Auditory cortex encodes lipreading information through spatially distributed activity

Ganesan Karthik[1], Cody Zhewei Cao[1], Michael I. Demidenko[1], Andrew Jahn[1], William C. Stacey[2], Vibhangini S. Wasade[3,4], David Brang*[1]

[1] Department of Psychology, University of Michigan, Ann Arbor, MI 48109
[2] Department of Neurology, University of Michigan, Ann Arbor, MI 48109
[3] Henry Ford Hospital, Detroit, MI 48202
[4] Department of Neurology, Wayne State University School of Medicine, MI 48201

***Corresponding Author:** djbrang@umich.edu

Data Availability Statement

The data that support the findings of this study will be made openly available through the University of Michigan Deep Blue Repository.

Conflict of Interest Statement

The authors declare no competing financial interests.

Acknowledgements

This study was supported by NIH Grants R00DC013828 and R01NS094399. The authors report no conflicts of interest.

Keywords

Multisensory; Audiovisual; Speech; ECoG; iEEG; sEEG

Summary

Face-to-face communication improves the quality and accuracy of heard speech, particularly in noisy environments. Silent lipreading modulates activity in auditory regions, which has been hypothesized to reflect the transformation and encoding of multiple forms of visual speech information used to support hearing processes. Evidence suggests visual timing information as one such signal encoded in auditory areas: seeing when a speaker's lips come together between words can help listeners parse word-level boundaries. However, it remains unclear how lipreading alters activity in the auditory system to improve speech perception at the single word-level. Using fMRI and intracranial electrodes in patients, here we show that silently lipread words can be classified from neural activity in auditory areas based on distributed spatial information. Lipread words evoked similar representations to the corresponding heard words, consistent with the prediction that automatic lipreading refines the tuning of auditory representations. Similar to heard words, lipread words varied in the distinctiveness of their neural representations in auditory cortex: e.g., the lipread words DIG and GIG evoked more similar neural activity in auditory cortex relative to the more perceptually distinct word FIG, suggesting that lipreading activity reflects probabilistic distributions as opposed to the unique identity of the lipread word. Notably, while visual speech has both excitatory and suppressive effects on auditory firing rates, classification was observed in both neural populations, consistent with the prediction that lipreading contributes to phoneme population tuning by both activating the corresponding representation and suppressing incorrect phonemic representations. These results support a model in which the auditory system combines the joint neural distributions evoked by heard and lipread words to generate a more precise estimate of what was said, particularly during noisy speech.

Introduction

Visual speech improves auditory speech perception during face-to-face conversations[1,2]. These benefits are strongest in noisy situations[3] and in individuals with hearing loss due to healthy aging[4], intrinsic brain tumor[5], stroke[6,7], concussion[8,9], or cochlear implants[10]. However, there remains limited understanding of how the brain enables vision to facilitate hearing processes.

The ability to extract useful information from visual speech signals (e.g., lipreading) is an implicit behavior that is rooted in the statistical relationship between auditory and visual cues in the natural environment[11]. Lip dynamics are strongly correlated with different features of speech including temporal information (onset of words, rate of speech, and the boundaries between words) and relative spectral pitch based on the acoustics of the oral cavity[1]. Most recognizably, the shape of the lips during speech is reliably associated with corresponding speech sounds[12]; these simple lip shapes are described as visemes and are analogous to phonemes in the auditory domain (basic units of speech sounds).

Research has demonstrated that silent visual speech (e.g., lipreading) evokes activity within the auditory system[13,14]. Indeed, intracranial electroencephalography (iEEG) recordings indicate that visual speech influences processing in auditory regions through multiple temporal, spectral, and spatial configurations[15]. While these findings highlight the broad effect of visual information on auditory speech processing, differences in activity do not provide a mechanistic account for how visual speech signals integrate with auditory neuronal populations. Best understood among these mechanisms is how visual timing information during continuous speech biases auditory timing through phase-resetting mechanisms[16,17]. However, it remains unclear how lipreading information (visemes) is transformed into a signal used by the auditory system.

