 37
3.3.3 Interaction 38
3.3.4 Multivariate Pattern Analysis (MVPA) 39
3.4 Discussion 41
3.4.1 Behavioral Benefits of Visual Speech for Degraded Auditory Signals 42
3.4.2 Vision Provides General Restoration of Degraded Speech Information 42
3.4.3 Evidence for Distinct Neural Representations During Audiovisual Integration 43
3.4.4 Support for Phoneme-Level Restoration Mechanisms 44
3.4.5 Methodological Considerations and Limitations 45
3.4.6 Conclusion and Future Directions 46
Chapter 4 Viseme Categories in the Wild 47
4.1 Introduction 47
4.2 Method 49
4.2.1 Dataset 49
4.2.2 Model architecture 50
4.2.3 Distance Calculation and Statistical Testing 55
4.2.4 Viseme Clustering 56
4.2.5 Cluster Visualization 57
4.3 Results 57
4.3.1 Cosine Similarity Matrices Across Frames 58

4.3.2 Effect Size and Statistical Significance 59
4.3.3 Parametric Viseme Cluster Transitions 61
4.4 Discussion 63
4.4.1 From Coarse to Fine-Grained Visual Distinctions 64
4.4.2 Place vs. Manner Dynamics in Visemes 65
4.4.3 Implications for Models of Audiovisual Integration 65
4.4.4 From Static to Dynamic Visemes 66
4.4.5 Limitations 67
4.5 Conclusion 68
Chapter 5 General Discussion 69
5.1 Summary of Findings 69
5.2 Theoretical Integration and Implications 70
5.3 Limitations 71
5.4 Future Directions 72
5.5 Conclusion 73
Appendices 75
Bibliography 81

List of Tables

Table 2-1 Number of electrodes and participants who contributed data to each group-level time series analysis..... 10

Table 4-1 Phonemes and corresponding words used in the study. 53

Appendix Table A-1 Word Stimuli used in iEEG study..... 75

Appendix Table C-1 Two class SVM decoding accuracy between auditory degraded conditions against auditory-clear..... 77

Appendix Table C-2 Two class SVM decoding accuracy between auditory degraded conditions..... 77

Appendix Table D-1 Extensive list of all word used for training from the Lipreading in the Wild dataset. 80

List of Figures

Figure 2-1 iEEG Electrode Distribution 8

Figure 2-2 iEEG Experiment Paradigm 11

Figure 2-3 Group-level plots of ERSP 15

Figure 2-4 Group linear mixed-effect model estimates for theta ITPC between congruent audiovisual and auditory-alone 17

Figure 2-5 Single subject correlations of theta band ITPC and power differences 19

Figure 2-6 Group LME estimates for theta ITPC for congruent and incongruent audiovisual trials. 21

Figure 3-1 Spectrograms of example sentence A1 from the CID sentence set. 30

Figure 3-2 Schematic of auditory-alone and audiovisual trials. 31

Figure 3-3 Behavioral Performance for Subjects in the Scanner 36

Figure 3-4 Univariate Contrasts of Conditions Against Fixation. 38

Figure 3-5 Univariate Contrast Differences between Audiovisual and Auditory-Alone Conditions. 39

Figure 3-6 Univariate Contrasts of Conditions Against Auditory-Alone Clear. 40

Figure 3-7 Two class SVM decoding accuracies against auditory-alone clear condition for both STG and pSTS. 40

Figure 4-1 Exemplar frame from stimuli of speakers speakers from the Lipreading in the Wild dataset. 50

Figure 4-2 , Overview of the Model Structure for Video Training. 51

Figure 4-3 Visualization demo of the visual preprocessing process. 53

Figure 4-4 Distribution of auditory onset frame for stimuli onset. 55

Figure 4-5 Cosine distance between 18 selected phoneme categories across four consecutive frames. 59

Figure 4-6 Feature distinctiveness between 18 selected phoneme categories across four consecutive frames. 60

Figure 4-7 t-SNE cluster visualization of the average embeddings of the 18 phonemes. 62

Figure 4-8 Graph theory clustering of 18 selected phoneme transitions across frames, organized by place of articulation and manner of articulation. 63

Appendix Figure B-1 Individual subject electrode coverage demonstration..... 76

List of Appendices

Appendix A: Word stimuli used in the iEEG study.....75

Appendix B: iEEG Individual Subject Coverage 76

Appendix C: fMRI Decoding Results.....77

Appendix D: Viseme Study Word Stimuli List 78

Abstract

Speech perception and cognition are major cognitive functions necessary for everyday social and vocational life. Given the increasing prevalence of communicative disorders, including hearing impairments in the United States, it is important to understand the behavioral and neural mechanisms that support speech perception. This understanding could lead to improved diagnoses of communicative disorders, and more effective hearing aid development. While it is understood that people use lipreading to compensate for hearing difficulties, the specific speech information available in the visual speech signal, and the mechanisms underlying the benefit of visual speech on auditory speech perception, remain poorly understood. Literature suggests that visual lip movements provide at least three types of information to aid auditory speech perception: temporal, spectral, and categorical. Across three experiments, we examined how the auditory system uses these different types of information extracted from visual speech. In Study 1, using intracranial recordings from a large group of patients (n = 14), we demonstrated that distinct neuronal populations in the auditory cortex encode visual lip movement information both before and during the speech onset in the theta band, an oscillation whose frequency has been hypothesized to encode speech information. In ***Experiment 2***, we tested the hypothesis that visual features can selectively restore the representation of spectrally or temporally degraded auditory speech to the level of an undegraded auditory speech signal. While behavioral results indicate that listeners were able to use temporal and spectral information from visual speech, neuroimaging evidence indicates that temporal and spectral information may be restored in a non-selective manner. Finally, categorical visual phonemes (visemic information) can modulate auditory perception, yet the organization of these visual categories and their relationship to auditory phonemes remain unclear. In ***Experiment 3***, we used a deep convolutional neural network trained on a large dataset (Lipreading in the Wild) to investigate how many visemes categories exist in visual speech for a naive listener. Our results demonstrate that viseme categories do not stay static leading up to auditory onset as the field assumed previously. Instead, visemes dynamically change perceptual distance amongst themselves, and

reforms new clusters at different timepoints. Altogether, these findings advance our understanding of the mechanisms by which of how visual speech information contributes to the encoding of auditory speech during everyday communication.

Chapter 1 Introduction

Everyday communication is largely audiovisual. It has been known that being able to see the speaker's face helps us hear the speaker better across a variety of hearing environments (Sumby and Pollack, 1954; Ross et al., 2007). Among the wealth of visual information from the speakers, including hand gestures (Drijvers & Özyürek, 2017), head movements (Munhall et al., 2004), etc., researchers have focused on the mouth movements since they consistently provide the most information (Davis and Kim, 2005). While a lot of research has shown that vision modulates auditory speech perception, the exact types of information that vision provides to the auditory system remain unclear. Thus, this dissertation aims to provide clarity on the exact types of visual speech information that assist with speech perception. While previous studies have suggested that visual mouth movements mostly provide one type of information or another, we hypothesize that multiple types of auditory speech information are simultaneously available in visual speech. We hypothesize that both the speaker's face provides timing information (e.g., word onset and offset) for the listener both prior to and after auditory speech onset. Separately, we hypothesize that, separate from auditory timing information, spectral information in speech is also recovered in the auditory cortex. Since prior behavioral evidence suggests that listeners can use visible mouth width to recover auditory spectral information, we hypothesize that when such spectral information from the visual domain is being restored in the auditory cortex. Evidence in support of these hypotheses would lend credence to the idea that listeners are adaptive information-seekers that can exploit different kinds of auditory speech information from the visual domain.

1.1 Components of Visual Speech

Multiple different types of information exist in the visual domain during face-to-face communication, including head movements, hand gestures, and mouth movements that are

directly speech-related, as well as social expressions such as facial expressions that are indirectly related to speech perception.

Mouth movements have long been posited as critical to audiovisual speech perception. Particularly, listeners can use speakers' mouth movements to recover degraded speech information under various levels of noise (Sumby and Pollack 1954, Erbert, 1969). Multiple aspects of mouth movements potentially contribute to the speech perception benefit, including timing information (Ahissah et al., 2000), spectrotemporal information (Plass et al. 2020) and categorical viseme information (Gurler et al. 2015).

Head movements have been suggested to convey information related to speech rhythms, or prosody (Munhall et al, 2004). Specifically, this rhythmic information includes stress, prominence, and other aspects of prosody. Emphasis on different words in the same sentence can dramatically change the intended message of the speaker. Thus, head movement likely encodes higher-order semantic information about the ongoing speech. Consider the example, "Why do you think she took his umbrella?". Emphasis on "you", "she", "his" and "umbrella" can lead to very different meanings from the same combination of words. Thus, the listener would benefit greatly from correctly understanding the intent of the speaker. Indeed, studies have shown that subjects do take advantage of head movements to determine which word in a sentence is receiving emphatic stress (e.g., Bernstein, Eberhardt, & Demorest, 1998). Additional work has identified that the head movement induced speech benefit is present independent of being able to see mouth movements (Davis and Kim, 2005), further providing evidence that multiple aspects of the speaker's head and face provide speech-related information.

Lastly, concurrent hand gestures during speech, or iconic gestures (e.g. Goldin-Meadow & Wagner, 2005), are related to speech on a semantic level, due to the similarities between the gestures and what they represent. Consider the example of the speaker's hands moving further apart horizontally to symbolize a long object. Indeed, being able to see the speaker's gestures, in addition to the speaker's mouth movements, has been found to provide additional benefits to speech recognition and comprehension (Drijver and Özyürek, 2017).

Altogether, visual speech provides listeners with a wealth of information from the acoustic to semantic levels. In the current dissertation, we focus solely on the mouth movement aspect of visual speech.

1.2 Contributions of Lipreading to Speech Perception

The McGurk effect (McGurk and McDonald,1976) was the first to report that mouth movements can mold auditory speech outcomes. In particular, an auditory /ba/ phoneme when paired with a visual mouth movement of the phoneme /ga/, induces a percept of neither the auditory nor the visual stimulus, but a fused percept of /da/ in listeners. Critically, the McGurk effect is automatic. Even if the listeners are aware knowing that an incongruent audiovisual pair is being presented, the fusion of the audiovisual counterparts still occurs. The McGurk effect introduced the concept of visual phonemes, or visemes, and offered evidence that they have a strong modulatory effect on the perceptual fate of their auditory companion.

Visemic information encoding has been identified at multiple different sites in the brain, including the visual cortex, non-primary auditory cortex and prefrontal cortex. Curiously, during silent lipreading, even though there is no meaningful speech-related auditory input, the auditory cortex has been shown to be activated nonetheless (Calvert et. al. 1997; Pekkola et al. 2005). While this evidence suggests that information from the visual domain is being encoded in the auditory cortex, the exact type of information remains unclear. However, converging evidence from functional magnetic resonance imaging and intracranial EEG suggests that it is indeed the categorical visemic information that is encoded in the auditory cortex (Karthik et al. 2024, Bröhl et al., 2022).

Additionally, emerging evidence also suggests that primary visual cortex might already encode incoming visual input as categorical information that eventually arrives at the auditory cortex (Nidiffer et al., 2023). Indeed, studies have suggested that functional connectivity between visual and auditory cortices is modulated by hearing difficulties (Byczynski & Park, 2025) and thus abstracted visual linguistic information might directly modulate the encoding of incoming speech stimuli in the auditory cortex (Metzger et al. 2020).

1.3 Spectrotemporal Information in Audiovisual Speech

Visemic information isn't the only type of information present in mouth movements. Indeed, lip movements during speech have been suggested to carry important timing information about the ongoing speech. In particular, the mouth shape correlates strongly with the auditory speech envelope (Chandrasekaran et al., 2009, Schwartz & Savariaux, 2014). Evidence has

emerged that listeners are able to make use of this relationship between the audio and visual modalities, as seen in stronger audiovisual speech benefits for words that are mouth-led over those that are voice-led (Karas et al. 2019). Additionally, lip movements have also been suggested to carry spectral information. Specifically, mouth width has been suggested to correlate with the spectral content of the ongoing speech (Plass et al., 2020). Altogether, these findings illustrate that lip movements contain speech-related information beyond the visemic level.

1.4 Overview of the Dissertation

Across three experiments, we examined how the auditory system uses information extracted from visual speech. The current literature suggests that visual lip movements provide at least three types of information to assist with auditory speech perception. First, is temporal information about ongoing auditory speech. Specifically, visual lip movements provide both auditory onset information and ongoing timing information to the auditory cortex (Arnal et al., 2009, Karas et al., 2019, Crosse et al., 2015). However, little is known about whether these two are separate or the same processes in the auditory cortex. ***Experiment 1*** tested whether the same neuronal population in the auditory cortex encodes visual lip movement information before and during the speech onset.

Second is spectral information about the ongoing auditory speech. Evidence has shown that aspects of a speaker's face, especially the mouth width, are strongly correlated with the ongoing auditory spectral content (Chandrasekaran et al., 2009; Plass et al., 2020). Further, behavioral studies have shown that listeners are able to take advantage of the width/spectrum correlation to perceptually restore auditory speech signals (Plass et al., 2020). However, little work has examined the neural underpinnings of the auditory spectral restoration in the auditory cortex. ***Experiment 2*** tested whether visual features can selectively restore the representation of spectrally or temporally degraded auditory speech to levels of an undegraded auditory speech signal, or whether the presence of visual speech signals alters the representation of speech in a direction that is further away from the undegraded speech.

Finally, lip movements also influence higher-level phonemic representations as well. Recent work in the auditory cortex has shown that phonemic representations in the auditory

cortex are separate processes from those for phonetic features. While a growing body of literature has demonstrated the distributed nature of phoneme encoding in the auditory cortex, little is known about how visual lip movements modulate these representations. ***Experiment 3*** used deep learning to better understand the relationship between auditory and visual speech signals.

Chapter 2 Visual Speech Enhances Auditory Onset Timing and Envelope Tracking Through Distinct Mechanisms

2.1 Introduction

In challenging listening environments, the presence of a speaker's face facilitates speech recognition and comprehension (Sumby and Pollack 1959; Erber 1975; Ross et al., 2007), indicating that some information from the visual modality is relevant and beneficial to speech perception. Past research has examined what information is present in visual speech that aids speech perception processes, with studies indicating that visual timing information improves sensitivity to auditory speech onset (Abrams et al., 2008, Arnal et al., 2009, Karas et al., 2019) and entrainment to the speech envelope (for a review, see Obleser & Kayser,2019), that spectral information about the auditory signal can be recovered from mouth movements (Plass et al.,2020), and that lipreading can be used to bias phonemic processing (Karthik et al., 2024, Metzger et al., 2020). However, the independence of these processes, their neural mechanisms, and the functional roles of visual information on speech perception remain unclear.

Research has identified theta band oscillations as a candidate for how visual timing information influences auditory speech processes, with two potential mechanisms for brain-to-speech coupling. First, theta band oscillations are thought to encode the onset of auditory speech through phase-resetting mechanisms, leading to stronger oscillatory alignment in the theta band at speech onset (Giraud and Poeppel, 2012; Peelle and Davis, 2012; Ding and Simon,2014; Zoefel et al.,2018). Second, theta band power fluctuations correlate with the speech envelope, potentially reflecting enhanced time-locked processing of rhythmic speech information (e.g. Chalas et al., 2022, Ding et al. 2017; Keitel et al. 2018; Nourski et al., 2009). Thus, researchers have proposed that features of theta oscillations can encode both the onset of speech through phase-resetting mechanisms (onset encoding models) and the ongoing temporal dynamics of speech via sustained responses (envelope tracking models).

However, whether these two processes are supported by the same or distinct physiological processes remains a debate in the field. While some researchers have argued that brain-to-speech coupling in theta band only encodes onset as part of an evoked response model (Oganian et al 2023), others have presented evidence that changes in theta oscillations throughout the speech envelope explains the brain-to-speech coupling equally as well as the evoked model in which only speech onset is encoded (Duecker et al., 2024).

A parallel gap exists for understanding how visual speech timing information is used during speech perception. A large body of work has shown that audiovisual speech directly modulates activity in auditory areas, including the STG (e.g. Ghazanfar et al., 2005; Besle et al*.,* 2008; Banks et al*.* 2011). Two separate mechanisms of audiovisual enhancements have been observed: first, in an evoked fashion, preparatory mouth movements prior to auditory speech onset can modulate auditory cortical responses through phase-resetting mechanisms in the theta band (Arnal et al., 2009; Luo et al., 2010; Mégevand et al., 2020). These visual-induced phase resets, as observed by higher ITPC prior to speech onset, are thought to facilitate precise timing predictions of upcoming auditory speech (Chandrasekaran et al., 2009; Schwartz & Savariaux, 2014; Karas et al., 2019). Second, in an entrainment fashion, as a word and sentence are heard, the observation of mouth movements increases the correlation between evoked neural responses in the auditory system and the speech envelope, which correlates with both subjective and objective measures of speech intelligibility (Ahissar et al. 2001, Crosse et al., 2015; O'Sullivan et al., 2017; O'Sullivan et al., 2020).

However, it remains unclear whether the same or different physiological processes in auditory cortex make predictions for speech onset based on preparatory mouth movements and track mouth evolutions during speech. Accordingly, the present study tests the hypothesis that different crossmodal mechanisms in the STG encode information about preparatory mouth movements before speech onset and mouth evolution during the speech envelope. We recorded intracranial EEG (iEEG) data from $n = 21$ epilepsy patients with electrodes implanted within the STG and presented single syllable words both in auditory-alone and audiovisual speech contexts. Importantly, the audiovisual speech condition included both congruent and incongruent audiovisual combinations (such that on some trials the visual signal mismatched the auditory signal) but with preserved onset timing information (i.e., the visual signal correctly predicted the onset time of the auditory word even in the incongruent condition). Using group-level linear

mixed-effects modeling to compare auditory and audiovisual conditions, results showed that visual speech elicited a theta phase reset before auditory speech onset at the anterior, middle, and posterior STG. In parallel, we observed theta band power decreases in response to audiovisual speech after speech onset only in the posterior STG. Crucially, phase and power differences were dissociable at the group and individual subject levels, consistent with the hypothesis that visual speech encodes auditory onset and speech envelope through discrete mechanisms.

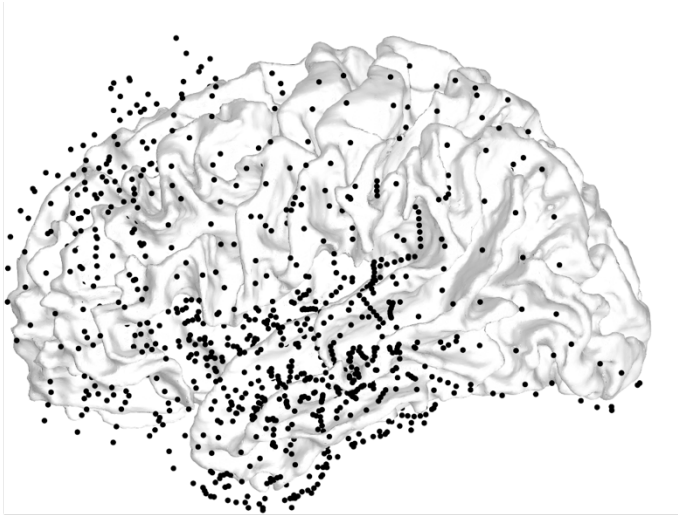


Figure 2-1 iEEG Electrode Distribution All electrodes from n=21 patients who participated in the study. A total of n=1664 electrodes were planted in these patients, shown on the Freesurfer cvs_mni_152 white matter surface.

2.2 Methods

2.2.1 Participants, Implants and Recordings

N = 21 epilepsy patients (7 female, 30 ± 10.18-year-old) undergoing clinical evaluation using iEEG participated in this study. Data from *n* = 8 of these patients were included in a prior publication, but those analyses examined different features of the task data than those reported here (Karthik et al., 2024). iEEG data were recorded from clinically implanted depth electrodes (5-mm center-to-center spacing, 2-mm diameter) and/or subdural grid electrodes (10-mm center-to-center spacing, 3-mm diameter): 6 patients had subdural electrodes, and 18 patients had depth electrodes. Across all patients, data was recorded from a total of 1664 electrodes (83.2 ± 5.9

electrodes per participant, range = 30 - 170 per participant). The number, location, and type of electrodes used were based on the clinical needs of the participants. Due to differences between amplifiers used at different recording sites, iEEG data were obtained at 1000 Hz (n = 7), 1024 Hz (n = 2), or 4096 Hz (n = 12). All participants provided informed consent under Institutional Review Board (IRB) approved protocols at the University of Chicago (UC), University of Michigan (UM), and Henry Ford hospitals (HF). Single subject electrode distribution could be found in Supplemental Information Figure 1.

2.2.2 iEEG Electrode Registration

Pre-operative T1-weighted magnetic resonance imaging (MRI) and a postoperative Computed Tomography (CT) scans were acquired for all participants in the study. Registration of the preoperative MRI to postoperative CT was performed using Statistical Parametric Mapping (Friston et al., 1994) and electrode localization was performed using custom software (Brang et al., 2016). 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 post-operative compression. The FreeSurfer image analysis suite (Dale, Fischl, and Sereno 1999; Fischl, Sereno, and Dale, 1999) was used for subsequent image processing procedures including cortical surface reconstruction, volume segmentation.

2.2.3 iEEG Preprocessing

Data were resampled to 1024 Hz during initial stages of processing for all participants. To ensure that the observed signals were analyzed from maximally local neuronal populations, data were referenced offline in a bipolar fashion (signals subtracted from each immediately adjacent electrode in a pairwise manner). Noisy electrodes were identified and rejected by estimating raw signal variance. Channels in which variability exceeded 5 SD were rejected. Data were then split into 2-second epochs (-1 to 1 seconds pre-stimulus time with t = 0 at the onset of auditory speech). Noisy trials were likewise identified and rejected when variability in the raw signal exceeded 3 SD. (across all trials within the electrode). Slow drift artifacts and power-line interference were

attenuated by high-pass filtering the data at 0.1 Hz and notch-filtering at 60 Hz (and its harmonics at 120, 180, and 240 Hz). Individual trials were then separately filtered into theta (3 - 7 Hz, wavelet cycles varied linearly from 3-5) using Morlet wavelet convolution and then power transformed.

2.2.4 Electrode Selection

Electrodes meeting both anatomical and functional criteria were included in analyses. Electrodes met the *anatomical selection criteria* if they were located within 10 mm of the FreeSurfer anatomical label "superiortemporal", derived from the MNI152 template (Collins et. al., 1994; Fonov et. al., 2011). Electrodes met the functional selection criteria if they produced a significant response in the theta band across all conditions in the [0, 500ms] time window, meaning that we observed a significant peak in the power spectral density plot (PSD) the theta band range (4-7 Hz). Among the electrodes that met the criteria for the anatomical localizer, a total of n=527 electrodes met the *functional* criteria, and were included in the study (see ***Table 2-1***).

Anterior STG		**Middle STG**		**Posterior STG**	
N. Elecs	**N Subs**	**N. Elecs**	**N Subs**	**N. Elecs**	**N Subs**
176	18	217	19	134	18

Table 2-1 Number of electrodes and participants who contributed data to each group-level time series analysis

2.2.5 Auditory-Visual Speech Paradigm

Audiovisual speech stimuli were modified from a prior study (Ross et al., 2007) in which a female speaker produced common monosyllabic words. The modified stimulus set contained 40 common monosyllabic words spoken by a female speaker. Each word began with one of four initial consonants: /b/, /f/, /g/, and /d/ (10 of each); phonemes in the second position of these words were additionally balanced across the four groups. Stimuli were recorded at 29.97 frames per second,

trimmed to 1100 ms (or 33 frames) in length, and adjusted so the first consonantal burst of sound occurred around t=500 ms during each video.

Patients were seated in a bed at the testing hospital (UM, HF & UC) where they performed a word identification task while auditory, visual, and audiovisual speech stimuli were presented via a laptop using Psychtoolbox (Brainard, 1997; Pelli, 1997). Subjects were presented with trials in one of four main conditions: auditory-alone, visual alone, congruent audiovisual or incongruent audiovisual. Each main condition had 40 trials per block for a total of 160 trials per block, and all subjects performed 2 blocks (320 trials total, 80 trials per main condition group). Half of the trials in each condition were presented as they were recorded, and in half, pink noise was added to the auditory channel to reduce the signal-to-noise (SNR) ratio of the signals to -6 db. SPL. In the current study, trials were analyzed without consideration of the clear and noisy distinction.