Within the auditory domain, phonetic and phonemic features are encoded by local and distributed populations of neurons, respectively[18]. Mesgarani and colleagues[18] used human iEEG recorded from high-density electrodes to demonstrate that phonemes are represented by distributed populations of neurons in the STG. Combined with past research, these data support a model in which the STG contains a patchy distribution of neurons that are tuned to specific phonetic features via their spectro-temporal profiles[18,19]. For example, research has reported spatially distinct responses in these regions to spectrally similar phonemes such as /ba/ and /da/[19-21], and clustered activities across a large phoneme-space (e.g., the distributed pattern of activity to /ma/ is more similar to /na/ than it is to the spectro-temporally distinct phoneme /ba/[18]. Indeed, the identities of different heard phonemes can be decoded by the distribution of activity in the auditory cortex[22], even when the physical auditory stimulus remains the same.

Building on this understanding of auditory perception, we proposed that activity from lipread visemes is relayed from visual regions to auditory cortex, preferentially modulating the same populations of neurons that encode matching phoneme responses[15]. In this hypothesis, heard and lipread activations in auditory cortex are combined through a winner-take-all mechanism, in which the phoneme population with highest activation profile leads to the phoneme that is perceived[23].

Here we test the hypothesis that the identities of individual visemes are represented in the auditory system through distributed patterns of activation, and these spatial distributions match corresponding phoneme representations. Auditory cortex activation magnitude and informational content were examined using functional magnetic resonance imaging (fMRI) in healthy individuals and iEEG recordings in patients with epilepsy during word perception tasks, in which patients either saw the lip movements or heard the speech sounds for the same groups of words.

The identities of the different words were classified from fMRI and iEEG signals in auditory cortex using support vector machines (SVMs). Results demonstrate that the auditory system reliably encodes the identity of visemes using spatially distributed activity in a similar manner to heard words. Moreover, visemes evoked spatially similar activity to matching phonemes, consistent with the hypothesis that visual speech targets corresponding phoneme representations

Results

fMRI Experiment

In Experiment 1 we presented subjects (n = 64) with consonant vowel (CV) syllables while fMRI activity was acquired. Each trial included the silent video or auditory stimulus taken from a speaker producing the CVs /mama/, /fafa/, or /kaka/ (Fig. 1a). Stimuli were presented using an optimized event-related design (pre-registered at OSF: https://osf.io/6fzwd/) and data were analyzed using univariate and decoding approaches to examine the activation and information present in heard and lipread signals.

Mean behavioral accuracy was high in both conditions: 95.67% (SD = 3.01%) in the listening condition and 92.31% (SD = 3.72%) in the lipreading condition. As expected, the mean accuracy in the listening condition was significantly higher than the lipreading condition; $t(63) = 6.57$, $p < 0.001$, Cohen's $d = 0.96$. None of the 64 subjects performed below the pre-registered exclusion threshold (accuracy in either condition below 75%).