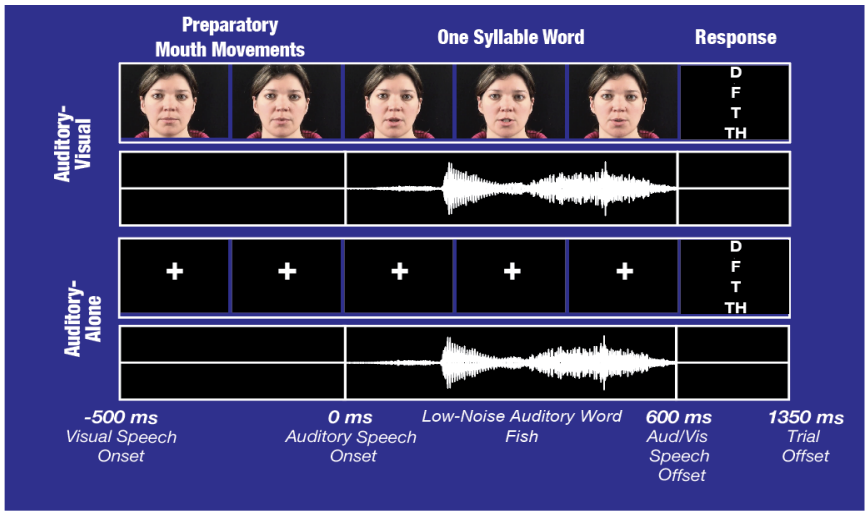


Figure 2-2 iEEG Experiment Paradigm The current study adopted a 4 x 2 design in which the task conditions orthogonally manipulated (1) the visual speech information on a trial (no visual [auditory-alone], no auditory [visual-alone], congruent AV and incongruent AV), and (2) the intelligibility of the auditory speech stimuli (low-noise or high-noise). As shown in Figure 1, subjects were presented with words and were asked to select the initial consonant heard from four options in the current trial. For example, on a trial with the auditory word 'FISH', the options presented to the participant are the letters 'D', 'F', 'T', and 'TH', which correspond to the words 'DISH', 'FISH', 'TISH', and 'THISH'.

Two task variants were used. In Task Variant 1, on each alternate trial, subjects (n = 3) were asked to speak aloud the last two words presented. In Task Variant 2 (see Fig. 2), subjects (n = 18) were asked to indicate the letter(s) associated with the initial sound of each word in a 4-alternative forced choice task. A pilot study was performed to validate stimuli use and study design where undergraduate students at the University of Michigan performed free response reports of the word they heard (Ahn et al., 2024). Based on piloting results, two sets of choices

were available for subjects depending on the word identity. For trial stimulus beginning with a consonant of “D” or ‘F’, like ‘FISH’, response options could be ‘D’, ‘F’, ‘T’, and ‘TH’, corresponding to ‘DISH’, ‘FISH’, ‘TISH’, and ‘THISH’. For trial stimulus beginning with a consonant of “B” or ‘G’, like ‘BUY’, response options would be ‘B’, ‘G’, ‘D’, and ‘TH’, corresponding to ‘BUY’, ‘GUY’, ‘DIE’, and ‘THIGH’. Participants responded using a joystick device. The full stimuli list can be found in Supplemental Information (SI) Table 1.

2.3 Analysis

2.3.1 Group-Level Time-Series Analyses

We used linear mixed-effects models to estimate general group-level differences between auditory-only and audiovisual conditions as done previously by our group (e.g., Karthik et al. 2021 and Karthik et al. 2024) and in a similar framing to other groups (Desai et. al., 2024). We created three regions of interest (ROIs) within the STG to provide coarse anatomical delineation while ensuring an adequate amount of data was present in the ROIs to fit the LME models. 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 assigned to more than one ROI.

Linear mixed-effects modeling of the time-series data was performed using the *fitlme* function in MATLAB R2021a (MathWorks Inc., Natick, MA). First, electrodes in the same ROI from the same participant were averaged at each timepoint. The justification for the averaging step is two-fold: one, neighboring electrodes share variance in the model, and two, to reduce complexity of the model. Two, individual trial theta power time-series were down-sampled to 100Hz, and trimmed to the [-1, 0.5] second window relative to auditory speech onset, leaving a total of 151 data points per trial per ROI per subject. Lastly, main-effect models were constructed at each of the 151 timepoints, in which differences between auditory-only and audiovisual trials, or congruent-audiovisual and incongruent-audiovisual conditions, were separately evaluated at each

of the three STG ROIs (anterior, middle, posterior). In Wilkinson notation (Wilkinson & Rogers, 1973), the linear mixed effect model is represented as:

Theta Power ~ Condition + (Condition|Subject).

Critically, we modeled both random intercepts and random slopes for trial conditions 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).

2.3.2 Analysis of Inter-Trial Phase-Coherence (ITPC)

For group-level analyses of transient inter-trial phase-coherence (ITPC), each epoch was convolved with 20 wavelets with center frequencies ranging from 1 to 20 Hz in 1-Hz intervals, with the number of cycles varying across this frequency range (3 cycles for the frequencies 1–3 Hz and 4–20 cycles varied linearly for the frequencies 4–20 Hz). For group-level comparisons against baseline, ITPC was calculated separately for each frequency band and each electrode during the pre-stimulus baseline period (−1,000 to −500 ms) and the 500 ms stimulus onset or offsets period (0–500 ms); we used a 500-ms window in the baseline period to ensure stable "null" ITPC values. Frequency-specific ITPC values for each period were subtracted and Fisher's Z-transformed before averaging across frequencies to produce electrode-specific difference scores. These difference scores were then averaged across each participant's auditory or visual electrodes, and these participant-level difference scores were compared to 0 using a one-tailed, one-sample *t*-test.

For group-level time-frequency analyses, ITPC values for each timepoint (−1 to 1 s) in the theta frequency band were first averaged across electrodes within each participant and then averaged across participants. Significant timepoints were identified in the -0.5 to 0.5 s range relative to the pre-stimulus baseline period (−1 to −0.6 s) using one-tailed one-sample *t* tests at each timepoint and frequency. Multiple comparison corrections were applied using cluster statistics, in which the time-frequency map of *t* value statistics is permuted 10,000 times to identify clusters of significant contiguous time-frequency points (at $p < 0.05$) that are greater than those in 95% of the permuted data.

For single-electrode analyses of transient ITPC, phase-shuffling was used to generate a null distribution of ITPC values for the 200-ms onset period in each electrode. For 10,000 iterations, a random value ranging from $-\pi$ to $+\pi$ (uniformly sampled) was added to each trial before calculation of ITPC. The distribution of ITPC values produced by this procedure is designed to reflect the expected distribution of values observed under purely spurious phase coherence. *p* values were determined by identifying the proportion of values from this distribution that equaled or exceeded observed ITPC values, and then were corrected for multiple comparisons using FDR correction ($q = 0.05$).

2.3.3 LME Correlations at the Group Level

To test whether the phase and power modulations are independent, we modeled the correlation between the conditional differences between congruent-audiovisual and auditory-alone conditions, for theta ITPC values in the -0.5 to 0 sec window (the preparatory mouth movement period), and theta power in the 0 to 0.5 sec window (the duration of auditory speech). The same time series data from congruent-audiovisual and auditory-alone conditions were used in a linear mixed-effects model to test for correlations between the two mechanisms separately at each of the three ROIs:

$$\text{Power}_{\textit{congruent-audiovisual - auditory-alone, }\theta} \sim 1 + \text{ITPC}_{\textit{congruent-audiovisual - auditory-alone, }\theta} + (1|\text{SubjectID})$$

2.3.4 Correlations at the Single Subject Level

To test for the presence of a correlation between phase and power metrics, a linear mixed-effects model, which accounts for inter-subject variability by adding subject as a random effect, was constructed to determine if there is a consistent pattern across participants. Time-series data from the theta phase and power analyses were modeled with the difference between the conditions as the fixed effect and participants as the random effect:

$$\text{Power}_{\textit{congruent-audiovisual - auditory-alone, }\theta} \sim \text{ITPC}_{\textit{congruent-audiovisual - auditory-alone ,}\theta} +$$

$$(\text{ITPC}_{\textit{congruent-audiovisual - auditory-alone, }\theta}|\text{Subject})$$

2.4 Results

2.4.1 Spectral Responses Across Conditions

We first examined theta band power across all electrodes to replicate findings of theta power increase following speech onset. Using a linear mixed effect model with subjects as the random effect with all electrodes throughout the STG (n=527) we quantified theta band power (4 - 8 Hz) after sound onset (0 ~0.5 sec) relative to the pre-stimulus baseline. Consistent with prior findings (e.g., Karthik et. al. 2021), we found a significant increase in theta band power at the group level in each of the three conditions (Figure 3): auditory-alone (F (1,26876) = 16.467, p<0.001); congruent-audiovisual (F (1,26876) = 7.07, p=0.007), incongruent-audiovisual (F (1,26876) = 6.182, p= 0.017).

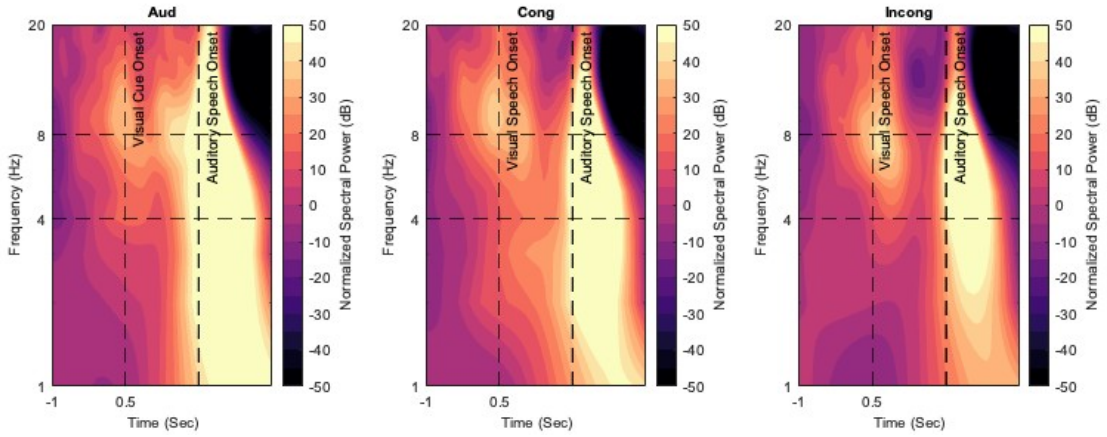


Figure 2-3 Group-level plots of ERSP Group-level plots showing event-related spectral power from 1 to 20 Hz (shown in log scale), from -1 to 0.5 secs, for all three conditions. Data reflect iEEG activity from all anatomically localized auditory electrodes (n=501), first averaged across electrodes within each participant, then averaged across participants. Vertical dotted lines denote visual mouth movement onset (-0.5 sec) and auditory speech onset (0 sec). Theta band (4-8 Hz) is highlighted by two horizontal lines. Color scale reflects normalized

2.4.2 Comparing Theta ITPC and Theta Power in Auditory-Alone and Congruent-Audiovisual Trials

First, we sought to replicate the finding that preparatory mouth movements elicit a theta phase reset (Luo et al., 2010; Mégevand et al., 2020; Biau et al., 2021). To do this, we compared

the time courses of inter-trial phase-coherence (ITPC) values for both auditory-alone and congruent-audiovisual conditions in auditory electrodes across the STG. This analysis was performed by pooling electrodes within the anterior, middle and posterior STG within each subject (Smith, 2013) and modeled both the slope and intercept using a linear mixed-effects model. Average ITPC responses from each condition and in each ROI were shown in Figure 4A. Prior to visual mouth movement onset, between -1 and -0.5 seconds, ITPC values for both congruent-audiovisual and auditory-alone remained non-significant from baseline throughout the STG. After mouth movement onset at -0.5 seconds, ITPC values in the congruent-audiovisual condition increased, peaking shortly after auditory speech onset, while ITPC values in the auditory-alone condition failed to increase until auditory speech onset. The conditional ITPC differences between congruent-audiovisual and auditory-alone were strongest during the preparatory mouth movement window in all three segments of the STG. Significantly greater ITPC ($p < .05$, FDR-corrected) was observed in the congruent-audiovisual condition compared to the auditory-alone condition across broad time ranges: anterior STG (-0.31 to 0.06 secs, and 0.42 to 0.50 sec, minimum $p = 0.0269$), middle STG (-0.45 to -0.09 sec, minimum $p < 0.001$) and posterior STG (-0.37 to -0.21 sec, minimum $p = 0.002$).

Second, we examined theta power changes across the two conditions. Prior research has demonstrated that following speech onset, congruent audiovisual trials elicit decreased theta power relative to auditory alone trials (Karthik et al., 2021). Accordingly, we compared the time courses of theta band power for both auditory-alone and congruent-audiovisual conditions in auditory electrodes across STG. As in the ITPC analyses, we pooled electrodes within each ROI per subject and modeled both the slope and intercept using a linear mixed-effects model.

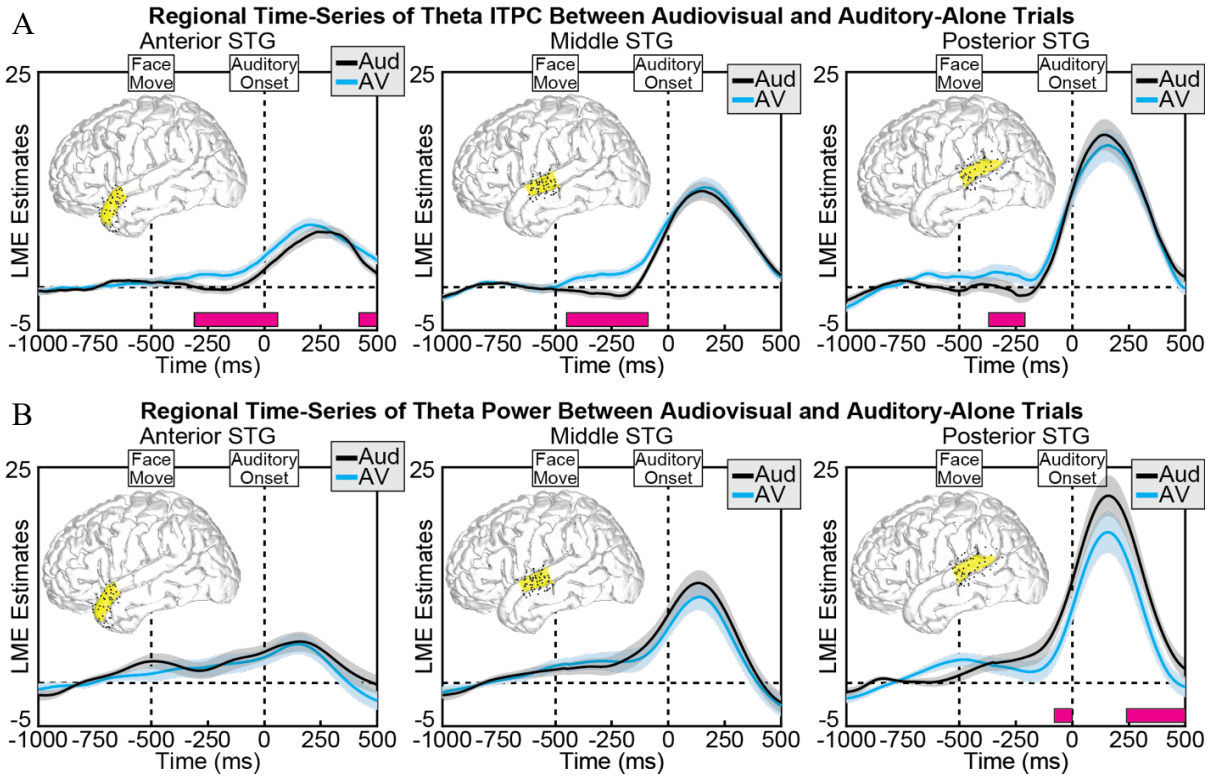


Figure 2-4 Group linear mixed-effect model estimates for theta ITPC between congruent audiovisual and auditory-alone (a) Group linear mixed-effect model (LME) estimates for each time-point of theta ITPC in auditory-alone (blue) and audiovisual (red) trials, calculated separately at anterior (left), middle (middle), and posterior (right) regions of the STG. Shaded areas reflect 95% confidence intervals. Black boxes reflect significant differences after correcting for multiple comparisons. Significant differences in theta power occurred after visual speech onset in anterior, middle and posterior STG. (b) LME estimates for each time-point of theta power in auditory-alone (blue) and audiovisual (red) trials, calculated separately at anterior (left), middle (middle), and posterior (right) regions of the STG. Shaded areas reflect 95% confidence intervals. Black boxes reflect significant differences after correcting for multiple comparisons. Significant differences in theta power occurred mainly in the posterior STG.

Results showed a similar pattern to those in Karthik et al. (2021), a study using syllable-level stimuli. Theta power within all ROIs of the STG increased steadily before auditory onset and peaked immediately after sound onset (Figure 4B). Consistent with previous results using audiovisual phonemes (Karthik et al., 2021), less theta power was observed in the congruent-audiovisual condition compared to the auditory-alone condition generally after auditory onset, with effects only present in the posterior superior temporal gyrus (-0.08 to 0 sec and 0.24 to 0.50 sec, minimum p=0.0026). No significant differences were observed after correcting for multiple comparisons at the anterior STG (minimum p = 0.557) and middle STG (minimum p = *0*.383).

2.4.3 Correlations between Measures and Conditions

The present results showed two effects of visual speech on theta band activity. First, congruent-audiovisual trials showed increased theta ITPC relative to auditory-alone trials before speech onset throughout the STG (though largest in the middle STG). Second, congruent-audiovisual trials showed decreased theta power relative to auditory-alone trials generally after speech onset and only within the posterior STG. However, these results in isolation do not directly test whether the same or different mechanisms contribute to conditional differences before and after speech onsets.

Instead, if these pre-auditory-onset and post-auditory-onset effects were dependent, we would expect to see a significant correlation between conditional differences (congruent-audiovisual vs auditory-alone) of both measures (ITPC and Power).

$$\text{Power}_{\textit{congruent-audiovisual - auditory-alone, }\theta} \sim 1 + \text{ITPC}_{\textit{congruent-audiovisual - auditory-alone, }\theta} + (1|\text{SubjectID})$$

Results showed that significant correlations between the conditional differences of the two measures in anterior STG (F (174) = 7.14, p<0.01), but not middle STG (F (215) =0.32, p=0.57), or posterior STG (F (132) =2.56, p=0.11). These results suggest that the two measures are likely correlated in anterior STG, but does not offer evidence that these two measures are correlated in the middle or posterior STG.

2.4.4 Correlations at the Individual Subject Level

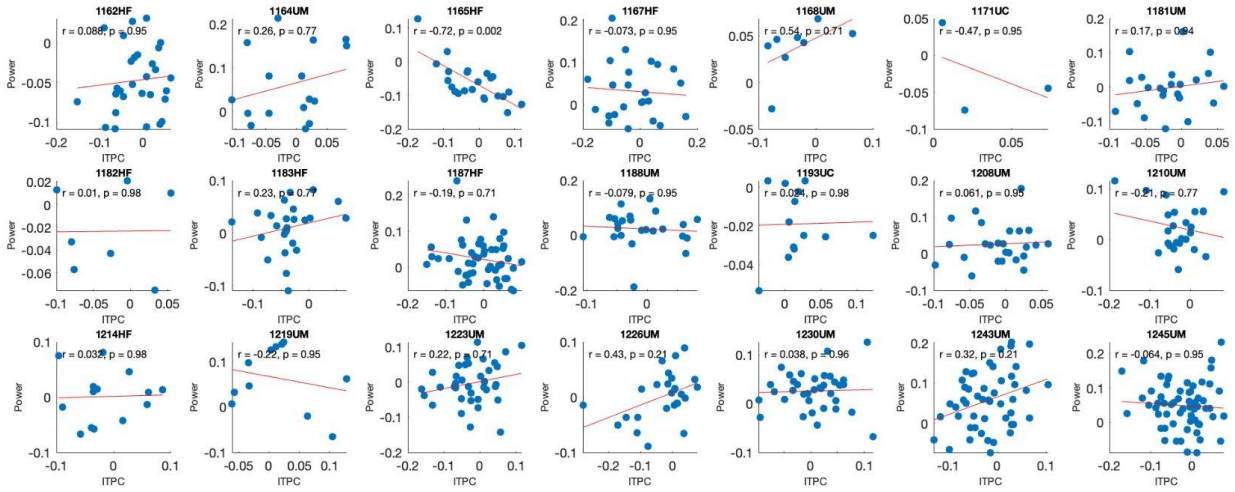


Figure 2-5 Single subject correlations of theta band ITPC and power differences Single subject correlations of theta band ITPC differences between congruent audiovisual and auditory alone conditions during preparatory mouth movement period and theta band power differences between the same two conditions following speech onset. Each blue dot represents a single electrode in the STG included in the analysis. A linear regression (red line) between the two difference values was fitted at the individual subject level to quantify the relationship between the two processes. 20 out of 21 subjects returned no significant correlation between the two metrics (p>.05 after FDR correction). One subject showed a significant relationship (r = -0.72, p = 0.002), which is marked in red.

The above test examined the relationship between the ITPC and power effects at three regions of the STG using a linear mixed effect model, showing a significant interaction between the two. However, this significant interaction does not preclude the possibility that both ITPC and power effects are still present at the same electrodes within a participant , which would run counter to our hypothesis that the effects arise from different neuronal populations. To test for a linear relationship between these effects, we calculated correlations between the ITPC and power differences at the single subject level using all electrodes within the STG) at the single subject level. Specifically, if it is true that these effects were generated by the same underlying mechanism, then we would expect a consistent pattern of negative correlations, observed at the single subject level. For example, an electrode showing a large positive effect for the ITPC analysis (i.e., $ITPC_{\theta, \text{congruent-audiovisual}} > ITPC_{\theta, \text{auditory-alone}}$ before speech onset) should also show a large negative effect for the power analysis (i.e., $Power_{\theta, \text{congruent-audiovisual}} < Power_{\theta, \text{auditory-alone}}$ after speech onset). Consistent with our prediction that the phase and power effects are generally unrelated at the electrode level, only 1 out of 21 participants showed a significant correlation between the two measures (FDR-correction applied). The scatter plots of individual-subject correlations are shown in Figure 5.

2.4.5 Correlations at the Group Level

The individual participant correlation analyses showed that only one out of the 21 participants showed a significant correlation between ITPC and power. While these single subject-level analyses can show specific instances when the ITPC and power effects co-varied due to subject specific differences or random variability, they do not capture the trends across the entire group. Using linear mixed-effect models that pool all electrodes included in the study, we modeled the correlation between pre-auditory-onset theta ITPC and post-auditory-onset theta power, with *subject* as the random variable. Across all electrodes in the STG, no relationship between power and ITPC effects was observed [F (525) = 0.06, p = 0.790]. Furthermore, restricting the model to electrodes based on anatomical region showed no significant relationships: anterior STG [F (132) = 0.963, p = 0.328], middle STG [F (215) = 1.095, p = 0.297], posterior STG [F (174) = -1.757, p = 0.071]. These results further reinforce the conclusion that pre and post speech onset effects stem from separate neuronal populations, and likely different mechanisms.

2.4.6 Comparison of Theta ITPC and Theta Power in Congruent-Audiovisual and Incongruent-Audiovisual Trials

The above analyses test the main proposed hypothesis: that congruent audiovisual ITPC and power effects are largely independent. A comparison of congruent and incongruent audiovisual speech allows a follow-up examination of how theta ITPC and power are modulated when the visual evolution of the word differs from the auditory envelope of the word. Specifically, stimuli were designed so that incongruent visual words nevertheless predicted the same onset timing of the mismatched auditory word, meaning that congruent and incongruent audiovisual stimuli should not differ before sound onset in either ITPC or power measures. Accordingly, a comparison of ITPC and power differences after speech onset enable a better interpretation of the functional significance of the two physiological responses. To this end, we compared the time courses of ITPC for congruent-audiovisual and incongruent-audiovisual speech conditions, similar to the analysis performed between auditory-alone and congruent-audiovisual. The analysis was performed by pooling electrodes within the anterior, middle and

posterior STG within each subject, and modeled both the slope and intercept using a linear mixed-effects model.

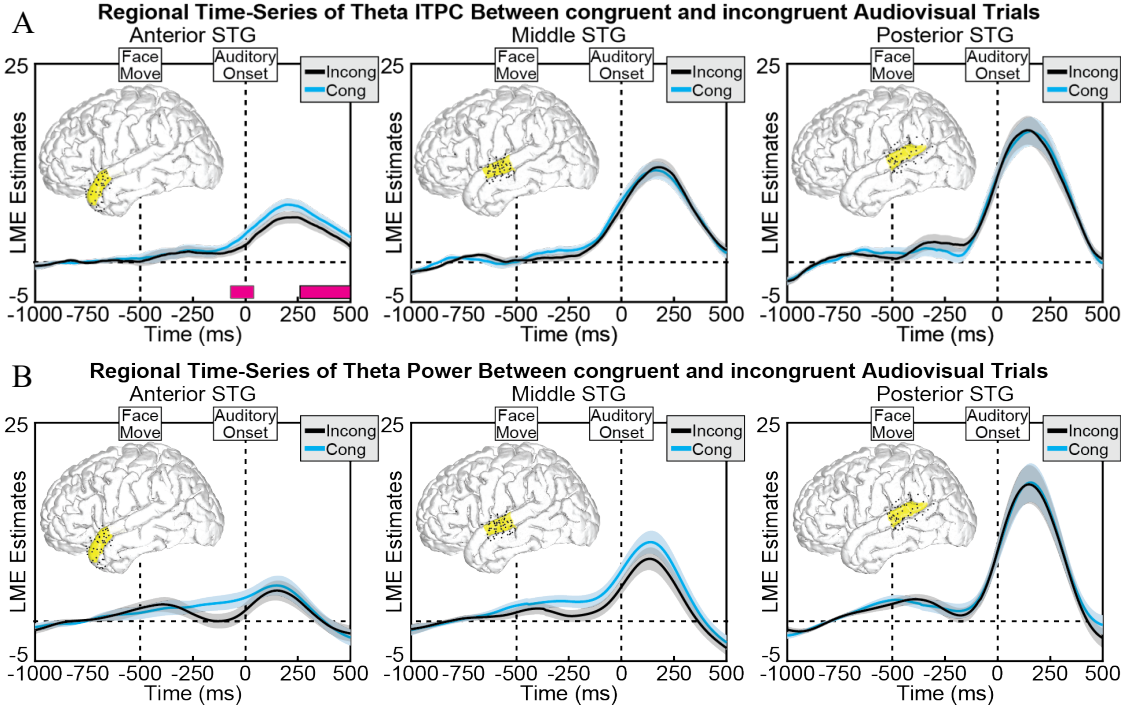


Figure 2-6 Group LME estimates for theta ITPC for congruent and incongruent audiovisual trials. (a) Group linear mixed-effect model (LME) estimates for each time-point of theta theta ITPC in congruent-audiovisual(blue) and incongruent-audiovisual (red) trials, calculated separately at anterior (left), middle (middle), and posterior (right) regions of the STG. Shaded areas reflect 95% confidence intervals. No Significant differences in theta ITPC emerged briefly after the onset of the auditory stimuli, during the time course of the stimuli in the anterior STG. (b) LME estimates for each time-point of theta power in congruent-audiovisual(blue) and incongruent-audiovisual (red) trials, calculated separately at anterior (left), middle (middle), and posterior (right) regions of the STG. Shaded areas reflect 95% confidence intervals. No Significant differences in theta power emerged during the time course of the stimuli.

As shown in Figure 6, for both congruent-audiovisual and incongruent-audiovisual conditions, theta band ITPC increased steadily after the onset of visual mouth movements and peaked quickly after sound onset. As expected, no significant differences in ITPC values were observed in response to visual speech onset. However, audiovisual congruency did introduce a significant difference in ITPC values in the anterior STG (-0.07 to 0.04 secs and 0.26 to 0.5 secs, minimum p= 0.015) generally following sound onset (the small differences present before sound onset are likely attributable to temporal smoothing due to spectral filtering). No significant ITPC differences were observed in the middle STG (minimum p = 0.999) or posterior STG (minimum p = 0.996) at any time point in the ITPC measure.

Following a similar pattern to the ITPC data, theta band power in both congruent-audiovisual and incongruent-audiovisual conditions increased steadily after mouth movement and peaked after auditory onset. Importantly, despite the differences in ITPC values between the two conditions, no significant differences were observed in theta band power in any of the three ROIs, anterior STG (minimum $p = 0.531$), middle STG (minimum $p = 0.558$) and posterior STG (minimum $p > 0.99$).

2.5 Discussion

Visual speech improves speech perception fluency, but the kinds of information that visual speech provides to the auditory system remains unclear. While prior studies have separately demonstrated that visual speech can modulate timing processes both before and after auditory speech onset, it remains unclear whether these processes are enabled by overlapping or independent neural mechanisms. In the current study, we recorded intracranial EEG data from participants as they listened to audiovisual or auditory-alone speech while performing a word detection task. Replicating prior work, we found evidence for visual modulation of speech both before and after speech onset in the STG but crucially showed a dissociation between the two. These data are consistent with a model in which speech onset information and envelope tracking from the visual stream are encoded by two distinct mechanisms in the STG.

2.5.1 Correlated and Complementary Modes of Audiovisual Speech Processing

Our findings suggest two distinct mechanisms at work; each associated with specific stages of speech processing. The first mechanism encodes predictive timing information, operating via theta phase reset during preparatory mouth movements and before auditory speech onset in the anterior and middle STG. This phase resetting enables alignment of neural activity to optimize processing in anticipation for the onset of speech. The second mechanism encodes correlated visual speech information during continuous speech in the form of increased theta band power when audiovisual speech is present.

Our findings align with the framework proposed by Campbell (2007), which suggests that visual speech operates in two modes: a "complementary mode" that provides predictive timing cues, such as mouth movements before speech onset, and a "correlated mode" where visual and auditory cues synchronize during ongoing speech. In the "complementary mode", preparatory mouth movements deliver critical timing information for auditory speech onset; however, once auditory input begins, visual cues become less crucial, as auditory information takes precedence. This interpretation aligns with our observation that visual speech modulates the predictive period before speech onset while exerting a reduced influence afterward.

2.5.2 Preparatory Mouth Movements Modulate Theta ITPC across the STG

Preparatory mouth movements led to increased theta ITPC across all STG sub-regions, without inducing changes in theta power. This finding replicates evidence that visual cues from mouth movements can reset theta phase in auditory areas (Biau et al., 2021; Mégevand et al., 2020), but with a larger sample of electrodes and patients, as well as using robust group-level statistics. This phase reset appears to contribute to predictive timing of speech onset, potentially optimizing the phase of auditory neuron activity for effective speech onset processing (Arnal et al., 2009; Karas et al., 2019). Indeed, studies suggest that the precise theta phase of auditory neurons can impact visual speech perception outcomes (Thézé et al., 2020), supporting a model in which theta phase reset in response to mouth movements facilitates optimal phase angles for processing auditory speech onset. Moreover, the observed ITPC differences in middle and posterior STG align with auditory speech studies indicating that these regions are functionally specialized for encoding speech onsets (Hamilton et al., 2018).

2.5.3 Sustained Theta ITPC Differences During Speech Envelope Encoding

We observed heightened ITPC in the congruent-audiovisual condition compared to auditory-alone in the anterior STG, particularly aligned with the auditory speech envelope. Recent evidence from auditory speech studies, such as Oganian et al. (2023), suggests that theta phase resetting occurs primarily as an evoked response to speech onset, rather than as a sustained

process throughout speech. Therefore, the ITPC differences observed during the speech envelope may be unique to audiovisual speech processing. However, comparing congruent-audiovisual and auditory-alone conditions alone limits our ability to determine whether these ITPC differences are due to broader attentional processes associated with the presence of a speaker or if they reflect more specific responses to the speech envelope itself.

However, if ITPC differences after speech onset indeed reflect attentional differences, then we would observe much weaker ITPC differences during the speech envelope between congruent-audiovisual and incongruent-audiovisual. That is not what we found. Differences between congruent-audiovisual and incongruent-audiovisual speech were significant after speech onset, suggesting that ITPC in the anterior STG may capture aspects of envelope tracking. Indeed, previous research has indicated that the anterior STG is functionally specialized for sustained speech encoding, as seen in its tracking of the speech envelope through high gamma power (HGp; a correlate of population spiking rates) (Hamilton et al., 2018).

Recent findings have suggested that HGp is modulated by theta oscillatory phases during audiovisual speech perception (Mégevand et al., 2020). Considering the sustained theta ITPC differences observed in the anterior STG, our results may offer a mechanistic basis the sustained speech encoding observed in HGp by Hamilton et al., 2018. This theta-to-HGp coupling could represent an adaptive process in the anterior STG, optimizing speech envelope tracking and potentially facilitating the integration of multisensory cues for more effective audiovisual speech processing.

2.5.4 Theta Power and Attention Modulation in Posterior STG

In addition to prominent ITPC differences, we observed theta power differences in the posterior STG (pSTG) after speech onset between the congruent-audiovisual and auditory-alone conditions, but not between the congruent-audiovisual and incongruent-audiovisual conditions. These findings suggest that the observed theta power differences in pSTG are not reflecting speech envelope related information in the visual stream but rather processes that are different only between auditory-alone and congruent-audiovisual. Thus, it is possible that theta power differences in pSTG are reflecting attentional modulation tied to the presence of multisensory input rather than processes unique to speech envelope encoding.

These findings are consistent with recent work suggesting that theta oscillations play a role in modulating attention during audiovisual speech. Specifically, Byczynski and Park (2024) demonstrated that faster theta rhythms coordinate attention by suppressing irrelevant input to optimize focus on relevant stimuli. Wang et al. (2016) further showed that theta oscillations engage cognitive control networks during cross-modal attention tasks, supporting a flexible and focused attentional state when congruent audiovisual cues are present. Together, our findings and these studies indicate that theta power in the posterior STG likely plays a broader attentional role, enhancing attention allocation in response to congruent visual and auditory signals.

While some studies have shown that congruent mouth movements aid auditory speech encoding significantly better than incongruent speech in broadband ERP signals (e.g. Crosse et al., 2015), which is largely represented by low-frequency signals including theta activity, the absence of theta power differences in our study does not contradict these findings. As broadband ERP signals reflect a composite of multiple frequency bands, it is possible that other frequencies, including delta activity which has been suggested to reflect comprehension level encoding of the speech envelope (Molinaro & Lizarazu, 2018), contributes to the envelope tracking differences seen in studies like Crosse et. al. (2015).

2.5.5 Limitations

While these findings offer valuable insights, the generalizability of our study is limited by the scope of stimuli used. Our analyses focused on whole-word stimuli grouped by four specific phonemes (/b/, /d/, /g/, /h/), raising the possibility that the observed effects may not apply to a broader range of consonants. Additionally, our use of word level stimuli, while optimal to capture pre-auditory-onset and post-auditory-onset effects for the study, does not reflect sentence level dynamics observed in more naturalistic stimuli seen in other studies (e.g. Oganian et al., 2023, Mégevand et al., 2020).

Additionally, while our finding of increased theta-band ITPC prior to speech onset is consistent with a phase-resetting interpretation, it may also reflect a stimulus-locked evoked response rather than a reset of endogenous oscillations. This distinction remains an active topic of debate in the field (Makeig et al., 2004; Sauseng et al., 2007). Notably, Oganian et al. (2024) recently demonstrated that evoked response models better explain the coupling between speech

and low-frequency neural responses, challenging the notion that sustained entrainment or phase reset is the primary mechanism during speech processing. Since we did not observe accompanying high-gamma activity or assess cross-frequency coupling in our data, we are cautious in labeling our effects as true phase reset. Instead, we interpret them more conservatively as increased phase alignment, consistent with a stimulus-locked response to visual speech cues.

2.6 Conclusion

In summary, the current study found pre-speech onset theta ITPC differences across the STG, as well as post-speech onset theta power differences, operating via two separate mechanisms. We replicate and expand on pre-speech-onset ITPC effects (Mégevand et al., 2020) as well the post speech onset theta band power differences in the posterior superior temporal gyrus (Karthik et al., 2021), providing a unified analysis of these separate processes. Together, these findings point to a dual mechanism in audiovisual speech processing: one that encodes predictive timing via phase reset and another that encodes speech correlation via sustained theta activity. Importantly, our data show that phase and power encoding are independent and are processed by largely non-overlapping neural populations within the STG.

Chapter 3 Vision Non-selectively Restores Speech Information in the Auditory Cortex

3.1 Introduction

Visual speech improves auditory speech perception accuracy in challenging listening conditions (Sumby and Pollack, 1959; Ross et al., 2007). Many studies identified different types of information that are extracted from a speaker's face that facilitate speech perception fluency. Given the correlation between lip shapes and spoken speech, lipreading has received much attention for the potential benefits it imparts on hearing (Chandrasekaran et.al., 2009). However, categorical extraction of a viseme is not the only information that visual speech provides. Aspects of a speaker's face also provide information about auditory temporal content (e.g., the onset of words, Ozker et al., 2017), as well as spectral content (relative pitch, Plass et al., 2020). However, it remains unclear how these visual signals perceptually restore information to the auditory system.

Most prior research on this topic has tested how basic units of visual speech, known as visemes, change activity and representations in the auditory system. Recent work demonstrated that silently lipread speech can be classified from auditory areas (Karthik et. al., 2024) consistent with a model in which visemic information primes phoneme sensitive populations in the superior temporal gyrus (STG) (Mesgarani et. al., 2012; Beauchamp et. al., 2010). This research is consistent with prior work indicating that visemes facilitate phoneme processing by sharing meaning information and statistical features between the visual and auditory systems (e.g., Nidiffer et al. 2023; Brohl & Kayser, 2024).

Conversely, less work has examined how spectral and temporal speech information is recovered from visual cues. Studies have shown that mouth width of the speaker is strongly correlated with the spectral content of the ongoing speech, and that listeners use that information to recover the degraded spectral content (Plass, et al. 2020). Additionally, when listeners have access to temporal information in the visual modality, speech recognition benefits significantly

(Karas *et. al.*, 2019), and the onset of visual speech modulates excitability of the auditory cortex (Mégevand *et. al.*, 2020).

Because of the wealth of relevant speech information that is available from visual speech cues, it is necessary to understand how this information is used by the auditory system to improve speech perception fluency. The first goal of the current research is to test whether the auditory system changes its use of visual speech information depending on the type of auditory information that has been degraded in the speech signal. For example, if spectral information from a speech signal is selectively degraded (making the speech sound robotic) does the auditory system selectively restore this missing spectral content from the visual speech? Alternatively, vision may have a non-selective beneficial effect on auditory speech processes that is indifferent to the fidelity of the specific auditory features, but rather a more holistic reconstruction of phoneme identity, and eventually the original speech stimuli. Indeed, there are studies that suggest that visual cues allow partial reconstruction of phonetic features crucial for identifying speech sounds, particularly in the presence of competing noise sources (O'Sullivan et al., 2021; Zhang et al, 2022, Shahin et al, 2018, Haider et al., 2024)

More generally, regardless of the way in which an auditory speech signal has been degraded, the goal is to recover missing information. A second goal of the current research is to understand how noisy audiovisual speech alters patterns of STG activity relative to clear auditory-alone speech. Specifically, does visual speech information functionally restore the pattern of activity in the STG so that it more closely matches the functional pattern that is evoked by clear auditory-only speech. Some evidence suggests that the addition of the speaker's mouth movements shifts the representations of noisy speech in the auditory regions closer to that of more intelligible, clearer speech (Rennig and Beauchamp, 2021, Zhang et. al., 2023). That said, there has been contradicting evidence from EEG literature suggesting that visual speech does not shift the auditory cortex's pattern to represent clearer speech (Shahin et al, 2013). Thus, it is possible that the spectrotemporal restoration of speech information occurs not via pattern shifts, but instead via separate neuronal populations in the STG that are involved in processing speech in auditory and audiovisual contexts. Indeed, a growing body of work has shown that posterior but not anterior parts of the auditory cortex are subject to modulation and encoding of visual speech information, especially categorical phoneme information (Metzger, et al. 2020, Karthik et al. 2022).

To test these two sets of hypotheses, we collected fMRI (functional magnetic resonance imaging) data from *n*=64 subjects while they listened to auditory and audiovisual sentences and identified keywords during each sentence. Behaviorally, keyword identification was more accurate in audiovisual sentences relative to auditory-alone sentences. Neural activity wise, audiovisual speech increased BOLD in the STG and pSTS relative to auditory-alone speech. Additionally, we found evidence to support a model in which auditory areas are insensitive to the kinds of speech information that is degraded. Critically, we did not find that visual speech shifts activation patterns in auditory regions to more closely mimic those evoked by clear speech. Instead, we provide evidence that such visual restoration occurs by the recruitment of additional auditory populations. Moreover, our findings are in line with a model in which vision has a general restorative effect on speech perception as opposed to a targeted effect that recovers only auditory features that have been degraded in a speech signal.

3.2 Methods

3.2.1 Subjects

N = 75 subjects underwent fMRI scans at the University of Michigan and data from n=64 subjects were included (n=11 were excluded either due to not completing 5 full runs of the task or due to poor task performance, i.e. having accuracy dip below chance in any one of the conditions, which is 25%; see specific details below for pre-registered exclusion criteria). Among the 64 included subjects, there were 38 females. Age range was 18-39, std dev 6.20, mean 26.43.

The sample size was calculated based on a similar phoneme-identification fMRI study conducted by our group (Karthik et al., 2024) using a similar event-related fMRI design and including a similar total scan time for each condition. The justification for the sample size can be found in our study preregistration (*DOI: 10.17605/OSF.IO/7RBGS*).

3.2.2 Stimuli

Audiovisual recordings of participant F10 speaking 100 Central Institute for the Deaf

(CID) Everyday Sentences (Lists A to J) were retrieved from the Audiovisual Database of Spoken American English (Richie *et al.,* 2009). Auditory streams for each sentence were spectrally and temporally degraded using ModFilter Sound Tools version 5 (Elliott & Theunissen, 2009); In the temporal degradation conditions (AM), temporal modulations were low pass filtered either below 2 Hz (subjects 1–17) or 4 Hz (subjects 18–64), obscuring amplitude modulations in the 2 to 7-Hz range (subjects 1–17) or 4 -8 Hz range(subjects 18–64). In the spectral degradation (FM) conditions, information along the spectral axis was low-pass filtered below 0.01 cycles/kHz, obscuring spectral details associated with formants and formant transitions (4 cycles/kHz and lower) while preserving the broadband amplitude envelope. The change in amplitude frequency filtered stimuli was made so that subjects would have similar word recognition performance in the two auditory conditions (Aud_AM and Aud_FM).

Twenty unique sentences were presented in each of five conditions (100 sentences total): unimodal unfiltered auditory, unimodal auditory with temporal degradation, audiovisual with temporal degradation, unimodal auditory with spectral degradation, and audiovisual with spectral degradation. Each condition used two of the 10 lists from the CID Everyday Sentences corpus and the lists assigned to each condition were counterbalanced across subjects. Each list was presented twice within the same condition and only after all the unique sentences were presented, for a total of forty sentences per condition (200 sentences total).

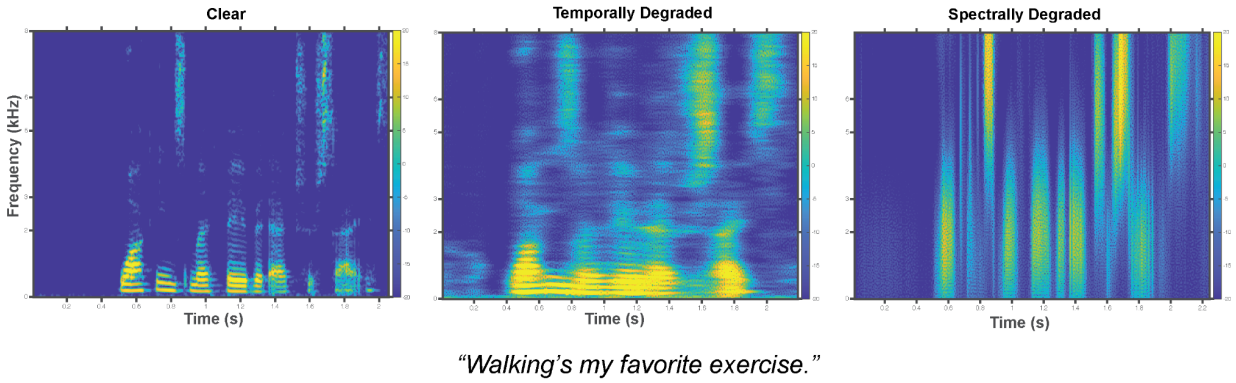


Figure 3-1 Spectrograms of example sentence A1 from the CID sentence set. 'Walking is my favorite exercise'. From left to right, the same sentence was plotted in three of the auditory conditions, clear (non-degraded), temporally degraded (AM) and spectrally degraded (FM).

3.2.3 Task

Stimuli from each condition were presented in randomized order. After each stimulus, subjects were asked to identify a target word within a given time using a joystick device in their dominant hand. Four keywords were shown on the screen: one was a word in the current sentence, and the other three were randomly selected from words present in other sentences. The order of the options was randomized on each trial. Participants were encouraged to select their best guess if they are unsure of what they heard.

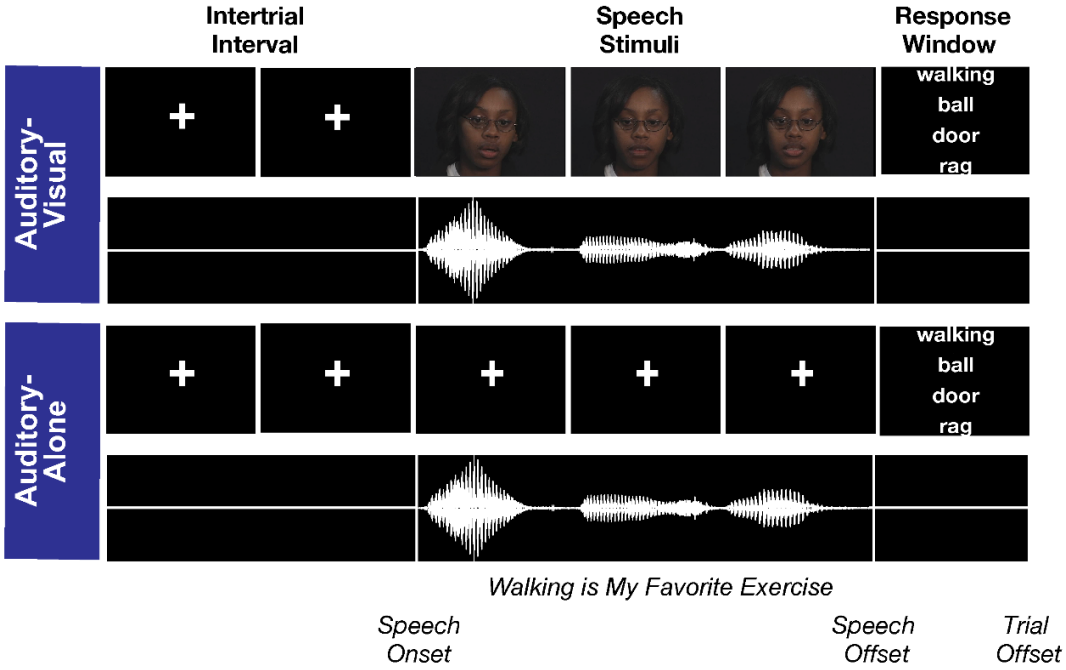


Figure 3-2 Schematic of auditory-alone and audiovisual trials. Auditory trials began with a fixation cross followed by the sentence stimulus. Visual trials presented the visual components of these same recordings with the corresponding audio track. After stimulus offset, subjects were cued to identify which one of the four keywords they heard.

3.2.4 fMRI Data Collection

Subjects were scanned in a GE scanner with a 3T magnet and Nova 32 channel Brain coil. The five functional data acquisition runs varied in length according to run number: run 1 was set to 9 minutes in length, and runs 2-5 were set to 8.5, 8, 7.5, and 7.5 minutes respectively.

The variable run times were established based on running a test version with separate pilot subjects, in which subjects took longer to perform the task in the earlier blocks than later ones.

3.2.5 fMRI Data Preprocessing

3.2.5.1 Anatomical Reconstruction

MRI and CT Acquisition and Processing. Preoperative T1-weighted magnetic resonance imaging (MRI) and postoperative computer tomography (CT) scans were acquired for all subjects. The preoperative T1 MRI was registered to the postoperative CT using SPM12 using the 'mutual information' method. The CT was not resliced or resampled. The localization of each electrode was performed using custom software. 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 post-operative compression. For all subsequent analyses including reconstruction of cortical surfaces, volume segmentation and anatomical labeling, the FreeSurfer image analysis suite was utilized (*http://surfer.nmr.mgh.harvard.edu/*).