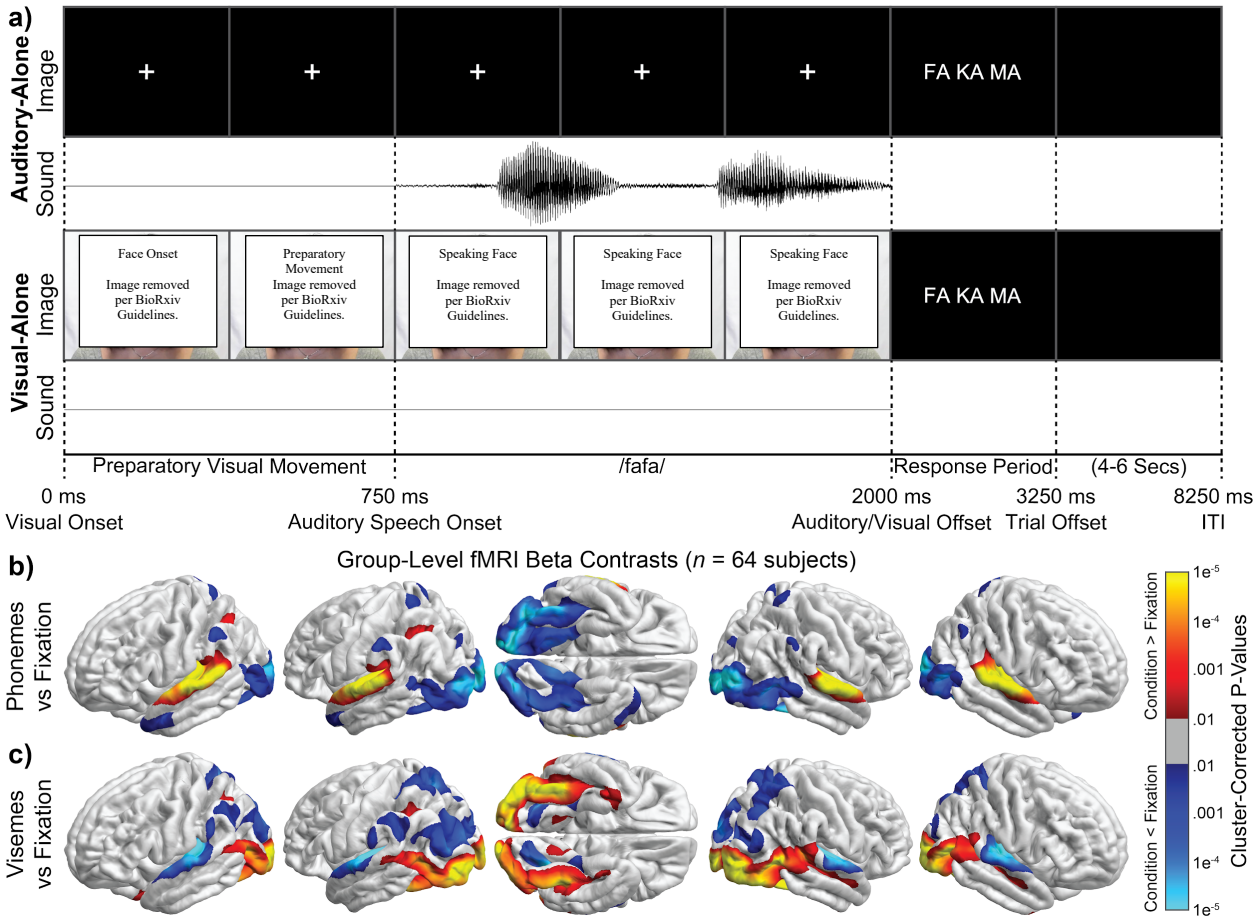

Figure 1. fMRI task schematic and univariate activations. (a) Schematic of auditory and visual trials. Auditory trials began with a fixation cross followed by a CV stimulus (either /mama/, /fafa/, or /kaka/). Visual trials presented the visual components of these same recordings without the corresponding audio track. After stimulus offset, subjects were cued to identify which of the three phonemes (or visemes) they saw (or heard) via button press. (b-c) Univariate group-level analyses of (b) phonemes vs fixation or (c) visemes vs fixation. Phonemes evoked maximally increased activity in the STG bilaterally. Visemes evoked increased activity within bilateral visual cortex, left pSTS, right MT, and posterior STG bilaterally, along with suppression through middle STG regions (including Heschl's gyrus). Colored regions reflect significant increases (red and yellow) or decreases (blues) in task-related activation (thresholded at p<.001 and corrected for multiple comparisons using cluster-statistics).