3.2.5.2 Functional Data Preprocessing

Each subject's functional data went through slice-timing correction and spatial realignment at the scanning center. Preprocessing was completed using the preprocesses function in the FreeSurfer FS-FAST software package (version 7.4.1) (Charlestown; http://surfer.nmr.mgh.harvard.edu/). The process includes the following steps: Timing correction/reslicing, projection of the functional data onto the ***fsaverage*** surfaces in both left and right hemisphere. The normalized data was then spatially smoothed using a full-width half maximum (FWHM) kernel of 5mm or 0mm (no smoothing).

3.2.6 Analysis

3.2.6.1 *Trial-Exclusion*

An error in the generation of the stimuli led to an audiovisual asynchrony greater than 400 ms in a minority of the stimuli (17 of 100 sentences in AV_AM condition, 42 out of 100 sentences in AV_FM condition). These trials were removed from all analyses. The inclusion of all trials in behavioral and neural analyses did not change any of the reported patterns of results. Additionally, in rare instances the block recording stopped prior to the completion of a trial (1.2% of trials overall) requiring the exclusion of those trials from all fMRI analyses.

3.2.6.2 Single Subject Univariate Analysis

In the first level analysis, whole brain beta maps were generated individually for all five conditions for each of the 64 subjects using the ***mkcontrast-sess*** command within the FreeSurfer module (Fischl 2013). Within each run of each subject, a voxel-level generalized linear model was used to identify voxel-level activation during each condition. We utilize a canonical hemodynamic response function with event duration set to the length for each of the auditory or audiovisual speech stimulus and 5.5 seconds for the fixation periods (Fixation). Event onsets times are defined as the moment when the fixation cross (for auditory trials) or face (audiovisual trials) appeared on the screen. The GLM generated beta maps individually for each condition, within each block for each subject while accounting for motion correction, by including six mean-subtracted motion estimates from the co-registration routine (roll, pitch, yaw; x-, y-, z-translations) in the model.

3.2.6.3 Group Level Univariate Analysis

FreeSurfer's group-analysis pipeline was used for second level analyses (Hagler, et. al. 2006). Each subject's functional data was first concatenated using the ***isxconcat-sess*** command. Then, group-level general linear models were estimated with the command ***mri_glmfit*** separately for each hemisphere and experimental condition. Lastly, ***mri_glmfitsim*** identified significant

voxels at the group level using vertex threshold of $p < .001$ and cluster-level threshold of $p < .05$ estimated with 1000 permutations to control for multiple comparisons; p-values were adjusted for separate tests of the two hemispheres. All tests were two-tailed. The permuted statistics are plotted using the Freeview visualization tool (*tksurferfv*) on the inflated *fsaverage* surface.

3.2.6.4 Regions of Interest (ROI)

We selected two regions of interest (ROI) from both left and right hemispheres based on results from literature. Two ROIs (STG and pSTS) were pre-registered. ROIs were identified at the individual subject level based on the surface labels, '*superiortemporal'* and '*bankssts'*, from FreeSurfer *aparc-aseg* labeling (Desikan et al. 2006).

3.2.6.5 Group Level Univariate Analysis (Interactions)

Due to the similarity in the univariate contrast between AM/FM contrasts, it is important to understand whether temporal and spectral information were encoded differently in the brain. We ran a 2x2 interaction analysis between the two degraded information types (AM/FM) and two modality types (Audiovisual/Auditory-alone). Conditional differences between audiovisual and auditory-alone conditions for each degraded information type was first calculated (respectively *AV_AM_Minus_Aud_AM*, *AV_FM_Minus_Aud_FM*). A paired t-test using the *ttest* function in MATLAB was calculated for each vertex included on the fsaverage pial, and multiple comparison correction was performed using the fdr_bh function (Benjamini & Hochberg, 1995; Groppe et al., 2011).

3.2.6.6 Binary Support Vector Machine ROI Condition Decoding

To understand whether and where temporal dynamics in the stimuli were encoded in the brain in a spatially distributed manner, we performed a support vector machine (SVM) analysis on regions of interest (ROIs) on the cortical surface at the individual subject level. Two ROIs were included: the superior temporal gyrus (STG) and the posterior superior temporal sulcus

(pSTS), defined by the corresponding FreeSurfer anatomical labels, *superiortemporal* and *bankssts*.

The BOLD time series data for each subject was first detrended (polynomial order 3) and standardized by subtracting the mean across time points for each voxel. For each trial, we extracted data from baseline (frames 1-3, corresponding to 0-2.4s post-stimulus onset) and post-stimulus periods (frames 4-13, corresponding to 2.4-10.4s). For classification analyses, we used the average BOLD activity during the post-stimulus period (frames 4-13).

We conducted six pairwise condition classifications: (1) Clean Auditory vs. Auditory AM noise, (2) Clean Auditory vs. Auditory FM noise, (3) Clean Auditory vs. Audiovisual AM noise, (4) Clean Auditory vs. Audiovisual FM noise, (5) Auditory AM vs. Audiovisual AM, and (6) Auditory FM vs. Audiovisual FM. These classifications allowed us to examine neural representations across unimodal and multimodal conditions with different types of temporal disruptions.

For the SVM decoder, we implemented a binary classification approach using MATLAB's ***fitcsvm*** function with standardization. A 5-fold cross-validation procedure was employed to evaluate classification performance. Trials with timing issues or technical problems were excluded from the analysis, as were any trials with NaN values (when a scan was cut off). For each pairwise classification, we required a minimum of 5 trials per condition. The classification results were analyzed at both the individual subject level and group level, with statistical significance assessed using t-tests comparing accuracy against chance level (0.5). Multiple comparison correction, FDR, was applied to the t-test outputs similar to the interaction.

3.3 Results

3.3.1 *Behavioral*

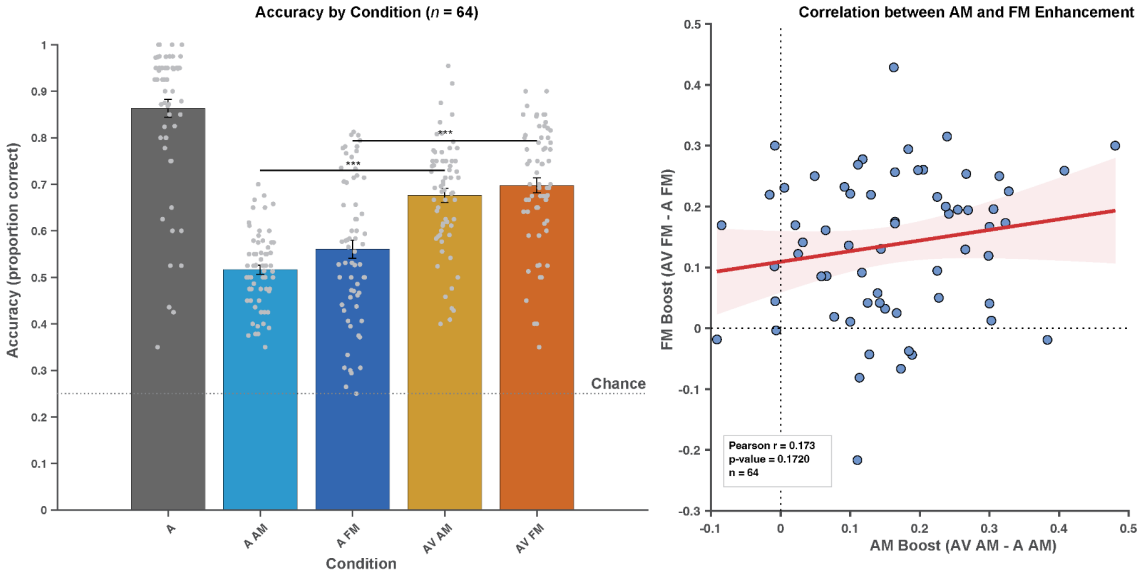


Figure 3-3 Behavioral Performance for Subjects in the Scanner *A. Barplot for the condition accuracy. Significant differences were found between AV and A conditions of both AM and FM pairs. B. Correlation between multisensory enhancement effects for amplitude modulation (AM) and frequency modulation (FM). Enhancement was calculated as the difference in accuracy between audiovisual and audio-only conditions for each modulation type. The red line represents the best-fit linear regression, with the shaded area indicating the 95% confidence interval.*

Subjects were included only if their performance in each condition was above chance (25%). For the temporally filtered (AM) conditions, subjects performed at an average accuracy of 51.64 % (SD = 8.0%, range = 35.0%~70.0%) in the audio-only AM condition (A AM) and 67.63% accuracy (SD = 12.0%, range = 40.0% ~ 95.5%) in the audiovisual AM condition (AV AM). A paired-samples t-test revealed a significant difference between these conditions ($t(63) = 10.726$, $p<0.001$, Cohen's $d = 1.341$).

For the spectrally filtered (FM) conditions, subjects performed at an average accuracy of 56.0% (SD = 15.4%, range = 25.0%~81.2%) in the audio-only AM condition (A AM) and 69.8% accuracy (SD = 12.8%, range =35.0% ~ 90.0%) in the audiovisual AM condition (AV AM). A paired-samples t-test revealed a significant difference between these conditions $t(63) = 9.105$, $p < 0.001$, Cohen's $d = 1.138$).

To examine whether the temporal enhancement and spectral enhancement were independent, we examined the relationship between amplitude modulation (AM) and frequency modulation (FM) multisensory enhancement effects by calculating a pair sample correlation between the enhancement scores (audiovisual - audio-only accuracy) for each modulation type. Pearson correlation analysis revealed no significant difference between the two types of enhancements (p=0.241); relationship between AM enhancement and FM enhancement (r = 0.173, p = 0.172, n = 64).

3.3.2 Univariate Contrast

We presented subjects (n=64) with sentences while fMRI signals were acquired. Each trial contained one sentence of varied length in one of the five conditions auditory-alone clear (Aud_Orig), auditory-alone temporally filtered (AUD_AM), auditory-alone spectrally filtered (AUD_FM), audiovisual temporally filtered (AV_AM), audiovisual spectrally filtered (AV_FM), presented in an event-related design. Data were analyzed to examine activation differences between auditory-alone and audiovisual counterparts and types of information available in the visual domain. Figure 3 shows the univariate contrasts for each of the five conditions relative to fixation. Compared to fixation, BOLD signals in all five experimental conditions showed significantly increased activity in STG (superiortemporal) and superior-frontal regions. The two audiovisual conditions, in addition to superiortemporal and superior-frontal regions, also show stronger activations in pSTS (banksSTS) and lateraloccipital and fusiform regions. Figure 3 shows the beta values of the multiple comparison corrected whole-brain analysis when comparing auditory alone and audiovisual conditions of the temporally degraded (AM) or spectrally degraded (FM) trials.

Comparing the degraded conditions between audiovisual and auditory-alone counterparts isolates regions that are modulated by the visual input. Whole-brain analyses showed that audiovisual speech elicited significantly increased BOLD activity within the STG, STS, lateral occipital cortex and fusiform area in both left and right hemispheres, all of which are statistically significant after cluster thresholding ($p < 0.05$, permutation $n = 1000$). The peak whole-brain difference was observed in lateraloccipital, superiortemporal and fusiform regions for both noise types.

3.3.3 Interaction

An interaction analysis was performed using mri_glmfit_sim in FreeSurfer. Using a cluster-forming threshold of $z = 3.0$ and $n = 1000$ permutations for Monte Carlo simulation, no clusters reached statistical significance at $p < 0.05$ (corrected for multiple comparisons, FDR). The maximum z-score observed was 4.06, which did not reach statistical significance ($p = 0.421$, corrected).

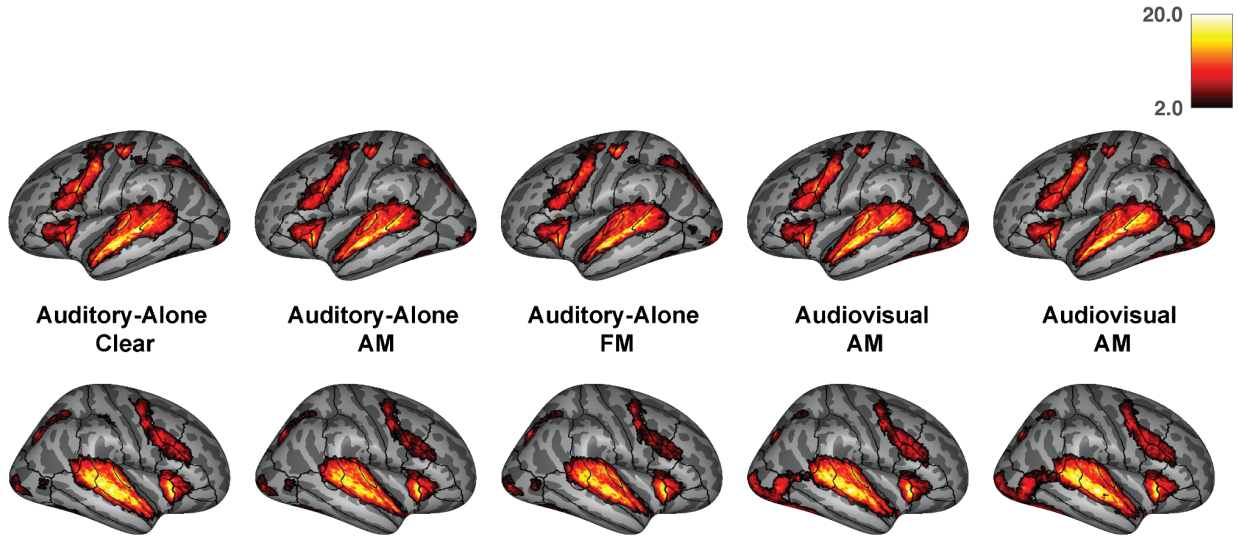


Figure 3-4 Univariate Contrasts of Conditions Against Fixation. Colored regions reflect significant increases in task-related activation for each of the five conditions against fixation period at the group level onto the fsaverage surface. aparc annotations were plotted onto the surface to mark boundary lines. The significance in BOLD difference between auditory and audiovisual conditions ($p<0.01$). Red/Yellow are for positive contrasts (e.g., Audiovisual AM > Fixation).

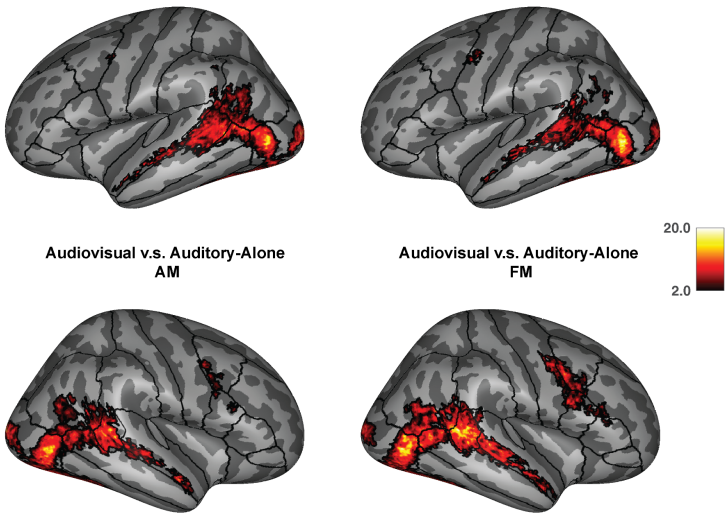


Figure 3-5 Univariate Contrast Differences between Audiovisual and Auditory-Alone Conditions. Colored regions reflect significant increases in audiovisual activation at the group level. The activation was cluster-corrected and clusters with significant p values (here shown in *-log10(p)*) were projected onto the fsaverage inflated pial surface. Notably, posterior STG and pSTS show significant activation in audiovisual conditions than auditory-alone

3.3.4 Multivariate Pattern Analysis (MVPA)

Univariate contrasts reveal activation magnitude but not informational content or representational structure. While the 2x2 interaction would inform us on whether specialized neuronal populations encode temporal and spectral information differently in a local manner, it remains possible that such information could be encoded in a spatially distributed manner. Indeed, previous decoding-based approaches using iEEG have identified auditory cortex encode sustained information, in contrast with onset/offset information, in more spatially distributed than the ones specialized in encoding speech onsets (Hamilton et al, 2018).

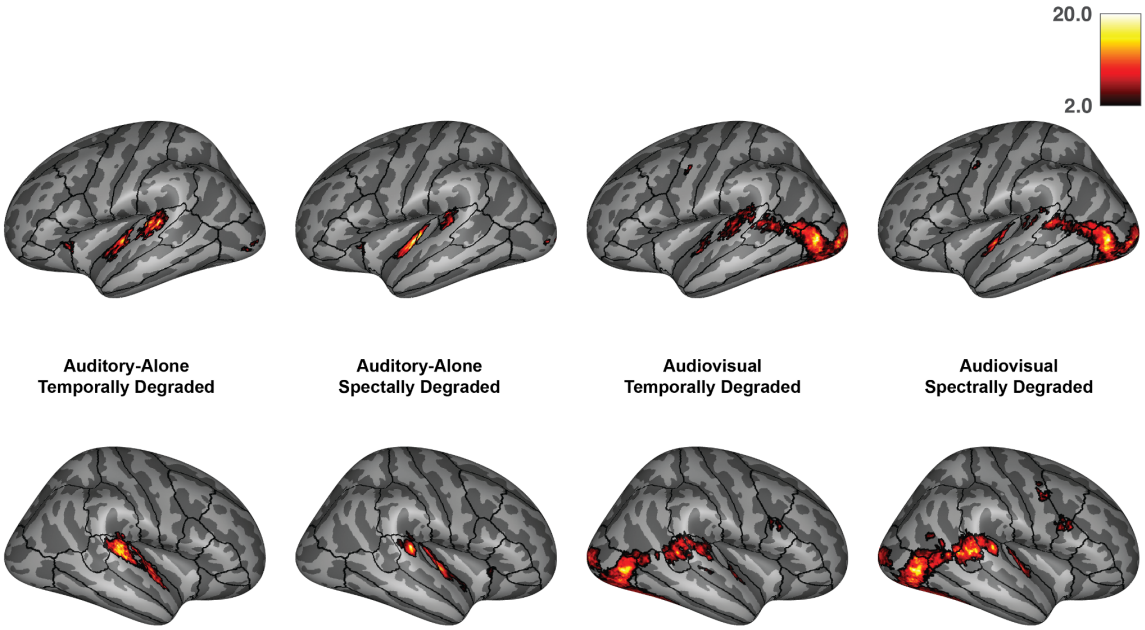


Figure 3-6 Univariate Contrasts of Conditions Against Auditory-Alone Clear. Colored regions reflect significant increases in task-related activation for each of the five conditions against the auditory-alone clear condition at the group level onto the fsaverage surface. Annotation labels "aparc" were plotted onto the surface to mark boundary lines. The significance in BOLD difference between auditory and audiovisual conditions ($p<0.01$). Red/Yellow are for positive contrasts (e.g., Audiovisual AM > Auditory-alone AM).

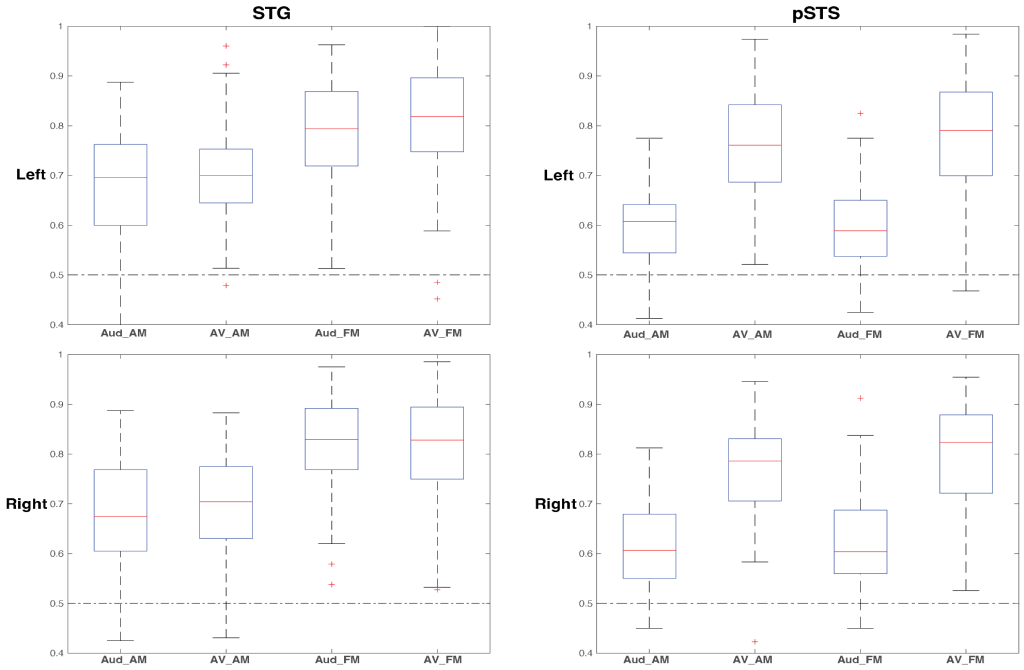


Figure 3-7 Two class SVM decoding accuracies against auditory-alone clear condition for both STG and pSTS. Decoding accuracies for all degraded conditions are significantly above chance (50%) for both STG and pSTS ($p <0.001$). Critically, STG showed significant differences between the noise types, but not within the

auditory-alone and audiovisual pairs. On the contrary, pSTS showed significant differences within the auditory-alone and audiovisual pairs, but not between the noise types.

In both left and right STG, classification accuracy for all condition pairs were significantly above chance (50%, see stats in complimentary). Critically, both AV conditions are not significantly different from their Aud condition counterparts. Nor are the AV conditions significantly different from the Aud ones. Further, no significant interaction was found between the decoding accuracies amongst the noise conditions (AM/FM) and modality types (Auditory-alone/ Audio-visual) in both hemispheres (Left, p = 0.86, t (63)= -0.18, std. dev. = 0.12; Right, p =0.47, t (63)= -0.73, std. dev. = 0.13).

In both left and right pSTS, classification accuracy for all condition pairs were significantly above chance (50%, see stats in complimentary). Both AV conditions are significantly different from their Aud condition counterparts, and the AV conditions as a whole are significantly different from the Aud ones. Interestingly, the AV conditions were more distinguishable from the Auditory-alone clear condition, compared to their auditory filtered counterparts. Lastly, no significant interaction was found between the decoding accuracies amongst the noise conditions (AM/FM) and modality types (Auditory-alone/ Audio-visual), in both hemispheres (Left, p=0.28, t (63) = -1.08, std. dev. = 0.13; Right, p =0.19, t (63) = 1.33, std. dev.= 0.13).

3.4 Discussion

A comprehensive body of research has demonstrated that visual speech information significantly modulates functional responses in the auditory system during audiovisual speech perception. In the current study, we investigated whether the auditory system adaptively utilizes visual speech information based on the specific type of auditory information that has been filtered from the speech stream, and how this recovered information is represented in auditory cortical areas. In a large sample of participants (n=64), we recorded behavioral performance and BOLD activity during a keyword detection task while either clear auditory speech or auditory/audiovisual speech was presented with either temporal or spectral information degraded in the auditory signal. Our findings support a model in which vision provides a general rather

than feature-specific restorative effect on auditory speech processing, engaging distinct neuronal populations rather than shifting activation patterns across the broader auditory cortex.

3.4.1 Behavioral Benefits of Visual Speech for Degraded Auditory Signals

Consistent with previous research (Sumby and Pollack, 1959; Ross et al., 2007), our behavioral results showed enhanced speech perception accuracy on audiovisual trials compared to auditory-only trials. Importantly, this benefit was comparable across both types of noise (spectrally or temporally degraded speech), supporting the conclusion that visual speech provides a compensatory mechanism restoring degraded auditory information. These findings align with research indicating that visual cues provide complementary information about both temporal (Ozker et al., 2017) and spectral (Plass et al., 2020), or phonemic (Shahin et al., 2018) aspects of speech.

3.4.2 Vision Provides General Restoration of Degraded Speech Information

The univariate contrasts between audiovisual and auditory-alone conditions revealed significantly stronger activations in STG and STS bilaterally for audiovisual conditions, replicating previous work demonstrating visual modulations in auditory regions (e.g., Calvert et al., 2005). Importantly, our analysis revealed equivalent enhancement of neural activity for audiovisual speech regardless of filtering type, with no significant interaction between modality (Audiovisual vs. Auditory-Alone) and information type (Temporally Filtered vs. Spectrally Filtered). This pattern of results provides compelling evidence for a model in which visual information enhances auditory speech processing in a feature-agnostic manner rather than through selective restoration targeting specific degraded features.