Previous research demonstrated that silent lipreading modulates fMRI BOLD responses within auditory regions[13]. First, we replicated this finding using univariate contrasts in listening and lipreading conditions (phonemes vs fixation and visemes vs fixation, respectively). Results of the whole-brain analysis, corrected for multiple comparisons using cluster statistics (vertex-wise threshold of $P < 0.001$, cluster-corrected to $P < 0.05$) are shown in Fig. 1b-c; full statistics for each analysis is reported in Supp. Tables 1 and 2 and beta estimates extracted from auditory and visual regions of interest (ROIs) are shown in Supp. Fig. 1. Phonemes elicited significantly increased

BOLD activity within the STG bilaterally and decreased BOLD within visual regions. Visemes similarly modulated activity broadly through the STG, with increased BOLD at the posterior STG and decreased BOLD in the middle to anterior STG, along with increased BOLD in visual regions. The finding that lipreading suppresses neural activity within portions of the auditory system is consistent with prior reports from fMRI[24] and iEEG[25], which has been theorized to reflect the optimized tuning of neurons specialized for auditory speech[26].

Univariate contrasts reveal activation magnitude but not informational content or representational structure. To examine whether visual speech is represented in the auditory system by distributed patterns of activity, we used multivariate pattern analysis (MVPA)[27] to classify individual phoneme and viseme labels. Previous decoding-based approaches using fMRI[28,29] and iEEG[30,31] demonstrated that speech patterns could be reconstructed from spatially distributed activity in auditory cortex.

Whole-brain searchlight-based MVPA was applied at the individual-subject level [32] conducted separately for each of the two conditions of interest (phonemes and visemes). Results of the whole-brain analysis, corrected for multiple comparisons using cluster statistics are shown in Figure 2 (vertex-wise threshold of $P < 0.001$, cluster-corrected to $P < 0.05$); full statistics for each analysis is reported in Supp. Tables 3 and 4. In the auditory-only condition, peak decoding accuracy was observed bilaterally in the STG and pSTS. This is consistent with previous studies demonstrating phonetic representations in the STG using MVPA[33]. In the visual-only condition, peak decoding accuracy was observed within the STG bilaterally, the left pSTS, visual cortex bilaterally, and right hMT+.