This finding is particularly interesting in light of the auditory speech literature, which has established functional specialization within auditory cortex, with anterior regions showing greater sensitivity to temporal information and posterior regions to spectral information (Hullett et al., 2016; Hamilton et al., 2018). Our results suggest that visual speech engages with auditory

cortex through mechanisms that transcend this specialization, providing a more generalized enhancement across the auditory processing hierarchy.

3.4.3 Evidence for Distinct Neural Representations During Audiovisual Integration

Rather than restoring auditory-like representations, our analyses revealed that visual speech creates distinct neural activation patterns during degraded speech perception. If visual input were simply restoring missing acoustic information, we expected audiovisual-degraded conditions to elicit neural patterns more similar to clear auditory speech. Instead, our multivariate analyses demonstrated that audiovisual-degraded speech was more distinguishable from clear auditory speech than auditory-alone degraded speech. This finding strongly supports a model in which visual speech engages additional or distinct neural populations rather than merely shifting existing auditory representations toward clear speech patterns.

Our data further demonstrate that STG and STS encode visual information differently during audiovisual speech processing. While STG showed sensitivity to noise type in the auditory-alone speech condition, visual input to STG did not enhance differentiation between audiovisual filtered conditions and auditory clear speech. In contrast, within STS—a region known for multisensory integration—the presence of visual information improved decoding accuracy between auditory clear and filtered conditions. This significant shift in decoding accuracy introduced by visual information in STS aligns with models proposing that pSTS receives visual inputs and serves an integrative role during audiovisual speech processing (Arnal et al, 2009; Nath et al., 2011; Ahn et al., 2024).

The increased separability of audiovisual filtered speech conditions from auditory-clear speech supports a model in which additional neuronal populations in pSTS are recruited during audiovisual processing, creating unique activation patterns specific to multimodal integration rather than simply restoring auditory representations.

3.4.4 Support for Higher-Level Speech Restoration Mechanisms

While our decoding analyses yielded null results with respect to differential restoration of spectral versus temporal information, these findings provide meaningful support for models of audiovisual speech integration that posit restoration occurs at a more abstract, categorical level rather than by reconstituting low-level acoustic features. Specifically, our findings align with theories proposing that the auditory system leverages visual input to infer or reconstruct phonemic representations, bypassing the degraded auditory signal (Shahin et al., 2018; Brohl & Kayser, 2024).

Several studies have demonstrated that phoneme identity and articulatory features can be decoded from auditory cortex during silent lipreading (Karthik et al., 2024; Nidiffer et al., 2023, Audenhaege et al., 2025), suggesting that visual input may serve to pre-activate or scaffold phoneme-level representations. This phonemic scaffolding may then support top-down predictive processes that facilitate lexical access, even when spectral or temporal fidelity of the speech information is poor. Such mechanisms are consistent with hierarchical predictive coding models, in which higher-order linguistic expectations help fill in missing sensory details (e.g., Arnal et al., 2009; Giraud & Poeppel, 2012).

Mechanistically, this higher-level restoration could occur via feedback connections from multisensory integration regions such as the posterior superior temporal sulcus (pSTS) to phoneme-selective populations in STG, enhancing their activation in the presence of congruent visual input. This process would not require the restoration of fine-grained spectrotemporal detail per se, but rather the recovery of sufficient phonological category information to support intelligibility. Notably, our data revealed increased decoding separability in pSTS but not STG, supporting the idea that distinct or additional populations are engaged under audiovisual conditions, consistent with a higher-level representational shift.

It is also possible that visual cues facilitate the restoration of even higher-level linguistic constructs, such as syntax or semantics. While our experiment did not directly assess these levels, prior studies show that intelligibility judgments and neural representations in the STG are modulated by audiovisual congruence and sentence-level coherence (Rennig & Beauchamp, 2020; Zhang et al., 2023). Our findings—that decoding accuracy was not significantly affected

by whether spectral or temporal features were degraded—imply that such restoration mechanisms may operate on a representational level above acoustics.

Critically, the null interaction in our decoding and univariate analyses should not be interpreted as the absence of restoration, but rather as evidence against feature-selective restoration. The failure to observe different neural signatures for spectral versus temporal restoration suggests that the auditory system may not attempt to reconstruct specific acoustic features from visual input. Instead, it might rely on shared visemic or phonological mappings that are functionally equivalent across different types of acoustic degradation. This non-selective enhancement would still support intelligibility by facilitating category-level identification (e.g., place and manner of articulation), enabling downstream linguistic processes to proceed despite noisy input.

3.4.5 Methodological Considerations and Limitations

While our results strongly support a general rather than feature-specific restoration model, certain aspects of our study design may have affected our ability to detect more subtle feature-selective effects. The sentence stimuli used in the current study varied in length (ranging from 1.53 - 4.64 secs, or 2-6 TRs), which, while enhancing ecological validity, complicated direct comparisons between conditions. Additionally, our design used 100 unique sentences across conditions without repetition, which prevented potential practice effects but limited our ability to perform single-trial analyses that might have revealed more subtle spectrotemporal restoration patterns.

The lack of repetition of sentences across conditions means that our ability to detect subtle representational shifts may have been limited by the inherent linguistic and acoustic variability between items. Future studies could address this limitation by using identical sentence stimuli across conditions within a participant, enabling more direct comparison of neural representations for the same linguistic content under different sensory conditions.

3.4.6 Conclusion and Future Directions

In the current study, we examined how auditory cortex processes visual speech information in the context of spectrally versus temporally degraded auditory signals. Our findings provide strong evidence for a model in which vision has a general restorative effect on auditory speech encoding rather than selectively targeting specific degraded features. Additionally, we demonstrate that audiovisual integration engages distinct neural populations, creating unique multisensory representations rather than simply restoring clear speech patterns.

These results advance our understanding of how the brain integrates visual and auditory speech information under different types of noise conditions, suggesting that the neural mechanisms supporting audiovisual speech perception operate at a more abstract representational level than direct feature-by-feature restoration. Instead, our findings point to a model where additional neuronal populations are recruited during audiovisual speech perception to form unique representations that support multimodal integration.

Future research using scalp-recorded or intracranial electroencephalography with higher temporal resolution could provide further insights into the neural dynamics of audiovisual speech integration that were not detectable with fMRI in the current study. Additionally, encoding models that explicitly represent spectral and temporal speech features could help elucidate how these features are integrated in multisensory speech processing. Such approaches may reveal more fine-grained patterns of integration in how visual speech information enhances auditory speech perception under different types of degradation.

Chapter 4 Viseme Categories in the Wild

4.1 Introduction

Seeing a speaker's face significantly enhances our ability to understand speech, particularly in challenging listening environments. This visual speech benefit has been well-documented across a wide range of contexts, from everyday conversations in noisy environments to clinical applications for individuals with hearing impairments (Sumby & Pollack, 1954; Grant et al., 2013). While multiple aspects of visual information contribute to this enhancement, one of the most extensively investigated components is the role of visemes—categories of visual mouth movements that correspond to phonetic units in spoken language.

Visual speech information influences not only low-level phonetic processing but also higher-level phonemic representations. The McGurk effect (McGurk & McDonald, 1976), in which incongruent auditory and visual syllables produce fused percepts, dramatically demonstrates how categorical visual information (visemes) can fundamentally alter phoneme perception. However, despite decades of research, the exact categorization of visemes remains a subject of ongoing debate in the field. Various classification schemes were proposed (Fisher, 1968; Walden et al., 1977; Owens & Blazek, 1985; Auer & Bernstein, 1997), but no single investigation has achieved consensus.

The persistent challenge in establishing definitive viseme categories stems largely from methodological limitations. Traditional approaches have relied on human perception studies, which are constrained by the sheer number of possible phoneme-viseme combinations—with 44 phonemes in English and at least 12 proposed viseme classes, there are minimally 528 possible combinations to investigate. The complexity increases further when considering evidence that some individuals can perceive sub-visemic differences (Files et al., 2015), potentially expanding the combination space to 1936 (44 × 44). Moreover, individual differences in speaking and perceiving styles introduce additional variability, making comprehensive human studies impractical.

Recent auditory research has demonstrated that phonetic and phonemic information are encoded differently in the brain. Phonetic features appear to be processed at the local, single-neuron level (Leonard et al., 2024), while phonemic information is represented through distributed activation patterns, with similar phonemes evoking similar neural responses (Mesgarani et al., 2012). However, far less is known about how categorical mouth movements (visemes) are encoded and how they interact with phoneme processing during audiovisual speech perception.

Our research group and others have proposed that visemic information extracted from visual speech selectively modulates phonetic features in the auditory system, biasing what listeners perceive. For example, seeing lip movements for the word "bat" while hearing the same word should enhance neural tuning to those specific phonetic features. This model further predicts that processing the word "bat" should benefit more from seeing lip movements for "pat" than for "mat," due to the greater phonetic similarity between "bat" and "pat." However, systematic tests of such predictions have been hindered by the lack of comprehensive data on the relationships between visemes and phonemes.

In this study, we address these challenges by leveraging deep learning approaches trained on a large-scale audiovisual dataset to analyze the structure of viseme representations and their relationship to phonemic representations. State-of-the-art deep learning models (Martinez et al., 2020) trained on datasets like Lipreading in the Wild (Chung et al., 2017) have demonstrated remarkable success in capturing statistical relationships between auditory and visual speech components, achieving high accuracy in lipreading novel visual stimuli. Interestingly, while human listeners struggle with lipreading tasks, machine learning algorithms have consistently achieved better results performing the same tasks (Hilder et. al., 2009). We take this as evidence that these models have learned the mapping between visual stimuli and the categorical phoneme identities. Whereas prior research has focused on the goal of word-level recognition, such practice might have introduced confounds from word lengths, making it unclear how much categorical information is present at the phoneme level. Thus, we retrained one deep learning model to allow the extraction of reliable phoneme-level probability distributions, enabling us to develop comprehensive viseme confusion matrices like those established for phonemes.

The ability of these neural networks to provide viseme estimates for each frame of a video provides a unique opportunity to examine not only what viseme categories emerge, but

also when they become distinguishable during the articulation process. This temporal dimension has been largely unexplored in previous research but is potentially crucial for understanding how visual information is integrated with auditory speech processing. Evidence suggests that visual speech benefits are not uniform across phoneme-viseme pairings (Ross et. al., 2007), and neural encoding of these pairings shows differential enhancement patterns (O'Sullivan et. al., 2020; Nidiffer & Cao et. al., 2023; Karthik et. al., 2024). By characterizing the temporal dynamics of viseme formation and organization, we aim to identify which phoneme-viseme combinations provide maximal benefit under various conditions.

Our computational approach enables a data-driven estimation of the number and boundaries of viseme categories through hierarchical clustering analysis of neural network-derived confusion matrices. This offers a crucial test of how many phonemes can be reliably identified based solely on statistical information from visual speech, without relying on human perception constraints.

4.2 Method

4.2.1 Dataset

Lipreading in the Wild (LRW), a 500-keyword audiovisual speech dataset was used in the current study. Each keyword appeared frequently in the British Broadcasting Company (BBC)'s program, and its occurrence was temporally sliced from the recording. Each instance of the keyword was 1.16 seconds long and resampled to 25 Hz (video) and 640 Hz (audio). The slicing of the keyword instance from the original video was performed in such a way so that the keyword is centered temporally, i.e. half of the keyword duration occurs before the video midpoint (t=0.58 sec), and half after. Lastly, f or training the model, each keyword had about 900-1000 instances each and divided into three splits for model training purposes -- train (n= 800 to 1000 instances), validate (n= 50) and test (n= 50).



Figure 4-1 Exemplar frame from stimuli of speakers from the Lipreading in the Wild dataset. Each stimulus is 1.16 seconds and 29 frame long for videos, and 18560 in length for the audio counterparts.

4.2.2 Model architecture

Parts of an existing model originally designed to decode word identity was used as the primary model architecture (Martinez et al., 2020). Given a video sequence, ***the first layer (a)*** is an input layer to the network is a 4D tensor, organized as B (batch) × T (frames) × H (height) ×W(width), with gray images in the videos. The input images are fed into ***the second layer (b),*** a 3D convolution with kernel 5 × 7 × 7; ***and third layer (b)***, a standard ResNet18 network, as the video feature encoder, with first convolution substituted by. No temporal down sampling is performed throughout the network, and a global spatial average pooling is applied after the last convolution, resulting in a feature map of dimensions B × C × T, where C indicates the channel dimensionality, which is 32 x 512 x 29 in the case of the current study.

4.2.2.1 Model Training

The model is trained for 80 epochs with a batch size of 32. An AdamW optimizer with an initial learning rate of 3e-4 was built into the model, whose learning rate is decayed using a cosine annealing strategy without a warm-up phase. The best model is deemed as one that yields the least loss and the highest validation accuracy. Only the best model was tested on the held-out test split of stimuli set.

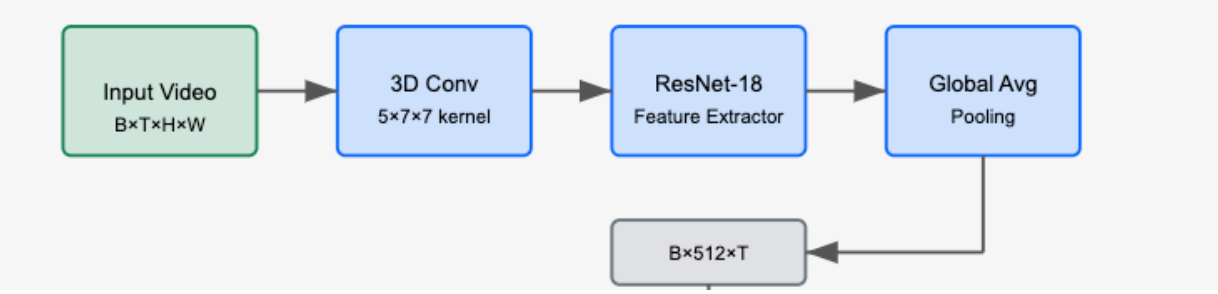


Figure 4-2 , Overview of the Model Structure for Video Training. The model training received inputs of Batch x Time (Frames) x Height x Width for each video. The visual features were extracted in two stages. One, to capture low-level motion patterns, a 3D convolution was performed on the input using a kernel with dimensions of (Frames x Height x Width). Two, to capture higher-level features, ResNet18 retains the temporal resolution and extracts the higher-level, abstract features from mouth movements. After the ResNet-18 layer, the feature maps have dimensions B×C×T×H×W (Batch size × Channels × Time frames × Height × Width). The Global Average Pooling(GAP) layer averages each of the 512 channels' values across the spatial dimensions (H×W), resulting in output with dimensions B×C×T as a feature map.

4.2.2.2 Phoneme Identification

This model was previously trained on visual mouth movements to predict word level identities using the Lipreading in the Wild dataset with over 500,000+ videos (Chung and Zisserman, 2016). Here we retrained the model on the same stimulus set using phoneme-level timing (onset, offset, duration) and labeling that are annotated using the Montreal Forced Aligner (McAuliffe et al., 2017). Specifically, to minimize effects from neighboring phonemes and visemes, the current work focuses on the first phoneme within each word.

Among the first phonemes in the 500 words in the LRW dataset, there are a total of 48 phonemes; the distribution of the number of words in each phoneme can be seen in Table 1. To ensure that the model has no interference from the frequency of occurrence from the frequency of words in the 500 words in the dataset, we selected initial phonemes that have 5 words in the dataset or more. Aligning with past viseme literature that have reported difficulties to categorize vowels(Auer & Bernstein, 2007; Binnie et al., 1974), we are retaining only the consonants in the current study. Lastly, we merged consonant variants with different aspiration properties (i.e. /k^{h}/, /p^{h}/, /t^{h}/), different palatalization (i.e. /b^{j}/, /m^{j}/,/f^{j}/, /d^{j}/), and dentalization variants (i.e., /d̪/, /t̪/, /k^{w}/). To minimize influence from mouth movements from the second phoneme, we mixed and matched the vowels in the second position.

After merging consonant variants, we identified 18 consonant phonemes which have at least 5 corresponding keywords that start with them, and supplemented with each phoneme's place and manner of articulation (Stevens, 1998; Catford, 1988)

Phoneme	Place of Articulation	Manner of Articulation	Selected Words
b	Bilabial	Stop (Plosive)	BECAUSE, BLACK, BUDGET, BIGGEST, BANKS
d	Alveolar	Stop (Plosive)	DEATH, DIFFERENCE, THERE, DAVID, DOING
dʒ	Postalveolar	Affricate	GENERAL, GEORGE, GERMANY, JAMES, JUDGE
f	Labiodental	Fricative	FOCUS, FIGHT, FRANCE, FACING, FURTHER
g	Velar	Stop (Plosive)	GREECE, GIVEN, GUILTY, GREAT, GIVING
h	Glottal	Fricative	HAPPEN, HEALTH, HOMES, HIGHER, HUNDREDS
j	Palatal	Glide (Approximant)	EUROPE, EUROPEAN, UNION, UNITED, USING
k	Velar	Stop (Plosive)	COURT, KILLED, CALLED, CAMERON, CAMPAIGN
l	Alveolar	Lateral Approximant	LABOUR, LOCAL, LEVEL, LIKELY, LARGE
m	Bilabial	Nasal	MIDDLE, MAJOR, MEANS, MOMENT, MEMBER
n	Alveolar	Nasal	NORTH, NUMBER, KNOWN, NATIONAL, NEVER
p	Bilabial	Stop (Plosive)	PLACE, POINT, PARENTS, PEOPLE, PUBLIC
s	Alveolar	Fricative	CENTRAL, SECOND, SIDES, SOCIAL, SUNDAY
t	Alveolar	Stop (Plosive)	THING, TODAY, TAKEN, TEMPERATURES, TIMES
tʃ	Postalveolar	Affricate	CHALLENGE, CHANCE, CHANGE, CHANGES, CHARGE
w	Labio-velar	Glide (Approximant)	WEAPONS, WAITING, WOMEN, WHERE, WINDS

ɹ	Postalveolar	Approximant	RATES, RATHER, REALLY, REASON, RECENT
θ	Dental	Fricative	THINGS, THINK, THOUGHT, THOUSANDS, THREAT

Table 4-1 Phonemes and corresponding words used in the study. 18 consonants with more than 5 corresponding words were selected for the current study. Places of articulation of the phoneme, manner of articulation of the phoneme and the corresponding are annotated aside.

4.2.2.3 Video Stream Preprocessing

Each video in the LRW dataset is 1.16 seconds and uniformly sampled at 25fps, making all the videos 29 frames long. Each keyword onset and duration were given by the dataset. Each video was temporally cropped so that the duration of the keyword is centered in the middle of each video.

Within each video, the stimulus was preprocessed for training efficiency. Each video is first cropped to a square containing the mouth region. Due to the variability of mouth positions across different videos, the mouth region was pre-determined by a selection of key mouth regions (points 49-68) pre-selected by the 68-point multi-PIE face-detection algorithm (Gross et al. 2010), with points 49-60 corresponding to outer lip contour and corners, and points 61-68 corresponding to inner month opening and shape. The cropped mouth region is a square of 88 x 88 pixels for each video example and was then converted to black and white for dimensionality reduction. Within an 88 x 88 matrix, each value is the grayscale values of the pixels.



Figure 4-3 Visualization demo of the visual preprocessing process. The original video undergoes four steps: 1. Face detection using the multi-pie face detection process. 2. Rotate the frame based on the detection outcome. 3. Crop out the mouth ROI using landmarks 49-68. 4. Making the frame black and white.

4.2.2.4 Temporal Realignment

Each video/audio stimulus in the dataset was centered around the auditory keyword duration, so that half of the keyword duration occurred before the midpoint of the stimulus, and half after. The setup is optimal for word level detection. However, due to the varying length of keywords, the centering of keywords is not optimal for phoneme-level feature embedding extraction. For example, when "ACCESS" and "ABSOLUTELY" both starts with the same phoneme, /ae/, due to the difference in length of the keywords, the embedding features are likely in different frames. To facilitate universal comparisons across keywords, we would need to align the keywords to start at the same time.

To implement this, using the auditory onset data given in the dataset, we aligned all instances of keywords to begin at frame #4. Specifically, if the keyword in the pre-aligned stimuli starts after frame #4, for example, if the auditory component at the video starts at frame #6, frame 1 and 2 from the video were moved to the end of the stimuli so auditory onset begins at frame 4. By doing so, we can compare feature maps across different keywords within the same frame, e.g. frame of the auditory onset, one frame prior to auditory onset, one frame after the auditory onset, etc.

To ensure that the model does not access future frames during training and inference—thereby avoiding violations of causality assumptions and artificially inflated performance—the preprocessing pipeline explicitly realigned the auditory onset to a fixed frame index. Additionally, the model input was restricted to a strictly causal five-frame window, comprising only the current frame and the four preceding frames. These changes ensured that the model's predictions were based solely on information available up to the current frame of analysis, enabling more valid, time-resolved interpretation of the learned feature embeddings.

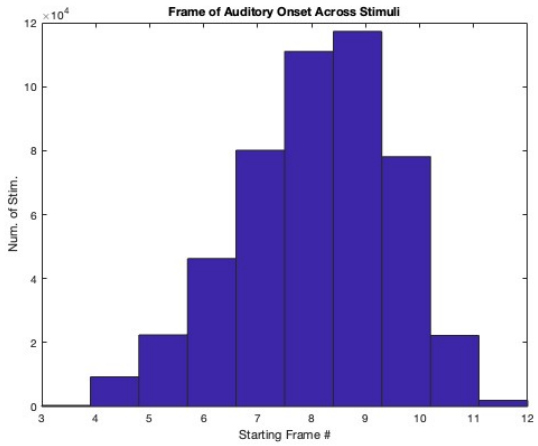


Figure 4-4 Distribution of auditory onset frame for stimuli onset. Across the 500,000 instances in the LRW audiovisual dataset., the earliest onset frame was #4 and the lastest was #12. All of the stimuli were shifted so that they would auditory onset would start on frame #4.

4.2.3 Distance Calculation and Statistical Testing

4.2.3.1 Phoneme Centroid Generation

To identify potential viseme categories, we computed pairwise distances between embeddings in the 512-dimensional feature space. For each of the 18 selected phonemes, we calculated a centroid by averaging all embeddings from test instances of words beginning with that phoneme. For example, the centroid for the phoneme /b/ represents the average of all test instances (50 examples each) of words like 'BECAUSE', 'BLACK', 'BUDGET', 'BIGGEST', and 'BANKS' - approximately 250 total examples. This process was repeated separately for each frame offset relative to speech onset (-3, -2, -1, and 0), allowing us to track the temporal evolution of phoneme representations.

4.2.3.2 Principle Component Analysis

For dimensionality reduction and to focus on the most informative features, we applied Principal Component Analysis (PCA) to the 512-dimensional embeddings before statistical

analysis. We retained the top 20 principal components in each of the selected phonemes, which captured most of the variance in the original feature space, respectively 74.63%, 70.41%, 70.28% and 71.33% across the four frames. This dimensionality reduction step was necessary as it improved computational efficiency but also mitigated potential noise in the high-dimensional embedding space, introduced by speaker differences etc. All subsequent statistical analyses and visualizations were performed on these PCA-transformed representations while maintaining the phoneme-level organization of the data.

4.2.3.3 Multivariate ANOVA

To determine whether different phoneme categories are statistically distinguishable in the visual domain, we employed Multivariate Analysis of Variance (MANOVA) on the cosine distance of the PCA-transformed embeddings. Cosine distance was selected over Euclidean metrics because it focuses solely on the directional patterns of feature activation rather than magnitude. This approach is critical when working in high-dimensional spaces where Euclidean distances can be disproportionately influenced by vector magnitude differences that may not be linguistically relevant.

4.2.4 Viseme Clustering

To identify viseme categories in a statistically rigorous manner, we employed a parametric approach based on multivariate analysis of variance (MANOVA) and graph theory. This method identifies clusters based on statistical discriminability between phonemes in the visual embedding space. For each frame relative to speech onset, we performed the following analysis.

- Previously, we applied MANOVA to test whether each pair of phonemes was significantly different in the 20-dimensional PCA-reduced visual embedding space. This resulted in a pairwise p-value matrix where each cell represented the statistical significance of differences between two phoneme categories.

- We applied Bonferroni correction to control for family-wise error rate across all pairwise comparisons (153 tests per frame).
- We converted the p-value matrix into an adjacency matrix by setting connections between phonemes that were not significantly different ($p > 0.01$) to 1, and significantly different pairs to 0. The diagonal was set to 1 since each phoneme is identical to itself.
- We constructed an undirected graph from this adjacency matrix using MATLAB's *graph* function, where connected nodes (phonemes) represent visually similar articulations.
- We identified connected components in this graph using the *conncomp* function, which assigns each phoneme to a cluster where members are not significantly different from each other.