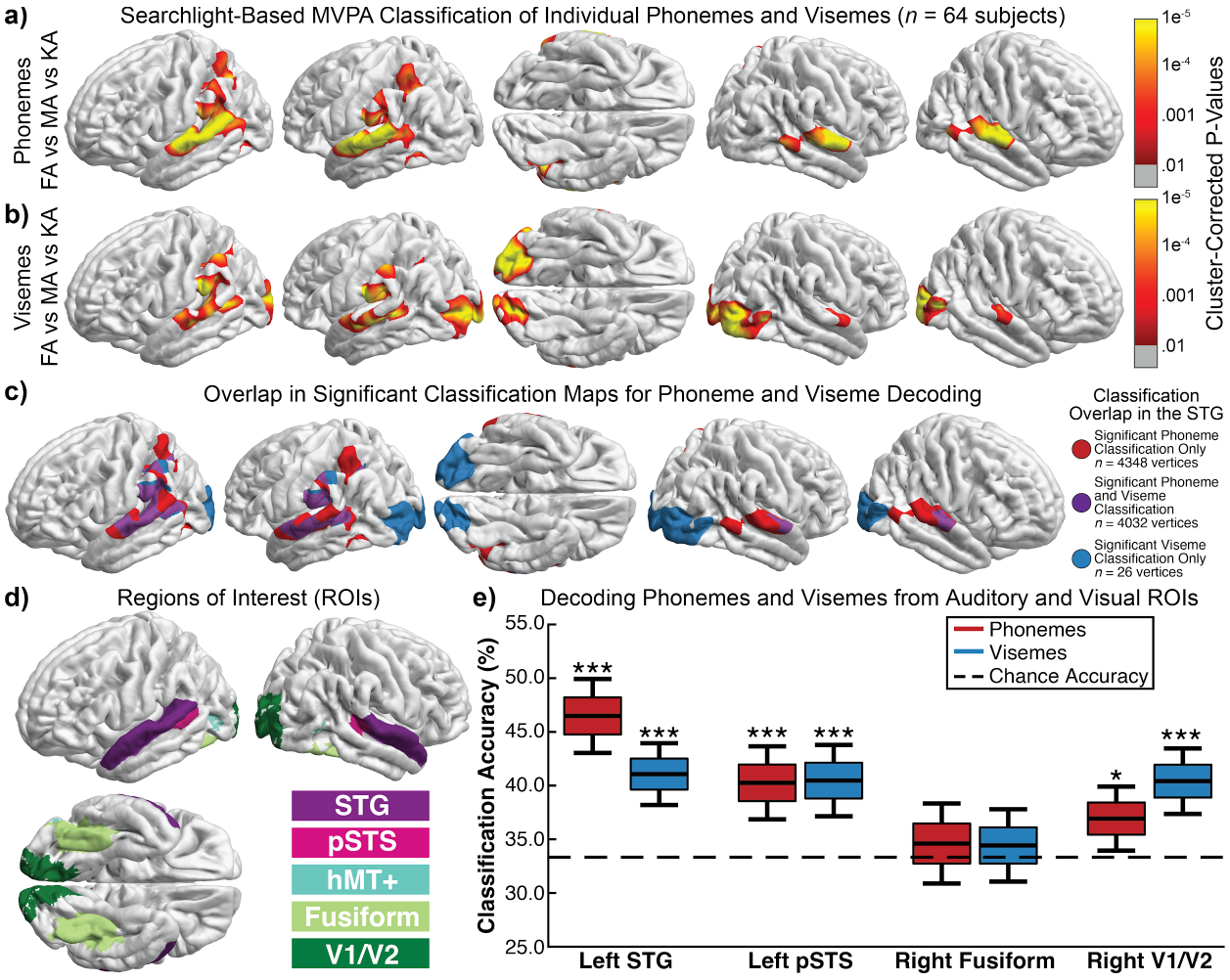

Figure 2. fMRI decoding of phoneme and viseme information in an event-related design. (a-c) Searchlight-based MVPA classification in $n = 64$ subjects. Classifiers were trained to identify (a) the phoneme heard (/fafa/, /mama/, or /kaka/) in the auditory-only condition, (b) the viseme seen in the visual-only condition, or (c) condition differences between auditory-only and visual alone trials. Decoding was conducted at the individual subject level and only group-level differences greater than chance (thresholded at p<.001 and corrected for multiple comparisons using cluster-statistics) are shown. (a) Peak phoneme decoding was observed in the bilateral STG. (b) Significant viseme decoding was observed in the bilateral STG, left pSTS, and visual regions. (c) Vertices with significant classification of phonemes but not visemes (red), visemes but not phonemes (blue), or with significant classification of both phonemes and visemes (purple). There is a large overlap in the vertices at which visemes and phonemes could be classified. Restricted to the just the STG, vertices at which viseme classification was significant covered roughly half of the area in the STG that phonemes were classified successfully at (48.1% overlap) with negligible area uniquely able to classify visemes. (d) Regions of interest (ROIs) used for hypothesis driven classification at the single-subject level. (e) Results of classification at selected ROIs. Phonemes were significantly classified from the left STG and pSTS. Visemes were significantly classified from the left STG (consistent with the hypothesis that information about visemes is represented within the STG), pSTS, and visual cortex. Center line reflects the mean, colored box SE, and the tails 95% confidence intervals. *p<.05, ***p<.001. Chance accuracy is 33.3%.

To understand the spatial overlap of phoneme and viseme representations in the auditory system we compared the spatial distribution of the classification maps. Results showed that a majority of vertices contained either only phoneme information or both phoneme and viseme information, with very few vertices representing viseme information alone. Across the left and

right STG, phonemes (but not visemes) were significantly classified at 27.2% of vertices, visemes (but not phonemes) were significantly classified at 0.16% of vertices, and phonemes and visemes were jointly classified from 25.2% of vertices. In total, phonemes were classified at twice as many vertices compared to visemes within the STG (52.4% vs 25.4%). Thus, STG classification was most prominent in vertices where lipreading produced BOLD suppression effects, consistent with predictions that lipreading regionally suppresses auditory activity to improve phoneme tuning responses.

To further quantify the relative information across target regions, we performed individual-subject SVM classification in five regions of interest (ROIs) in each hemisphere (dimension of vertices within the ROI; Fig. 2d). As shown in Fig 2e and Supp. Fig. 2 Phoneme classification accuracy was strongly above chance (33.3%) in the left and right STG, and the left pSTS (all $p <$ .001) with more moderate classification observed in the right pSTS, left and right hMT+, and left and right V1/V2 (all $p < .05$). Viseme classification accuracy was strongly above chance in the left STG, left pSTS, and right V1/V2 (all $p < .001$) with more moderate classification observed in the right STG, right hMT+, and left V1/V2 (all $p < .05$).

The univariate analysis showed that visual speech modulated activity in auditory regions: visemes suppressed activity in the middle STG and increased activity in the posterior STG. Viseme decoding analyses identified above-chance classification accuracy broadly throughout the STG. Comparing the two results, the univariate visual-only analysis showed four times as many significant vertices in the STG (bilaterally) compared to the area with significant viseme classification in the MVPA analysis (64.3% vs 15.7% of STG vertices). This is consistent with the prediction that only a restricted proportion of the STG encodes visemic information, while other regions reflect domain general responses to the visual signals or the presentation of other visual

information (e.g., temporal or spectral information; [1]. To better understand the relationship between these results we compared areas of significant classification relative to areas of significant activity (either increased or decreased BOLD for visemes relative to fixation). Across the left and right STG, significant viseme classification was observed in 27.3% of vertices with decreased BOLD during the visual-only condition (relative to fixation). Conversely, significant viseme classification was observed in only 8.74% of vertices with increased BOLD during the visual-only condition. This is consistent with a model in which visemes activate the correct representation in a minority of vertices in the posterior STG and suppress incorrect representations throughout the STG broadly. In contrast, viseme classification within the left pSTS was present only within vertices that showed increased BOLD activation during lipreading (87.7% of vertices with a BOLD increase).

While the decoding analyses provide information about which regions of the brain encode the identities of individual phonemes and visemes, it is not possible to directly investigate similarities between how these phonemes and visemes are represented in these regions. For example, an examination of the spatial and temporal (dis)similarities for phonemes and visemes would aid in the interpretation of how visemic identities are transformed and encoded in the auditory regions. Using the same data as in the classification analysis, we separately averaged beta estimates for each phoneme and viseme and then compared the spatial distribution of activations at the individual subject level. We restricted vertices to those with significant classification in both auditory-only and visual-only conditions (purple vertices in Fig. 2c) within the STG and calculated the correlation between vertex-wise beta estimates for phoneme and viseme pairs. We averaged correlations across matching pairs (e.g., the phoneme /ma/ and the viseme /ma/) and separately mismatching pairs (e.g., the phoneme /ma/ and the visemes /ka/ and /fa/), then compared

correlations at the group level. Across subjects we observed a small but reliable increase in the correlation between associated phonemes and visemes compared to mismatched phonemes and visemes in the left STG, $t(63) = 2.190$, $p = .032$, $d = 0.274$, but not the right STG, $t(63) = -0.026$, $p = .979$, $d = -0.003$. Repeating this analysis on all vertices within the anatomically defined STG ROI at the individual subject level, we observed a similar result, $t(63) = 2.493$, $p = .015$, $d = 0.312$. This result demonstrates that visemes evoke similar patterns of activity within the STG to those of phonemes. This is consistent with the prediction that automatic lipreading refines the tuning of auditory representations.