4.2.5 Cluster Visualization

To visualize relationships between different phonemes in the visual speech space, we created matrices that represent both the effect size (Wilks' Lambda, Johnson & Wichern, 2007) and statistical significance (p-values) between each phoneme pair. We employed t-SNE (van der Maaten & Hinton, 2008) to project the high-dimensional embeddings into a 2D space for visualization, both at the individual example level and at the phoneme cluster centroid level.

Additionally, we developed alluvial diagrams to track how phoneme cluster memberships change across frames relative to speech onset. This visualization reveals when viseme distinctions become most prominent during articulation and how they evolve as the mouth transitions through articulatory movements.

The temporal progression of these visualizations across different frames provides insight into the dynamic nature of viseme formation, showing that certain phoneme distinctions are more salient at specific points in the articulation sequence.

4.3 Results

To understand how distinguishable visual phonemes, or visemes are from each other at each time point, we extracted the feature space for each test instance (n=50) of the keywords

(n=500). We then calculated cosine distances between all pairs of keywords within the same frame. Next, we aggregated all keywords within the same initial phoneme class, and performed pairwise statistical comparisons between phoneme classes using multivariate ANOVA with Bonferroni correction for multiple comparisons.

4.3.1 Cosine Similarity Matrices Across Frames

The cosine similarity matrices revealed an evolution in the representational structure of visual phonemes across temporal frames. At the earliest pre-articulatory stage (Frame -3, approximately 120ms before articulation onset), the similarity matrix showed a relatively undifferentiated pattern with weak phoneme-specific clustering (Figure 4). The dominant diagonal structure indicates that while each phoneme is self-similar, the distinctions between different phonemes are less pronounced, with similarity values largely falling within a narrow range (0.2-0.6).

As articulation approaches (Frame -2), more structured patterns begin to emerge. Notable is the increasing separation between certain phonetic categories, particularly between labial consonants ('b', 'p', 'm') and coronal consonants ('t', 's'). However, many phonemes still show moderate cross-similarity, suggesting that visual distinctiveness remains partially ambiguous at this stage.

By Frame -1, approximately 40 ms before articulation onset, the similarity structure becomes considerably more organized. Clear blocks of higher similarity emerge, particularly among phonemes that share similar place of articulation features. The most prominent pattern is the high similarity among velar and glottal consonants ('g', 'h', 'j', 'k', 'l'), which form a distinct yellow-orange block indicating strong representational similarity. Simultaneously, bilabial consonants ('b', 'm', 'p') show increasing within-category similarity while maintaining clear boundaries from other phonetic groups.

At articulation onset (Frame 0), the similarity matrix exhibits its most differentiated structure. The diagonal dominance is strongest, indicating maximally distinctive phoneme-specific representations. The block structure is further refined, with clear organization around both place and manner of articulation. The affricates ('dzh', 'tsh') show increased similarity to

each other while maintaining clear boundaries from other categories. Similarly, the alveolar consonants form a more coherent group with higher within-category similarity.

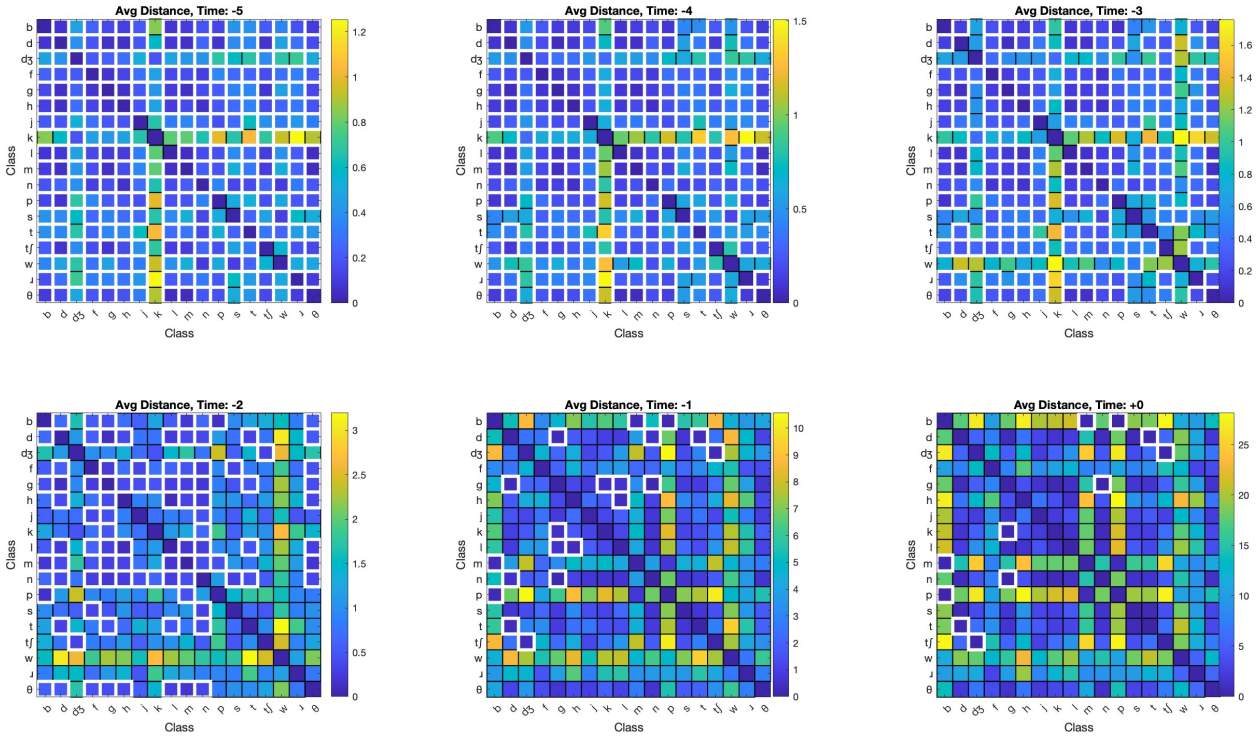


Figure 4-5 Cosine distance between 18 selected phoneme categories across six consecutive frames. Cosine similarity between phoneme class embeddings across time frames. Each cell represents the cosine similarity between the mean embeddings of two phoneme classes. The order of phonemes has been re-ordered based on clustering outcome. Phoneme pairs that are significantly different from null (10^4 permutations) are marked in white brackets. Higher values indicate greater similarity in embedding direction, suggesting reduced visual discriminability. Temporal progression shows increased structure and clustering around phoneme onset.

4.3.2 Effect Size and Statistical Significance

The effect size matrices (Wilks Lambda, shown Figure 6) revealed the magnitude of phoneme distinctiveness across frames. At Frame -3, effect sizes are relatively modest (predominantly blue, indicating values above 0.6), suggesting limited discriminability between most phoneme pairs. By Frame -2, the effect sizes strengthen slightly but remain moderate.

A substantial increase in effect size is observed at Frame -1, with many phoneme pairs showing strong effects (yellow regions, values below 0.4), particularly between phonemes from

different articulatory categories. At Frame +0, the effect sizes reach their strongest values, with numerous phoneme pairs showing highly significant differences (extensive yellow regions).

The pairwise p-value matrices confirmed this pattern, with increasing statistical significance of phoneme distinctions as articulation approaches. At Frame -3, relatively few phoneme pairs show significant differences ($p < 0.05$, blue squares). By Frame +0, most cross-category comparisons reach statistical significance, indicating robust differentiation between phonetic categories.

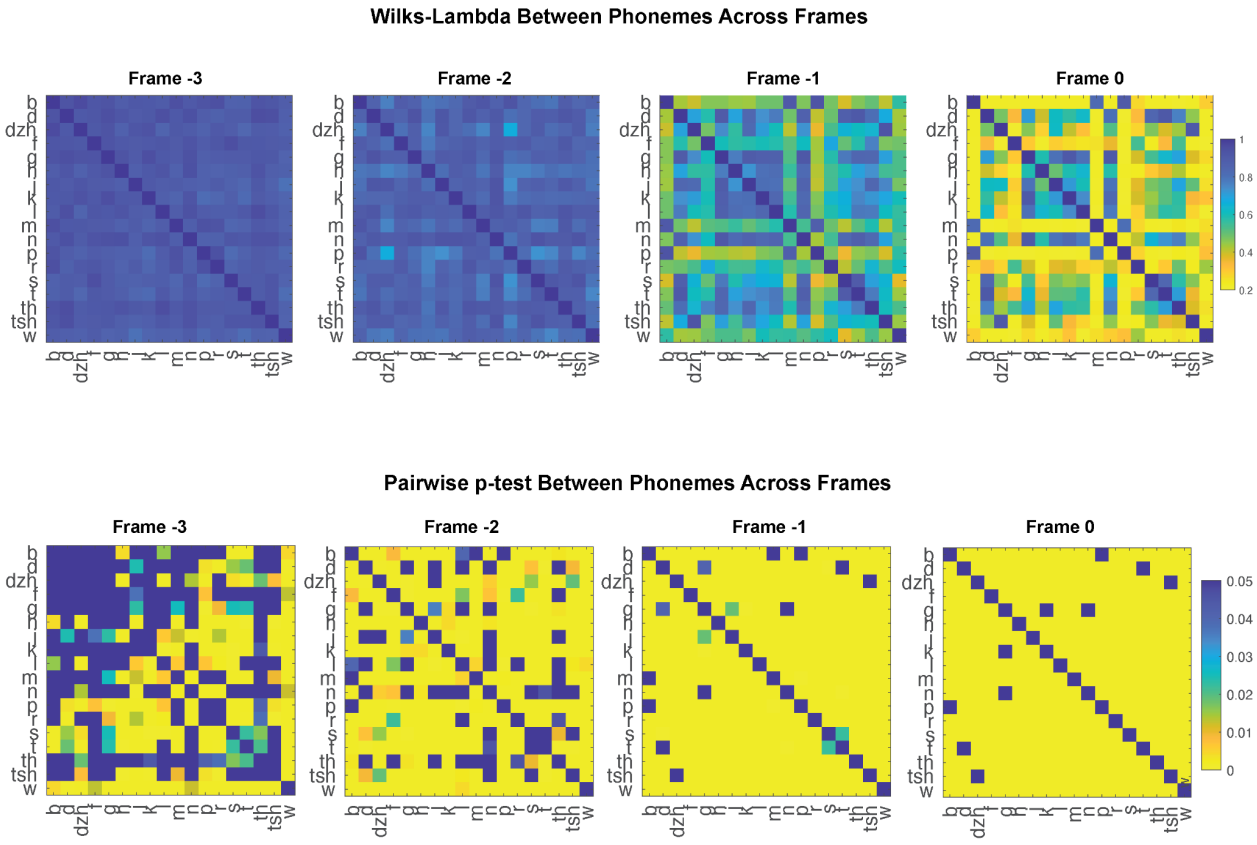


Figure 4-6 Feature distinctiveness between 18 selected phoneme categories across four consecutive frames. Wilks-Lambda effect sizes were calculated amongst all selected phonemes after aggregating keywords for each frame. Darker, colder colors indicate greater Wilks-Lambda values, and indicate lower distinctiveness (i.e. more similar to the rest). Reversely, lighter, warmer colors indicate lower Wilks-Lambda values, and indicate higher distinctiveness (i.e., more distinct). A multivariate ANOVA (MANOVA) was performed on the pairwise Wilks-Lambda scores, and corrected for multiple comparisons. Similarly, darker colors indicate more similar pairs, while lighter warmer colors indicate more distinct pairs.

4.3.3 Parametric Viseme Cluster Transitions

Our statistical significance-based clustering with a stringent threshold ($p < 0.01$) revealed a clear progression in the emergence of viseme categories as speech onset approached. The number of statistically distinct clusters identified at each time point were Frame -3 (120ms before onset): 1 cluster; Frame -2 (80ms before onset): 3 clusters; Frame -1 (40ms before onset): 10 clusters; Frame 0 (speech onset): 13 clusters. The exact membership of visemes at each frame can be seen in Figure 7. The transitioning of cluster memberships can be seen in Figure 8. This pattern demonstrates that visual distinctions between phonemes emerge gradually but rapidly in the 80ms window before auditory onset, with maximum differentiation achieved precisely at the moment of speech articulation.

Viseme categorization remained unstructured in early leadup to auditory onset. At Frame -3, all phonemes belonged to a single undifferentiated cluster, indicating minimal visual information for discrimination at this early preparatory stage. By Frame -2, three distinct clusters emerged: (a) a large group containing most consonants, and two singleton clusters for /h/ and /w/. This suggests that these two phonemes—the glottal fricative /h/ and labio-velar approximant /w/—have uniquely distinguishable visual properties even at early stages of articulation.

A substantial reorganization occurred by Frame -1, where the number of statistically distinguishable clusters increased to 10. At this point, several coherent articulatory groupings emerged: bilabials (/b/, /m/, /p/) clustered together; alveolar stops and nasals (/d/, /n/, /t/) formed another group; affricates (/dzh/, /tsh/) maintained a consistent pairing; several phonemes (/f/, /h/, /k/, /l/, /r/, /w/) formed their own singleton clusters. At Frame 0 (speech onset), further differentiation led to 13 distinct clusters, with all but five phonemes forming their own singleton categories. The remaining grouped phonemes were: bilabials (/b/, /p/); affricates (/dzh/, /tsh/), and alveolars /d/ and /t/.

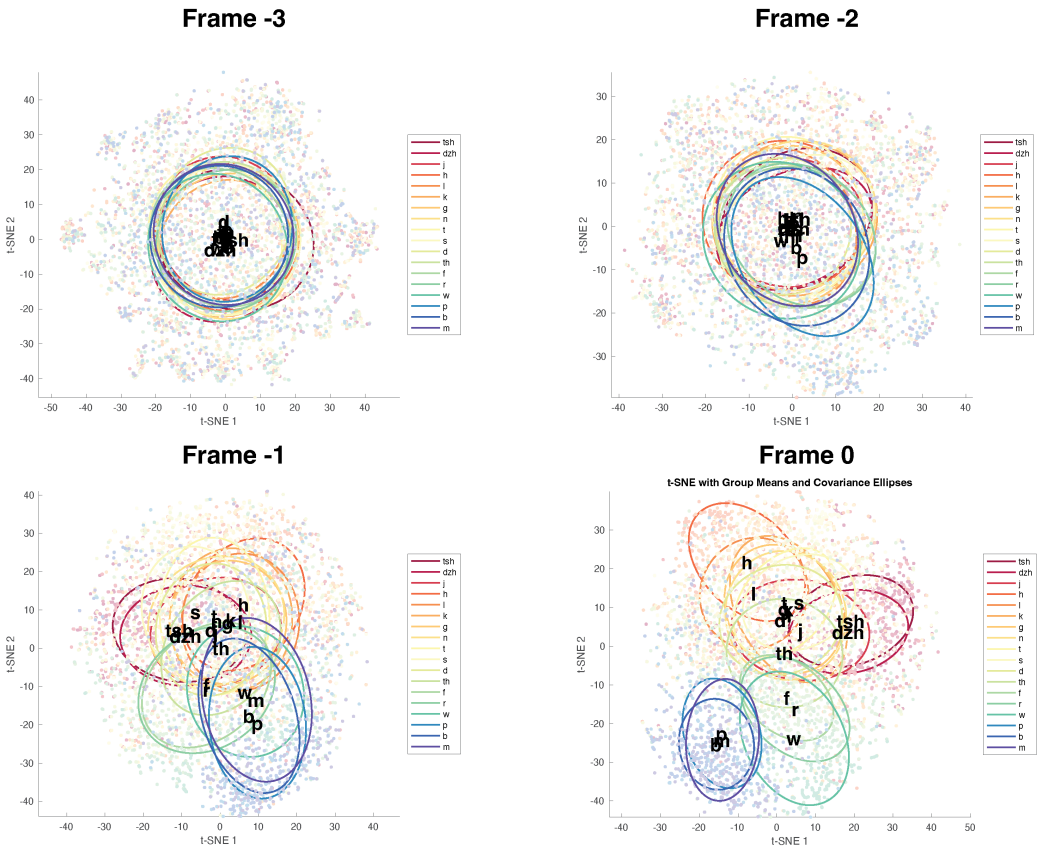


Figure 4-7 t-SNE cluster visualization of the average embeddings of the 18 phonemes. Across the 4 included frames, the phonemes were respectively grouped into 1,3, 10 and 13 clusters. Different colors in each plot symbolizes a different cluster membership. Gaussian eclipses were overlaid on top of the instances, representing covariance of all data points for the phoneme.

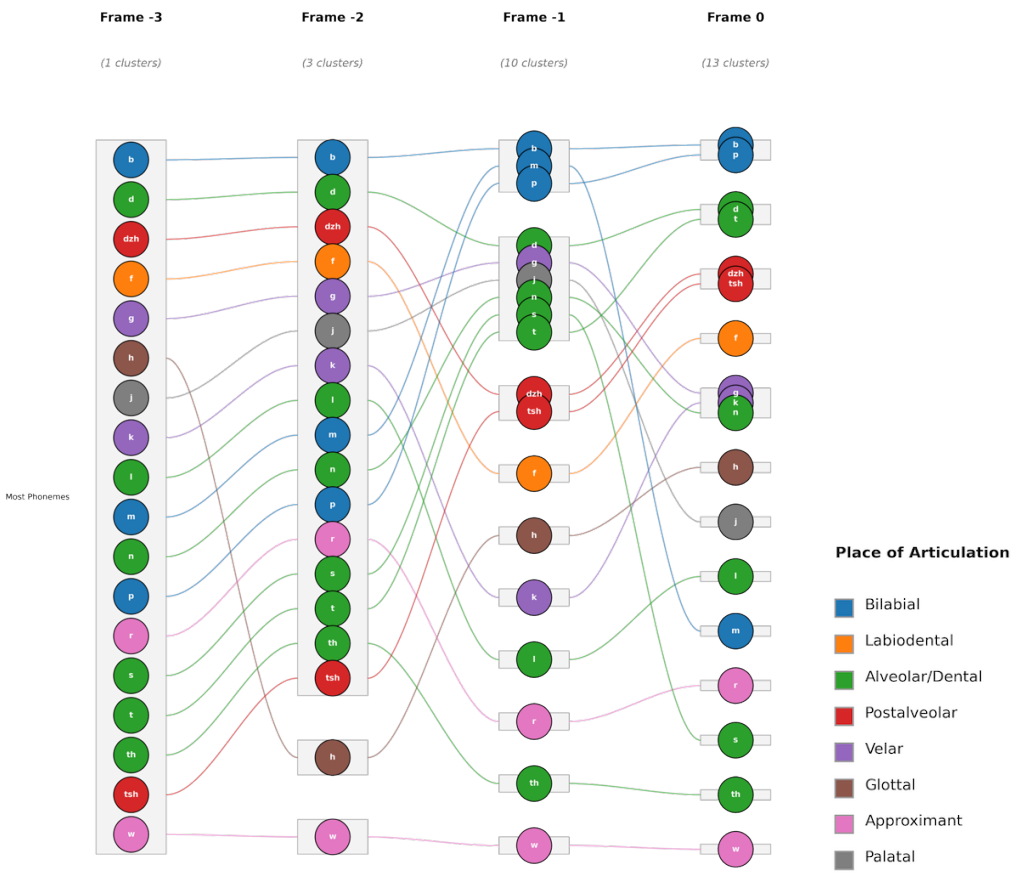


Figure 4-8 Graph theory clustering of 18 selected phoneme transitions across frames, organized by place of articulation and manner of articulation. The alluvial diagram visualizes how phonemes migrate between clusters across four frames relative to speech onset (120ms, 80ms, 40ms before onset, and at articulation onset). Each colored circle represents a phoneme, with color indicating place of articulation. The diagram shows that the information content present in the visual stream progress from coarse articulatory categories to refined phonetic clusters as articulation approaches. Place features (e.g., bilabial) become distinguishable earlier than manner features (e.g., stops, fricatives). The optimal number of clusters at each frame was determined by graph theory.

4.4 Discussion

Seeing the speaker's mouth movements help us hear better. Visible mouth movements during phoneme production, or visemes, have long been posited as a critical contributor to the visual speech benefit. However, the exact structure and similarity between visemes remain unclear due to the difficulty to perform thorough human testing. In the current study, we leverage a large audiovisual speech dataset, Lipreading in the Wild, and a deep learning network and investigated how many viseme categories there are in the physical stimuli space, to simulate the lipreading information available to a naïve listener. Among the 18 consonants we selected, our

results reveal a systematic progression in how visual phoneme representations in the physical stimulus space evolve from relatively undifferentiated patterns to highly organized, phonetically coherent structures as articulation unfolds. This temporal trajectory suggests that a hierarchical organization of visual speech information is available to listeners during audiovisual speech, wherein coarse articulatory distinctions precede fine-grained phonetic categorization.

4.4.1 From Coarse to Fine-Grained Visual Distinctions

Our parametric clustering approach shows a clear progression from undifferentiated visual information to highly specific phoneme distinctions. The increase from 1 to 13 clusters demonstrates that visual features for phoneme discrimination emerge predominantly in the 80ms window before auditory onset, with maximal differentiation achieved by the moment of speech production. This finding aligns with the visual speech head-start effect observed in neuroimaging studies, where visual information begins influencing auditory processing approximately 50-100ms before auditory onset (Karas et al., 2019).

The emergence of specific phoneme groupings in our analysis reveals that place of articulation is highly salient in visual speech. By Frame -1, bilabials (/b/, /m/, /p/) already form a distinct cluster, as do alveolars (/d/, /n/, /t/) and affricates (/dzh/, /tsh/). This pattern suggests that the visual system have access to information about articulatory configurations before auditory information becomes available.

By speech onset, our analysis identified 13 statistically distinct categories from 18 phonemes, suggesting that visual speech provides more fine-grained phonetic information than traditionally thought. However, not all phonemes were visually distinguishable - certain groups remained clustered together, particularly those sharing similar place of articulation features. This pattern of results supports a view where visual speech information is rich but still inherently ambiguous, explaining why context and auditory information remain crucial for accurate speech perception.

4.4.2 Place vs. Manner Dynamics in Visemes

Our study, for the first time, revealed differential temporal dynamics for two different types of visual phonetic information, place versus manner of articulation. Place distinctions emerged earlier (80 ms prior in current study) and show stronger organization throughout the sequence, reflecting the primary role of spatial configuration in visual speech perception. Bilabial articulations are particularly distinctive, showing clear separation even in early frames. This finding is consistent with research on the neural basis of visual speech perception, which has demonstrated that for mouth-leading words (e.g., '*chief*') vs. voice-leading words (e.g. '*knock*'), the early availability of about the place of articulation enhances responses in neurons representing phonemes that are compatible with phonemes sharing similar places of articulation and suppresses neurons representing phonemes that are incompatible (Karas et al., 2019). Manner distinctions, by contrast, emerge at later frames. The affricates (/dzh/, /tsh/), for example, remained relatively undifferentiated from other coronals in early frames but forming a distinct category by articulation onset. This pattern suggests that the dynamic components of manner features require more extensive articulatory movements to become visually salient.

4.4.3 Implications for Models of Audiovisual Integration

The progressive differentiation observed across frames has important implications for models of audiovisual speech integration. Traditional accounts (e.g. Fisher, 1968; Walden et al., 1977; Owens & Blazek, 1985) have often treated visual speech information as a static supplement to auditory processing. Our results challenge this view by demonstrating that visual speech representations are dynamically structured, with different phonetic distinctions becoming available at different time points.

These differential dynamics align with theoretical models of visual speech perception that emphasize the complementary and redundant information of auditory speech information, as suggested by Campbell (2008). Our findings suggest that this complementarity extends to temporal dynamics, with place of articulation information becoming available earlier than auditory speech onset, providing complimentary information to the auditory speech, and manner of articulation information emerges later that provides redundant information to the auditory

speech. This is in line with previous evidence in EEG audiovisual speech research that suggest that encoding of labial phoneme groups (/b/,/p/,/m/ & /f/, /v/) in electroencephalography (EEG) benefit the most from visual speech presence (O'Sullivan et al., 2020). Indeed, (/b/,/p/,/m/ & /f/, /v/) both belong to a labial phoneme category (bilabial for /b/,/p/,/m/ & labial-dental for /f/, /v/) whose place of articulation becomes available earlier in the visual stream.

Altogether, the early availability of visual place information may facilitate predictive processing, allowing the perceptual system to generate expectations about upcoming phonemes based on preparatory movements. Our finding is in line with a model where early phonetic processing occurs in the visual cortex (Nidiffer & Cao et al., 2023, Audenhaege et al., 2025), the strong visual speech benefit likely emerges from the encoding of early encoding of place of articulation in the visual cortex, that eventually reaches auditory cortex and tunes auditory perception (Metzger et al., 2020), hence the visual head-start effect (Karas et al., 2019).

4.4.4 From Static to Dynamic Visemes

Traditional approaches to visual speech have often conceptualized visemes as static articulatory configurations corresponding to groups of phonemes that are visually indistinguishable (e.g., /p/, /b/, /m/). This view suggests that "a difference between visemes is significant, informative, and categorical to the perceiver; a difference within a viseme class is not" (Massaro et al., 2012, Files et al., 2015). Our temporal analysis challenges this established notion by demonstrating that viseme organization is fundamentally dynamic, with discriminability between visually similar phonemes emerging at different time points relative to articulation.

Interestingly, our findings appear to contradict Files et al. (2015), who found that natural within-viseme stimulus pairs were discriminable in a psychophysical study. While their study suggested that humans can discriminate phonemes within the same viseme category, our analysis shows that this discriminability is likely not based on information present in the auditory stream and may be highly dependent on other linguistic information available to the listeners. This supports the hypothesis that human perceivers may rely on linguistic priors and contextual cues

from following vowels to disambiguate visually similar phonemes, a capability that purely visual analysis may not replicate (Yehia *et al.*, 1998; Munhall *et al.*, 2004).

4.4.5 Limitations

While our analysis provides novel insights into the temporal dynamics of visual speech representations, several methodological considerations warrant discussion. One, the current approach labelled visemes by their corresponding phoneme counterparts. This approach, while beneficial and necessary for hypothesis testing, does not rule out the possibility that the number of visemes might extend beyond the number of all phonemes, e.g., one phoneme might have different viseme counterparts due to the dentalization differences due to the vowel that follows. Indeed, an unsupervised classification approach unbounded by the number of phoneme classes would be more beneficial at revealing the relationship between viseme classes and phoneme identities. The k-means clustering approach imposes a fixed number of clusters at each time point, which may not perfectly capture the actual dimensionality of the embedding space. Alternative approaches, such as hierarchical clustering with variable cut points, might reveal more nuanced organizational patterns.

Additionally, our focus on isolated phonemes in controlled contexts may not fully capture the complexity of connected speech, where coarticulation effects substantially influence visual dynamics. Future research should extend this approach to continuous speech to examine how context modulates the temporal evolution of phoneme representations, as suggested by studies showing that "including wide phoneme and viseme contexts is crucial for predicting realistic lip motions".

Most importantly, while machine learning work provides important insight about existing statistical information in the physical stimuli, results in the current study are no substitute for human listening behavior in the real world. Indeed, human subjects might use additional cues like context, which results in different statistical priors about word frequency that aids speech recognition. Human listeners likely leverage additional top-down cues such as lexical expectations, syntactic structure, and world knowledge. These cues give rise to statistical priors—e.g., about phonotactic constraints or word frequency—that shape audiovisual integration in ways that purely data-driven models may not capture.

Future work would benefit from incorporating human behavioral data, such as eye-tracking or reaction time measures during audiovisual word recognition tasks, to determine how visual representations unfold over time in perception. For example, temporal generalization analyses using magnetoencephalography (MEG) or electroencephalography (EEG) could assess whether the temporal boundaries observed in our model-based embeddings align with neural or behavioral markers of perceptual decision-making. Additionally, comparing model-derived embeddings to representational similarity matrices derived from human judgments (e.g., using behavioral confusion data or fMRI-based RSA) could clarify the extent to which machine-derived visual dynamics map onto perceptual organization in the human brain.

4.5 Conclusion

Our analysis demonstrates that visual speech representations undergo a systematic evolution from coarse, broadly defined categories toward increasingly refined, phonetically coherent structures as articulation unfolds. This temporal trajectory reveals a hierarchical organization of visual speech processing, with broad distinctions preceding fine-grained categorization.

The differential dynamics for place versus manner features highlight the complex relationship between articulatory movements and visual salience, with spatial configurations becoming distinctive earlier than dynamic features. These findings enhance our understanding of how the visual system extracts phonetic information from the continuously changing patterns of articulatory movements and provide support for dynamic viseme models that capture the temporal nature of visual speech units.

Beyond theoretical implications, these results have potential applications for audiovisual speech recognition systems and visual speech synthesis, suggesting that temporal dynamics should be explicitly incorporated into computational models. By accounting for the progressive differentiation of visual speech representations, such models might achieve more robust performance, particularly in challenging acoustic environments.

Chapter 5 General Discussion

5.1 Summary of Findings

The overarching goal of this dissertation was to examine the role of visual modulations on auditory speech perception. We first tested how auditory speech onset information from the visual domain is encoded in the auditory cortex. We then established that temporal and spectral information from the auditory speech is restored in the auditory cortex. Lastly, we attempted to settle a debate regarding how many viseme categories are available to listeners in naturalistic speech. Together, these studies provide a comprehensive view of the neural mechanisms underlying audiovisual speech integration and the functional role of visual speech in auditory perception.

In Chapter 2, we investigated how visual speech modulates temporal dynamics both before and after speech onset in the superior temporal gyrus (STG) using intracranial EEG recordings. We found that visual information has two distinct effects on auditory processing: (1) a phase reset in theta band oscillations across the STG before auditory speech onset, and (2) decreased theta power in the posterior STG after speech onset. Critically, we revealed that these effects were dissociable at both group and individual levels, with no significant correlation between them, suggesting independent neural mechanisms supporting audiovisual integration before and after speech onset.

In Chapter 3, we examined how the auditory system processes visual speech information differentially based on the type of degraded auditory information. With fMRI and a large pool of subjects, we acquired the necessary spatial resolution to test hypotheses on the spatial distribution of activity. Using degraded speech stimuli (either temporal or spectral information), we found that visual speech improved behavioral recognition of speech embedded in both noise types. However, the neural responses in auditory regions showed a general rather than feature-specific restoration effect, with audiovisual trials engaging distinct neuronal populations rather than shifting activation patterns to resemble clear speech. These findings suggest that visual

speech provides general compensatory benefits regardless of which acoustic features are degraded.

In Chapter 4, we investigated the emergence of viseme categories (visual units of phonemes) by leveraging deep learning approaches applied to a large audiovisual dataset. Our results revealed a dynamic progression from coarse articulatory distinctions to refined phonetic categories as articulation unfolds. Place features (e.g., bilabial, coronal) became distinguishable earlier than manner features (e.g., stops, fricatives), with optimal category discrimination occurring immediately before auditory onset. We identified 10 statistically distinguishable viseme clusters, supporting a model where visual speech provides complementary information to the auditory stream that aids speech perception.

5.2 Theoretical Integration and Implications

Collectively, these studies advance our understanding of audiovisual speech processing in several important ways. First, they establish that visual influences on auditory processing operate through at least two distinct neural mechanisms rather than a single integrated process. The iEEG findings in Chapter 2 reveal separate pathways for pre-speech visual modulation (theta band phase reset) and post-speech integration (theta band power modulation), which operate independently. This functional separation is further supported by the fMRI results in Chapter 3, showing that visual speech engages distinct populations in the auditory cortex rather than shifting activity patterns to mimic that of clear speech. Together, these two studies support a multi-stage model of audiovisual speech integration, in which: visual timing information primes auditory cortex for incoming speech (Study 1), and that vision recruits additional auditory cortex populations when the signal is degraded (Study 2).

This model is consistent with predictive encoding accounts of speech processing, in which vision provides prior constraints not only on when the speech will occur, but also on what speech sounds are likely, tuned by perceptual and linguistic experience. Visual speech thus not only enhances low-level feature recovery, but potentially higher-level speech features (e.g. syntactic, semantic) as well.

Second, this research clarifies the temporal dynamics of visual speech information processing. Chapter 4 demonstrates that visual articulatory information unfolds progressively,

with place features becoming available earlier than manner features. This timing aligns with the theta phase reset observed in Chapter 2, which occurs before auditory onset and likely enables the auditory system to leverage early place-of-articulation information from visible mouth movements. The subsequent decreases in theta power after speech onset may reflect the integration of more fine-grained manner features that become available later in the articulation sequence.

Third, these findings challenge prevailing models of audiovisual integration that posit feature-specific restoration mechanisms. Rather than selectively compensating for specific degraded acoustic features, visual speech appears to provide a general enhancement that operates at a more abstract representational level. This is evidenced by the similar behavioral benefits observed for both temporally and spectrally degraded speech in Chapter 3, and by the engagement of distinct rather than overlapping neural populations during audiovisual processing.

This work also has important methodological implications. By combining multiple techniques (iEEG, fMRI, computational modeling) and levels of analysis (neural oscillations, BOLD signals, viseme categories), this dissertation demonstrates the value of a multi-method approach to understanding complex cognitive processes. Each method reveals different aspects of audiovisual integration that would not be available from a single methodological perspective.

5.3 Limitations

While this dissertation provides significant advances in understanding audiovisual speech processing, several limitations should be acknowledged.

In Chapter 2, while our analyses focused on inter-trial phase coherence (ITPC) in the theta band (4–8 Hz) prior to speech onset, the use of a 500 ms pre-onset window provides only 2–4 cycles for the theta rhythm. This limited cycle count may constrain the precision and reliability of phase-based metrics (Lachaux et al., 1999), and future studies should explore longer preparatory windows or complementary frequency bands (e.g., delta) for more robust phase estimates.

Second, we did not observe consistent modulations in high gamma power (HGp), which is more directly associated with neuronal firing and often aligns better with BOLD responses in fMRI (Mukamel et al., 2005; Ray & Maunsell, 2011). The absence of HGp effects limits the

direct comparability of our iEEG findings to Study 2 and may suggest that the visual timing mechanisms identified here reflect non-spiking processes such as subpopulation synchronization.

Finally, the stimuli used in this study were restricted to words beginning with a subset of phonemes (/b/, /f/, /g/, /d/), selected to ensure sufficient trial counts and clarity in audiovisual alignment. While this design enables strong control over stimulus timing and phonemic balance, it limits the generality of our conclusions about audiovisual speech processing across the broader phoneme space. Future work should extend these findings using more ecologically valid and phonetically diverse stimuli.

In Chapter 3, while it was conceptually beneficial to label the conditions as temporally and spectrally degraded, the two conditions are not entirely orthogonal. Indeed, when temporal information is filtered out of the stimuli, the spectral content is slightly degraded as well, and vice versa. Additionally, while adept at capturing high spatial resolutions of activations of superior temporal cortices including the STG and pSTS, our study was limited by the temporal resolution of fMRI, which cannot capture the fine-grained temporal dynamics (e.g. phoneme-level) of audiovisual integration similar to what was revealed in the iEEG data from Chapter 2. Future studies could address this limitation by combining the spatial precision of fMRI with the temporal resolution of iEEG through the development of encoding models that map between these two domains. Such models could reveal how the rapid oscillatory dynamics observed in Chapter 2 relate to the activation patterns and neural population responses identified in Chapter 3.

For Chapter 4, while our computational approach allowed us to analyze viseme categories with unprecedented detail, we have yet to examine the auditory counterpart of this viseme analysis. Future work combining analyses of both visual and auditory features would reveal which viseme clusters benefit most from audiovisual integration and under what conditions. Additionally, it is important to reiterate that the findings from our computational modeling work are meant to guide, but directly substitute, human perceptual behavior.

5.4 Future Directions

Several promising directions for future research emerge from this work. First, examining how individual differences in lip-reading ability relate to the neural mechanisms identified in

Chapters 2 and 3 could clarify the relationship between behavioral skills and neural processes. Second, investigating how these audiovisual integration mechanisms are affected in clinical populations with speech or hearing impairments could lead to targeted interventions. Third, extending the temporal dynamics analysis to connected speech rather than isolated words would provide insights into how visual information influences auditory processing during continuous discourse.

Additionally, future studies could explore how attention modulates these audiovisual integration processes. The theta phase reset observed in Chapter 2 may be influenced by attentional allocation, and understanding this relationship could clarify the role of top-down processes in audiovisual speech perception. Finally, developmental studies examining how these integration mechanisms emerge and change throughout childhood could provide valuable insights into critical periods for audiovisual speech learning.

Finally, the computational modeling results will benefit greatly from being compared against human listener testing. Our results revealed that certain articulatory information, including place of articulation, is available to the listeners before auditory speech onset. However, we do not know whether or how listeners make use of that information. It would be informative to perform phoneme-level clustering on existing speech-in-noise behavioral data from human subjects. Likely, listeners do not use all the statistical information in the visual scene as it is not computationally efficient and instead rely on higher-level linguistic priors. It would be informative to modify the model (e.g., by degrading the visual input) and see which version yields outputs most similar to those of human listeners.

5.5 Conclusion

This dissertation provides a thorough examination of how visual speech information influences auditory processing across multiple levels of analysis. The findings reveal that visual speech affects auditory cortex through distinct, independent mechanisms that operate at different time points relative to speech onset. Rather than selectively restoring specific degraded acoustic features, visual speech appears to provide general enhancement through engagement of distinct neural populations. Furthermore, we revealed that the temporal dynamics of viseme category

discrimination align with these neural mechanisms, with place of articulation features becoming available earlier than manner features.

These insights advance our understanding of the fundamental processes underlying the behavioral and neural dynamics of audiovisual speech perception and have implications for theoretical models of multisensory integration. By clarifying how the brain leverages available visual information to enhance speech encoding, this work may ultimately inform the development of interventions for individuals with hearing impairments and contribute to improvements in speech recognition technologies that incorporate both auditory and visual information.

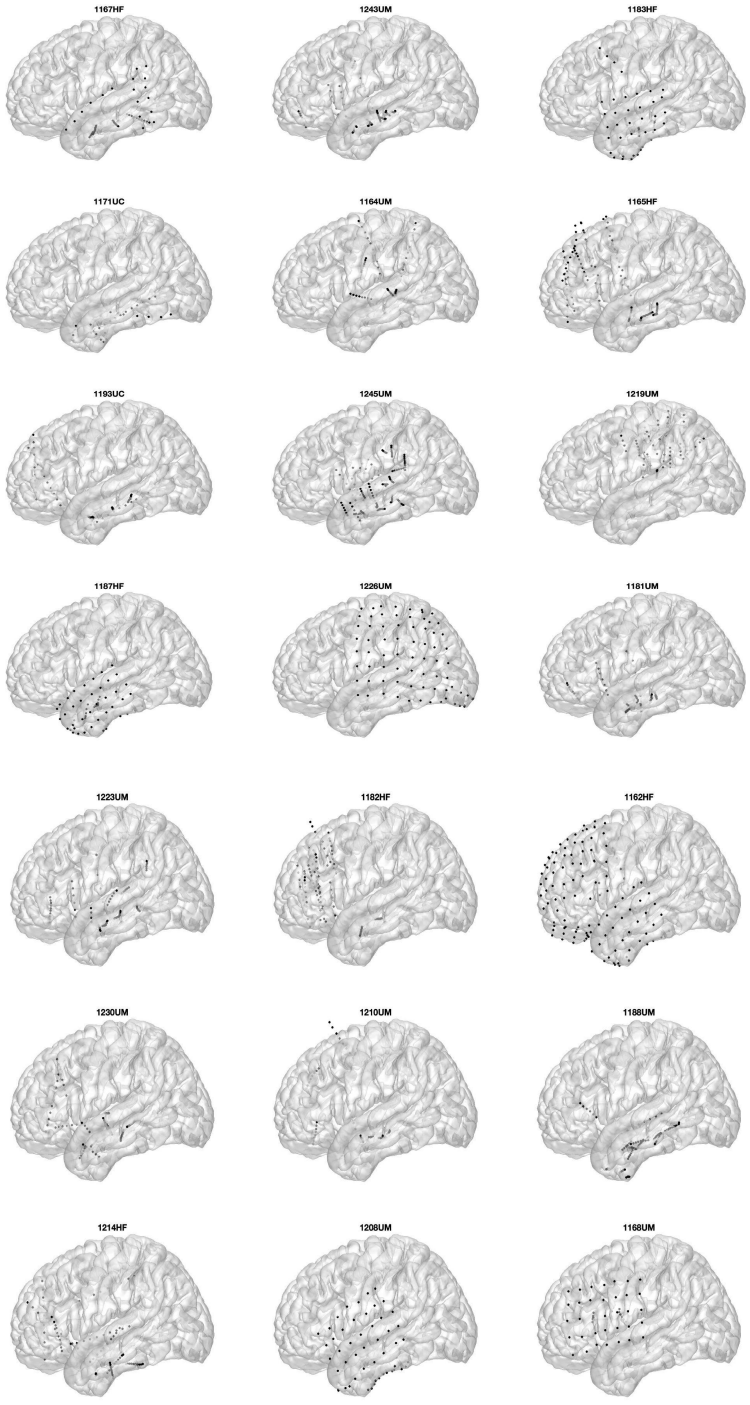
Appendices

Appendix A: Word stimuli used in the iEEG study

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

Appendix Table A-1 Word Stimuli used in iEEG study

Appendix B: iEEG Individual Subject Coverage



Appendix Figure B-1 Individual subject electrode coverage demonstration.

Appendix C: fMRI Decoding Results

Condition	Left STG	Right STG	Left pSTS	Right pSTS
Aud_AM	0.658 (p=6.49e-21)	0.663 (p=6.17e-18)	0.734 (p=2.28e-15)	0.774 (p=9.34e-14)
AV_AM	0.658 (p=1.26e-22)	0.663 (p=1.54e-21)	0.734 (p=4.79e-28)	0.774 (p=3.87e-30)
Aud_FM	0.658 (p=8.20e-31)	0.663 (p=1.47e-34)	0.734 (p=4.69e-12)	0.774 (p=6.12e-15)
AV_FM	0.658 (p=1.52e-29)	0.663 (p=6.17e-31)	0.734 (p=3.08e-28)	0.774 (p=2.76e-31)

Appendix Table C-1 Two class SVM decoding accuracy between auditory degraded conditions against auditory-clear in left and right STG and STS at the group level (*n*= 64).

Comparison	Left STG	Right STG	Left pSTS	Right pSTS
AV_AM vs Aud_AM	0.103 (p=0.436)	0.140 (p=0.235)	-0.002 (p=1.85e-15)*	0.012 (p=3.58e-17)*
AV_FM vs Aud_FM	0.106 (p=0.231)	0.119 (p=0.735)	0.015 (p=4.64e-16)*	0.024 (p=9.05e-19)*
Aud_AM vs Aud_FM	0.103 (p=1.83e-10)*	0.140 (p=1.67e-15)*	-0.002 (p=0.843)	0.012 (p=0.373)
AV_AM vs AV_FM	0.106 (p=6.33e-11)*	0.119 (p=1.91e-11)*	0.015 (p=0.187)	0.024 (p=0.033)*

Appendix Table C-2 Two class SVM decoding accuracy between auditory degraded conditions in left and right STG and STS at the group level (*n*= 64).

*Statistically significant at $p < 0.05$

Appendix D: Viseme Study Word Stimuli List

ABOUT	COUNCIL	HISTORY	OPPOSITION	SOMETHING
ABSOLUTELY	COUNTRIES	HOMES	ORDER	SOUTH
ABUSE	COUNTRY	HOSPITAL	OTHER	SOUTHERN
ACCESS	COUPLE	HOURS	OTHERS	SPEAKING
ACCORDING	COURSE	HOUSE	OUTSIDE	SPECIAL
ACCUSED	COURT	HOUSING	PARENTS	SPEECH
ACROSS	CRIME	HUMAN	PARLIAMENT	SPEND
ACTION	CRISIS	HUNDREDS	PARTIES	SPENDING
ACTUALLY	CURRENT	IMMIGRATION	PARTS	SPENT
AFFAIRS	CUSTOMERS	IMPACT	PARTY	STAFF
AFFECTED	DAVID	IMPORTANT	PATIENTS	STAGE
AFRICA	DEATH	INCREASE	PAYING	STAND
AFTER	DEBATE	INDEPENDENT	PEOPLE	START
AFTERNOON	DECIDED	INDUSTRY	PERHAPS	STARTED
AGAIN	DECISION	INFLATION	PERIOD	STATE
AGAINST	DEFICIT	INFORMATION	PERSON	STATEMENT
AGREE	DEGREES	INQUIRY	PERSONAL	STATES
AGREEMENT	DESCRIBED	INSIDE	PHONE	STILL
AHEAD	DESPITE	INTEREST	PLACE	STORY
ALLEGATIONS	DETAILS	INVESTMENT	PLACES	STREET
ALLOW	DIFFERENCE	INVOLVED	PLANS	STRONG
ALLOWED	DIFFERENT	IRELAND	POINT	SUNDAY
ALMOST	DIFFICULT	ISLAMIC	POLICE	SUNSHINE
ALREADY	DOING	ISSUE	POLICY	SUPPORT
ALWAYS	DURING	ISSUES	POLITICAL	SYRIA
AMERICA	EARLY	ITSELF	POLITICIANS	SYRIAN
AMERICAN	EASTERN	JAMES	POLITICS	SYSTEM
AMONG	ECONOMIC	JUDGE	POSITION	TAKEN
AMOUNT	ECONOMY	JUSTICE	POSSIBLE	TAKING
ANNOUNCED	EDITOR	KILLED	POTENTIAL	TALKING

ANOTHER	EDUCATION	KNOWN	POWER	TALKS
ANSWER	ELECTION	LABOUR	POWERS	TEMPERATURES
ANYTHING	EMERGENCY	LARGE	PRESIDENT	TERMS
AREAS	ENERGY	LATER	PRESS	THEIR
AROUND	ENGLAND	LATEST	PRESSURE	THEMSELVES
ARRESTED	ENOUGH	LEADER	PRETTY	THERE
ASKED	EUROPE	LEADERS	PRICE	THESE
ASKING	EUROPEAN	LEADERSHIP	PRICES	THING
ATTACK	EVENING	LEAST	PRIME	THINGS
ATTACKS	EVENTS	LEAVE	PRISON	THINK
AUTHORITIES	EVERY	LEGAL	PRIVATE	THIRD
BANKS	EVERYBODY	LEVEL	PROBABLY	THOSE
BECAUSE	EVERYONE	LEVELS	PROBLEM	THOUGHT
BECOME	EVERYTHING	LIKELY	PROBLEMS	THOUSANDS
BEFORE	EVIDENCE	LITTLE	PROCESS	THREAT
BEHIND	EXACTLY	LIVES	PROTECT	THREE
BEING	EXAMPLE	LIVING	PROVIDE	THROUGH
BELIEVE	EXPECT	LOCAL	PUBLIC	TIMES
BENEFIT	EXPECTED	LONDON	QUESTION	TODAY
BENEFITS	EXTRA	LONGER	QUESTIONS	TOGETHER
BETTER	FACING	LOOKING	QUITE	TOMORROW
BETWEEN	FAMILIES	MAJOR	RATES	TONIGHT
BIGGEST	FAMILY	MAJORITY	RATHER	TOWARDS
BILLION	FIGHT	MAKES	REALLY	TRADE
BLACK	FIGHTING	MAKING	REASON	TRIAL
BORDER	FIGURES	MANCHESTER	RECENT	TRUST
BRING	FINAL	MARKET	RECORD	TRYING
BRITAIN	FINANCIAL	MASSIVE	REFERENDUM	UNDER
BRITISH	FIRST	MATTER	REMEMBER	UNDERSTAND
BROUGHT	FOCUS	MAYBE	REPORT	UNION
BUDGET	FOLLOWING	MEANS	REPORTS	UNITED
BUILD	FOOTBALL	MEASURES	RESPONSE	UNTIL
BUILDING	FORCE	MEDIA	RESULT	USING
BUSINESS	FORCES	MEDICAL	RETURN	VICTIMS
BUSINESSES	FOREIGN	MEETING	RIGHT	VIOLENCE
CALLED	FORMER	MEMBER	RIGHTS	VOTERS
CAMERON	FORWARD	MEMBERS	RULES	WAITING
CAMPAIGN	FOUND	MESSAGE	RUNNING	WALES
CANCER	FRANCE	MIDDLE	RUSSIA	WANTED

CANNOT	FRENCH	MIGHT	RUSSIAN	WANTS
CAPITAL	FRIDAY	MIGRANTS	SAYING	WARNING
CASES	FRONT	MILITARY	SCHOOL	WATCHING
CENTRAL	FURTHER	MILLION	SCHOOLS	WATER
CERTAINLY	FUTURE	MILLIONS	SCOTLAND	WEAPONS
CHALLENGE	GAMES	MINISTER	SCOTTISH	WEATHER
CHANCE	GENERAL	MINISTERS	SECOND	WEEKEND
CHANGE	GEORGE	MINUTES	SECRETARY	WEEKS
CHANGES	GERMANY	MISSING	SECTOR	WELCOME
CHARGE	GETTING	MOMENT	SECURITY	WELFARE
CHARGES	GIVEN	MONEY	SEEMS	WESTERN
CHIEF	GIVING	MONTH	SENIOR	WESTMINSTER
CHILD	GLOBAL	MONTHS	SENSE	WHERE
CHILDREN	GOING	MORNING	SERIES	WHETHER
CHINA	GOVERNMENT	MOVING	SERIOUS	WHICH
CLAIMS	GREAT	MURDER	SERVICE	WHILE
CLEAR	GREECE	NATIONAL	SERVICES	WHOLE
CLOSE	GROUND	NEEDS	SEVEN	WINDS
CLOUD	GROUP	NEVER	SEVERAL	WITHIN
COMES	GROWING	NIGHT	SHORT	WITHOUT
COMING	GROWTH	NORTH	SHOULD	WOMEN
COMMUNITY	GUILTY	NORTHERN	SIDES	WORDS
COMPANIES	HAPPEN	NOTHING	SIGNIFICANT	WORKERS
COMPANY	HAPPENED	NUMBER	SIMPLY	WORKING
CONCERNS	HAPPENING	NUMBERS	SINCE	WORLD
CONFERENCE	HAVING	OBAMA	SINGLE	WORST
CONFLICT	HEALTH	OFFICE	SITUATION	WOULD
CONSERVATIVE	HEARD	OFFICERS	SMALL	WRONG
CONTINUE	HEART	OFFICIALS	SOCIAL	YEARS
CONTROL	HEAVY	OFTEN	SOCIETY	YESTERDAY
COULD	HIGHER	OPERATION	SOMEONE	YOUNG

Appendix Table D-1 Extensive list of all word used for training from the Lipreading in the Wild dataset.