iEEG Experiment

Results from the fMRI study demonstrated that viseme information is represented in auditory areas. Moreover, because visemes were classified based on the spatial distribution of vertices in the STG, this supports a model in which lipreading is represented through population-coded responses in the auditory system, similar to the neural representation underlying phonemes [18]. However, the slow temporal dynamics of fMRI signals prevent a fine-grained analysis of the time-course of lipreading activation to examine when this information is available to the auditory system. Additionally, the use of only three dissimilar CV stimuli prevented a more graded analysis of these population-coded responses, such as whether more perceptually similar phonemes (e.g., /ga/ and /da/) elicit more similar population-coded responses relative to perceptually distinct phonemes (e.g., /fa/ and /ba/). To answer both of these questions, next we collected data from a similar auditory-visual speech paradigm from $n = 6$ patients with epilepsy who had electrodes implanted within auditory areas of the brain (Fig. 3a). Patients were presented with 240 auditory-only (listening) and 240 visual-only (lipreading) trials containing a single 1-2 syllable word. Each

word began with one of four consonants ('B', 'F', 'G', or 'D') to enable the decoding of distinct phonemic patterns. 40 distinct words were used (10 containing each of the 4 initial consonants; Supp. Table 5) and each word was repeated 6 times within each condition. On each trial subjects selected the initial consonant heard or seen from four options (4-alternative forced choice). Subjects' mean behavioral accuracy across listening and lipreading trials was significantly above chance (25%) at the group level: listening (M = .919, SD = .081, $t(5) = 20.2$, $p < .001$, $d = 8.23$), lipreading (M = .674, SD = .179, $t(5) = 5.78$, $p = .002$, $d = 2.36$). As expected, listening trials were correctly identified significantly more often than visual trials, $t(5) = 5.93$, $p = .002$, $d = 2.42$.

Words in both the auditory-only and visual-only conditions evoked activity broadly throughout the STG and MTG consistent with prior work[15]. Examining the spectral breakdown of these responses (Fig. 3b-d), phonemes evoked increased theta and high gamma power (HGp) and suppressed beta power following word onset. In natural speech, visual articulations typically occur before the onset of speech-related sounds (typically within 40 - 200 ms of speech onset[34]). Because of this pre-articulatory visual information, visemes evoked increased beta suppression beginning before the expected phoneme onset time, consistent with past research[15] and indicative of feedback inputs into the auditory system[35,36]. Additionally, visemes evoked more moderate changes in HGp following sound onset, even though no sound was present. Viseme-related HGp increases were maximal at the posterior STG, consistent with past research[15,37,38]. Fig. 3e shows this pattern in single electrode HGp responses from two patients (both electrodes within the left posterior STG), with HGp changes occurring in response to visemes before phonemes. This pattern was distinct from responses in the fusiform gyrus, at which visemes evoked early HGp increases following the onset of the visual stimulus and no reliable response at any point during auditory-only trials (Fig. 3f).