Bibliography

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2008). Right-Hemisphere Auditory Cortex Is Dominant for Coding Syllable Patterns in Speech. *Journal of Neuroscience*, *28*(15), 3958–3965. https://doi.org/10.1523/JNEUROSCI.0187-08.2008

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, *98*(23), 13367–13372. https://doi.org/10.1073/pnas.201400998

Ahn, E., Majumdar, A., Lee, T. G., & Brang, D. (2024). *Evidence for a Causal Dissociation of the McGurk Effect and Congruent Audiovisual Speech Perception via TMS to the Left pSTS*. https://doi.org/10.1163/22134808-bja10129

Aiken, S. J., & Picton, T. W. (2008). Human Cortical Responses to the Speech Envelope. *Ear and Hearing*, *29*(2), 139–157. https://doi.org/10.1097/AUD.0b013e31816453dc

Aller, M., Økland, H. S., MacGregor, L. J., Blank, H., & Davis, M. H. (2022). Differential Auditory and Visual Phase-Locking Are Observed during Audio-Visual Benefit and Silent Lip-Reading for Speech Perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *42*(31), 6108–6120. https://doi.org/10.1523/JNEUROSCI.2476-21.2022

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398. https://doi.org/10.1016/j.tics.2012.05.003

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A.-L. (2009). Dual Neural Routing of Visual Facilitation in Speech Processing. *Journal of Neuroscience*, *29*(43), 13445–13453. https://doi.org/10.1523/JNEUROSCI.3194-09.2009

Audenhaege, A. V., Mattioni, S., Cerpelloni, F., Gau, R., Szmalec, A., & Collignon, O. (2025). Phonological representations of auditory and visual speech in the occipito-temporal cortex and beyond. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.1415-24.2025

Auer, E. T., & Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *Journal of Speech, Language, and Hearing Research: JSLHR*, *50*(5), 1157–1165. https://doi.org/10.1044/1092-4388(2007/080)

Banks, M. I., Uhlrich, D. J., Smith, P. H., Krause, B. M., & Manning, K. A. (2011). Descending Projections from Extrastriate Visual Cortex Modulate Responses of Cells in Primary Auditory Cortex. *Cerebral Cortex*, *21*(11), 2620–2638. https://doi.org/10.1093/cercor/bhr048

Bernstein, L. E., Eberhardt, S. P., & Demorest, M. E. (1989). Single-channel vibrotactile supplements to visual perception of intonation and stress. *The Journal of the Acoustical Society of America*, *85*(1), 397–405. https://doi.org/10.1121/1.397690

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00386

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M.-H. (2008). Visual Activation and Audiovisual Interactions in the Auditory Cortex during Speech Perception: Intracranial Recordings in Humans. *Journal of Neuroscience*, *28*(52), 14301–14310. https://doi.org/10.1523/JNEUROSCI.2875-08.2008

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research*, *17*(4), 619–630. https://doi.org/10.1044/jshr.1704.619

Bishop, C. W., & Miller, L. M. (2009). A Multisensory Cortical Network for Understanding Speech in Noise. *Journal of Cognitive Neuroscience*, *21*(9), 1790–1804. https://doi.org/10.1162/jocn.2009.21118

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-Reading Enables the Brain to Synthesize Auditory Features of Unknown Silent Speech. *Journal of Neuroscience*, *40*(5), 1053–1065. https://doi.org/10.1523/JNEUROSCI.1101-19.2019

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Brang, D., Plass, J., Sherman, A., Stacey, W. C., Wasade, V. S., Grabowecky, M., Ahn, E., Towle, V. L., Tao, J. X., Wu, S., Issa, N. P., & Suzuki, S. (2022). Visual cortex responds to sound onset and offset during passive listening. *Journal of Neurophysiology*, *127*(6), 1547–1563. https://doi.org/10.1152/jn.00164.2021

Bröhl, F., & Kayser, C. (2021). Delta/theta band EEG differentially tracks low and high frequency speech-derived envelopes. *NeuroImage*, *233*, 117958. https://doi.org/10.1016/j.neuroimage.2021.117958

Bröhl, F., Keitel, A., & Kayser, C. (2022). MEG Activity in Visual and Auditory Cortices Represents Acoustic Speech-Related Information during Silent Lip Reading. *eNeuro*, *9*(3). https://doi.org/10.1523/ENEURO.0209-22.2022

Buchan, J. N., Paré, M., & Munhall, K. G. (2008). The Effect of Varying Talker Identity and Listening Conditions on Gaze Behavior During Audiovisual Speech Perception. *Brain Research*, *1242*, 162–171. https://doi.org/10.1016/j.brainres.2008.06.083

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204–220. https://doi.org/10.1002/dev.20032

Byczynski, G., & Park, H. (2025). *Attentional inhibition by alpha power is modulated by faster theta rhythm and audio-visual congruency during natural speech perception* (p. 2024.02.25.581860). bioRxiv. https://doi.org/10.1101/2024.02.25.581860

Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of Audio-Visual Integration Sites in Humans by Application of Electrophysiological Criteria to the BOLD Effect. *NeuroImage*, *14*(2), 427–438. https://doi.org/10.1006/nimg.2001.0812

Campbell, R. (2008). The processing of audio-visual speech: Empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1001–1010. https://doi.org/10.1098/rstb.2007.2155

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The Natural Statistics of Audiovisual Speech. *PLOS Computational Biology*, *5*(7), e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Collins, D. L., Neelin, P., Peters, T. M., & Evans, A. C. (1994). Automatic 3D Intersubject Registration of MR Volumetric Data in Standardized Talairach Space. *Journal of Computer Assisted Tomography*, *18*(2), 192.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. *Journal of Neuroscience*, *35*(42), 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015

Crosse, M. J., Liberto, G. M. D., & Lalor, E. C. (2016). Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration. *Journal of Neuroscience*, *36*(38), 9888–9895. https://doi.org/10.1523/JNEUROSCI.1396-16.2016

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Davis, C., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition*, *100*(3), B21–B31. https://doi.org/10.1016/j.cognition.2005.09.002

Desai, M., Field, A. M., & Hamilton, L. S. (2024). A comparison of EEG encoding models using audiovisual stimuli and their unimodal counterparts. *PLOS Computational Biology*, *20*(9), e1012433. https://doi.org/10.1371/journal.pcbi.1012433

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), Article 1. https://doi.org/10.1038/nn.4186

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85*, 761–768. https://doi.org/10.1016/j.neuroimage.2013.06.035

Drijvers, L., & Özyürek, A. (2017). Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Duecker, K., Doelling, K. B., Breska, A., Coffey, E. B., Sivarao, D. V., & Zoefel, B. (2024). Challenges and Approaches in the Study of Neural Entrainment. *Journal of Neuroscience*, *44*(40). https://www.jneurosci.org/content/44/40/e1234242024.abstract

Elliott, T. M., & Theunissen, F. E. (2009). The Modulation Transfer Function for Speech Intelligibility. *PLOS Computational Biology*, *5*(3), e1000302. https://doi.org/10.1371/journal.pcbi.1000302

Erber, N. P. (1975). Auditory-Visual Perception of Speech. *Journal of Speech and Hearing Disorders*, *40*(4), 481–492. https://doi.org/10.1044/jshd.4004.481

Etard, O., & Reichenbach, T. (2019). Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *39*(29), 5750–5759. https://doi.org/10.1523/JNEUROSCI.1828-18.2019

Files, B. T., Tjan, B. S., Jiang, J., & Bernstein, L. E. (2015). Visual speech discrimination and identification of natural and synthetic consonant stimuli. *Frontiers in Psychology*, *6*. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00878

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Fisher, C. G. (1968). Confusions Among Visually Perceived Consonants. *Journal of Speech and Hearing Research*, *11*(4), 796–804. https://doi.org/10.1044/jshr.1104.796

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L., & Brain Development Cooperative Group. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, *54*(1), 313–327. https://doi.org/10.1016/j.neuroimage.2010.07.033

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210. https://doi.org/10.1002/hbm.460020402

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory Integration of Dynamic Faces and Voices in Rhesus Monkey Auditory Cortex. *Journal of Neuroscience*, *25*(20), 5004–5012. https://doi.org/10.1523/JNEUROSCI.0799-05.2005

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

Goldin-Meadow, S., & Wagner, S. M. (2005). How our hands help us learn. *Trends in Cognitive Sciences*, *9*(5), 234–241. https://doi.org/10.1016/j.tics.2005.03.006

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*, *103*(5), 2677–2690. https://doi.org/10.1121/1.422788

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. *Psychophysiology*, *48*(12), 1726–1737. https://doi.org/10.1111/j.1469-8986.2011.01272.x

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-PIE. *Proceedings of the ... International Conference on Automatic Face and Gesture Recognition. IEEE International Conference on Automatic Face & Gesture Recognition*, *28*(5), 807–813. https://doi.org/10.1016/j.imavis.2009.08.002

Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception, & Psychophysics*, *77*(4), 1333–1341. https://doi.org/10.3758/s13414-014-0821-1

Hagler, D. J., Saygin, A. P., & Sereno, M. I. (2006). Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. *NeuroImage*, *33*(4), 1093–1103. https://doi.org/10.1016/j.neuroimage.2006.07.036

Haider, C. L., Park, H., Hauswald, A., & Weisz, N. (2024). Neural Speech Tracking Highlights the Importance of Visual Speech in Multi-speaker Situations. *Journal of Cognitive Neuroscience*, *36*(1), 128–142. https://doi.org/10.1162/jocn_a_02059

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Current Biology: CB*, *28*(12), 1860-1871.e4. https://doi.org/10.1016/j.cub.2018.04.033

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582. https://doi.org/10.1080/23273798.2018.1499946

Helfrich, R. F., Fiebelkorn, I. C., Szczepanski, S. M., Lin, J. J., Parvizi, J., Knight, R. T., & Kastner, S. (2018). Neural Mechanisms of Sustained Attention Are Rhythmic. *Neuron*, *99*(4), 854-865.e5. https://doi.org/10.1016/j.neuron.2018.07.032

Hickok, G., Rogalsky, C., Matchin, W., Basilakos, A., Cai, J., Pillay, S., Ferrill, M., Mickelsen, S., Anderson, S. W., Love, T., Binder, J., & Fridriksson, J. (2018). Neural networks supporting audiovisual integration for speech: A large-scale lesion study. *Cortex*, *103*, 360–371. https://doi.org/10.1016/j.cortex.2018.03.030

Hilder, S., Harvey, R. W., & Theobald, B. J. (2009, September). Comparison of human and machine-based lip-reading. In *AVSP* (pp. 86-89).

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived

from Speech Stimuli. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *36*(6), 2014–2026. https://doi.org/10.1523/JNEUROSCI.1779-15.2016

Karas, P. J., Magnotti, J. F., Metzger, B. A., Zhu, L. L., Smith, K. B., Yoshor, D., & Beauchamp, M. S. (2019). The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. *eLife*, *8*, e48116. https://doi.org/10.7554/eLife.48116

Karthik, G., Cao, C. Z., Demidenko, M. I., Jahn, A., Stacey, W. C., Wasade, V. S., & Brang, D. (2024). Auditory cortex encodes lipreading information through spatially distributed activity. *Current Biology*, *34*(17), 4021-4032.e5. https://doi.org/10.1016/j.cub.2024.07.073

Karthik, G., Plass, J., Beltz, A. M., Liu, Z., Grabowecky, M., Suzuki, S., Stacey, W. C., Wasade, V. S., Towle, V. L., Tao, J. X., Wu, S., Issa, N. P., & Brang, D. (2021). Visual speech differentially modulates beta, theta, and high gamma bands in auditory cortex. *European Journal of Neuroscience*, *54*(9), 7301–7317. https://doi.org/10.1111/ejn.15482

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLOS Biology*, *16*(3), e2004473. https://doi.org/10.1371/journal.pbio.2004473

Lachaux, J. P., Rodriguez, E., Martinerie, J., & Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Human Brain Mapping*, *8*(4), 194–208. https://doi.org/10.1002/(sici)1097-0193(1999)8:4<194::aid-hbm4>3.0.co;2-c

Leonard, M. K., Gwilliams, L., Sellers, K. K., Chung, J. E., Xu, D., Mischler, G., Mesgarani, N., Welkenhuysen, M., Dutta, B., & Chang, E. F. (2024). Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature*, *626*(7999), 593–602. https://doi.org/10.1038/s41586-023-06839-2

Luo, H., Liu, Z., & Poeppel, D. (2010). Auditory Cortex Tracks Both Auditory and Visual Stimulus Dynamics Using Low-Frequency Neuronal Phase Modulation. *PLOS Biology*, *8*(8), e1000445. https://doi.org/10.1371/journal.pbio.1000445

Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. *Trends in Cognitive Sciences*, *8*(5), 204–210. https://doi.org/10.1016/j.tics.2004.03.008

Martinez, B., Ma, P., Petridis, S., & Pantic, M. (2020). *Lipreading using Temporal Convolutional Networks* (arXiv:2001.08702). arXiv. https://doi.org/10.48550/arXiv.2001.08702

Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *The Journal of the Acoustical Society of America*, *100*(3), 1777–1786. https://doi.org/10.1121/1.417342

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). *Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi*. 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Mégevand, P., Mercier, M. R., Groppe, D. M., Golumbic, E. Z., Mesgarani, N., Beauchamp, M. S., Schroeder, C. E., & Mehta, A. D. (2020). Crossmodal Phase Reset and Evoked Responses Provide Complementary Mechanisms for the Influence of Visual Speech in Auditory Cortex. *Journal of Neuroscience*, *40*(44), 8530–8542. https://doi.org/10.1523/JNEUROSCI.0555-20.2020

Metzger, B. A., Magnotti, J. F., Wang, Z., Nesbitt, E., Karas, P. J., Yoshor, D., & Beauchamp, M. S. (2020). Responses to Visual Speech in Human Posterior Superior Temporal Gyrus

Examined with iEEG Deconvolution. *Journal of Neuroscience*, *40*(36), 6938–6948. https://doi.org/10.1523/JNEUROSCI.0279-20.2020

Molinaro, N., & Lizarazu, M. (2018). Delta(but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience*, *48*(7), 2642–2650. https://doi.org/10.1111/ejn.13811

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., & Malach, R. (2005). Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. *Science (New York, N.Y.)*, *309*(5736), 951–954. https://doi.org/10.1126/science.1110913

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science*, *15*(2), 133–137. https://doi.org/10.1111/j.0963-7214.2004.01502010.x

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. *Interspeech 2017*, 2616–2620. https://doi.org/10.21437/Interspeech.2017-950

Nidiffer, A. R., Cao, C. Z., O'Sullivan, A., & Lalor, E. C. (2023). A representation of abstract linguistic categories in the visual system underlies successful lipreading. *NeuroImage*, *282*, 120391. https://doi.org/10.1016/j.neuroimage.2023.120391

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., Howard, M. A., & Brugge, J. F. (2009). Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex. *Journal of Neuroscience*, *29*(49), 15564–15574. https://doi.org/10.1523/JNEUROSCI.3065-09.2009

Obleser, J., & Kayser, C. (2019). Neural Entrainment and Attentional Selection in the Listening Brain. *Trends in Cognitive Sciences*, *23*(11), 913–926. https://doi.org/10.1016/j.tics.2019.08.004

Oganian, Y., Kojima, K., Breska, A., Cai, C., Findlay, A., Chang, E., & Nagarajan, S. S. (2023). Phase Alignment of Low-Frequency Neural Activity to the Amplitude Envelope of Speech Reflects Evoked Responses to Acoustic Edges, Not Oscillatory Entrainment. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *43*(21), 3909–3921. https://doi.org/10.1523/JNEUROSCI.1663-22.2023

O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2017). Visual Cortical Entrainment to Motion and Categorical Speech Features during Silent Lipreading. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00679

O'Sullivan, A. E., Crosse, M. J., Liberto, G. M. D., Cheveigné, A. de, & Lalor, E. C. (2021). Neurophysiological Indices of Audiovisual Speech Processing Reveal a Hierarchy of Multisensory Integration Effects. *Journal of Neuroscience*, *41*(23), 4991–5003. https://doi.org/10.1523/JNEUROSCI.0906-20.2021

O'Sullivan, A. E., Lim, C. Y., & Lalor, E. C. (2019). Look at me when I'm talking to you: Selective attention at a multisensory cocktail party can be decoded using stimulus reconstruction and alpha power modulations. *European Journal of Neuroscience*, *50*(8), 3282–3295. https://doi.org/10.1111/ejn.14425

Owens, E., & Blazek, B. (1985). Visemes Observed by Hearing-Impaired and Normal-Hearing Adult Viewers. *Journal of Speech, Language, and Hearing Research*, *28*(3), 381–393. https://doi.org/10.1044/jshr.2803.381

Ozker, M., Schepers, I. M., Magnotti, J. F., Yoshor, D., & 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. https://doi.org/10.1162/jocn_a_01110

Ozker, M., Yoshor, D., & 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*. https://doi.org/10.3389/fnhum.2018.00141

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. *Cerebral Cortex*, *23*(6), 1378–1387. https://doi.org/10.1093/cercor/bhs118

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., & Sams, M. (2005). Primary auditory cortex activation by visual speech: An fMRI study at 3 T. *Neuroreport*, *16*(2), 125–128. https://doi.org/10.1097/00001756-200502080-00010

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018). End-to-End Audiovisual Speech Recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6548–6552. https://doi.org/10.1109/ICASSP.2018.8461326

Plass, J., Brang, D., Suzuki, S., & Grabowecky, M. (2020). Vision perceptually restores auditory spectral dynamics in speech. *Proceedings of the National Academy of Sciences*, *117*(29), 16920–16927. https://doi.org/10.1073/pnas.2002887117

Ray, S., & Maunsell, J. H. R. (2011). Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. *PLOS Biology*, *9*(4), e1000610. https://doi.org/10.1371/journal.pbio.1000610

Rennig, J., & Beauchamp, M. S. (2022). Intelligibility of audiovisual sentences drives multivoxel response patterns in human superior temporal cortex. *NeuroImage*, *247*, 118796. https://doi.org/10.1016/j.neuroimage.2021.118796

Richie, Carolyn, Warburton, Sarah, & Carter, Megan. (2009). *Audiovisual Database of Spoken American English* (p. 7759462 KB) [Dataset]. Linguistic Data Consortium. https://doi.org/10.35111/XJ23-6G13

Rivet, B., Girin, L., & Jutten, C. (2007). Mixing Audiovisual Speech Processing and Blind Source Separation for the Extraction of Speech Signals From Convolutive Mixtures. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(1), 96–108. IEEE Transactions on Audio, Speech, and Language Processing. https://doi.org/10.1109/TASL.2006.872619

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(2), 318–331. https://doi.org/10.1037/0096-1523.22.2.318

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments. *Cerebral Cortex*, *17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Russo, M. J., Prodan, V., Meda, N. N., Carcavallo, L., Muracioli, A., Sabe, L., Bonamico, L., Allegri, R. F., & Olmos, L. (2017). High-technology augmentative communication for adults with post-stroke aphasia: A systematic review. *Expert Review of Medical Devices*, *14*(5), 355–370. https://doi.org/10.1080/17434440.2017.1324291

Sauseng, P., Klimesch, W., Gruber, W. R., & Birbaumer, N. (2008). Cross-frequency phase synchronization: A brain mechanism of memory matching and attention. *NeuroImage*, *40*(1), 308–317. https://doi.org/10.1016/j.neuroimage.2007.11.032

Shahin, A. J., Backer, K. C., Rosenblum, L. D., & Kerlin, J. R. (2018). Neural Mechanisms Underlying Cross-Modal Phonetic Encoding. *Journal of Neuroscience*, *38*(7), 1835–1849. https://doi.org/10.1523/JNEUROSCI.1566-17.2017

Shahin, A. J., Kerlin, J. R., Bhat, J., & Miller, L. M. (2012). Neural restoration of degraded audiovisual speech. *NeuroImage*, *60*(1), 530–538. https://doi.org/10.1016/j.neuroimage.2011.11.097

Shukla, A., Vougioukas, K., Ma, P., Petridis, S., & Pantic, M. (2020). Visually Guided Self Supervised Learning of Speech Representations. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6299–6303. https://doi.org/10.1109/ICASSP40776.2020.9053415

Son Chung, J., Senior, A., Vinyals, O., & Zisserman, A. (2017). *Lip Reading Sentences in the Wild*. 6447–6456. https://openaccess.thecvf.com/content_cvpr_2017/html/Chung_Lip_Reading_Sentences_CVPR_2017_paper.html

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A., & Williams Jr., R. M. (1949). *The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1* (pp. xii, 599). Princeton Univ. Press.

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

Thézé, R., Giraud, A.-L., & Mégevand, P. (2020). The phase of cortical oscillations determines the perceptual fate of visual cues in naturalistic audiovisual speech. *Science Advances*, *6*(45), eabc6348. https://doi.org/10.1126/sciadv.abc6348

Thomas, S. M., & Jordan, T. R. (2004). Contributions of Oral and Extraoral Facial Movement to Visual and Audiovisual Speech Perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(5), 873–888. https://doi.org/10.1037/0096-1523.30.5.873

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

van Wassenhove, V. (2013). Speech through ears and eyes: Interfacing the senses with the supramodal brain. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00388

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of Training on the Visual Recognition of Consonants. *Journal of Speech and Hearing Research*, *20*(1), 130–145. https://doi.org/10.1044/jshr.2001.130

Wang, W., Viswanathan, S., Lee, T., & Grafton, S. T. (2016). Coupling between Theta Oscillations and Cognitive Control Network during Cross-Modal Visual and Auditory Attention: Supramodal vs Modality-Specific Mechanisms. *PLOS ONE*, *11*(7), e0158465. https://doi.org/10.1371/journal.pone.0158465

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic Description of Factorial Models for Analysis of Variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *22*(3), 392–399. https://doi.org/10.2307/2346786

Woodward, M. F., & Barber, C. G. (1960). Phoneme Perception in Lipreading. *Journal of Speech and Hearing Research*, *3*(3), 212–222. https://doi.org/10.1044/jshr.0303.212

Yan, D., & Seki, A. (n.d.). Differential modulations of theta and beta oscillations by audiovisual congruency in letter-speech sound integration. *European Journal of Neuroscience*, *n/a*(n/a). https://doi.org/10.1111/ejn.16563

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*(1), 23–43. https://doi.org/10.1016/S0167-6393(98)00048-X

Zhang, L., & Du, Y. (2022). Lip movements enhance speech representations and effective connectivity in auditory dorsal stream. *NeuroImage*, *257*, 119311. https://doi.org/10.1016/j.neuroimage.2022.119311

Zhang, Y., Rennig, J., Magnotti, J. F., & Beauchamp, M. S. (2023). Multivariate fMRI responses in superior temporal cortex predict visual contributions to, and individual differences in, the intelligibility of noisy speech. *NeuroImage*, *278*, 120271. https://doi.org/10.1016/j.neuroimage.2023.120271

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase Entrainment of Brain Oscillations Causally Modulates Neural Responses to Intelligible Speech. *Current Biology*, *28*(3), 401-408.e5. https://doi.org/10.1016/j.cub.2017.11.071