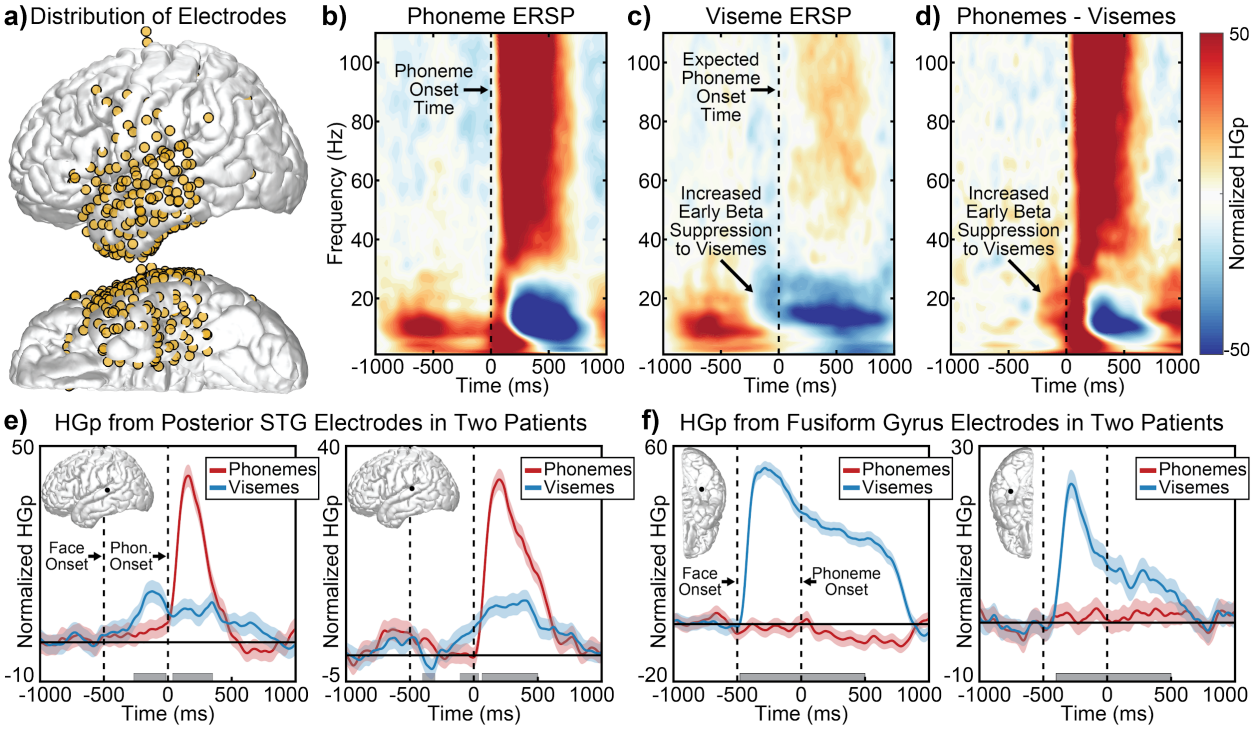

Figure 3. iEEG results during an auditory-only (listening) and visual-only (lipreading) speech perception paradigm. (a) Distribution of all recorded electrodes (those beneath the pial surface not shown) (n = 6 patients). (b-d) Event-related spectral perturbations (ERSP) plots from all STG electrodes, averaged across subjects. (b) Phoneme responses peaked after sound onset with theta and high-gamma power (HGp) increases, as well as beta suppression. (c) Viseme responses evoked maximal changes in beta power, with increased beta suppression starting before the expected time of sound onset. (d) Difference between phoneme and viseme ERSP plots. (e) HGp responses from two superior temporal gyrus (STG) electrodes in response to auditory-only trials (phonemes; red lines) and visual-only trials (visemes; blue lines). Posterior STG electrodes showed increased HGp responses to visemes before the time when speech sounds would be expected to begin. (f) HGp responses from two fusiform gyrus electrodes. Visemes evoked increased HGp shortly after onset of the face, with elevated HGp persisted throughout the visual movement period. Phonemes failed to evoke reliable changes in activity within the fusiform gyrus. Shaded regions reflect single condition 95% confidence intervals. Light gray boxes show significant between condition differences (multiple comparisons corrected using FDR).

To examine whether viseme information is represented within auditory regions, we decoded word information using spatial and temporal signals from iEEG electrodes. Fig. 4a shows group-level classification accuracy for decoding the initial word consonant for auditory-only and visual-only trials. Classification was conducted separately in each subject using SVM classifiers on single-trial event-related potential (ERP) responses (60 per consonant-initial auditory and visual words) using time points and electrodes as dimensions. We observed significant classification (evaluated using binomial statistics) in all six patients for both auditory-only (all $p < .05$) and visual-only conditions (all $p < .05$) (single-subject statistics shown in Supplemental Table 6). Similarly, at the group-level we observed classification accuracy reliably above chance for both

auditory-only ($t(5) = 5.39$, $p = .0030$, $d = 2.20$), and visual-only trials ($t(5) = 5.57$, $p = .0026$, $d = 2.27$). We additionally observed a trend towards greater classification in auditory-only trials relative to visual-only trials ($t(5) = 2.08$, $p = .0916$, $d = 0.851$).

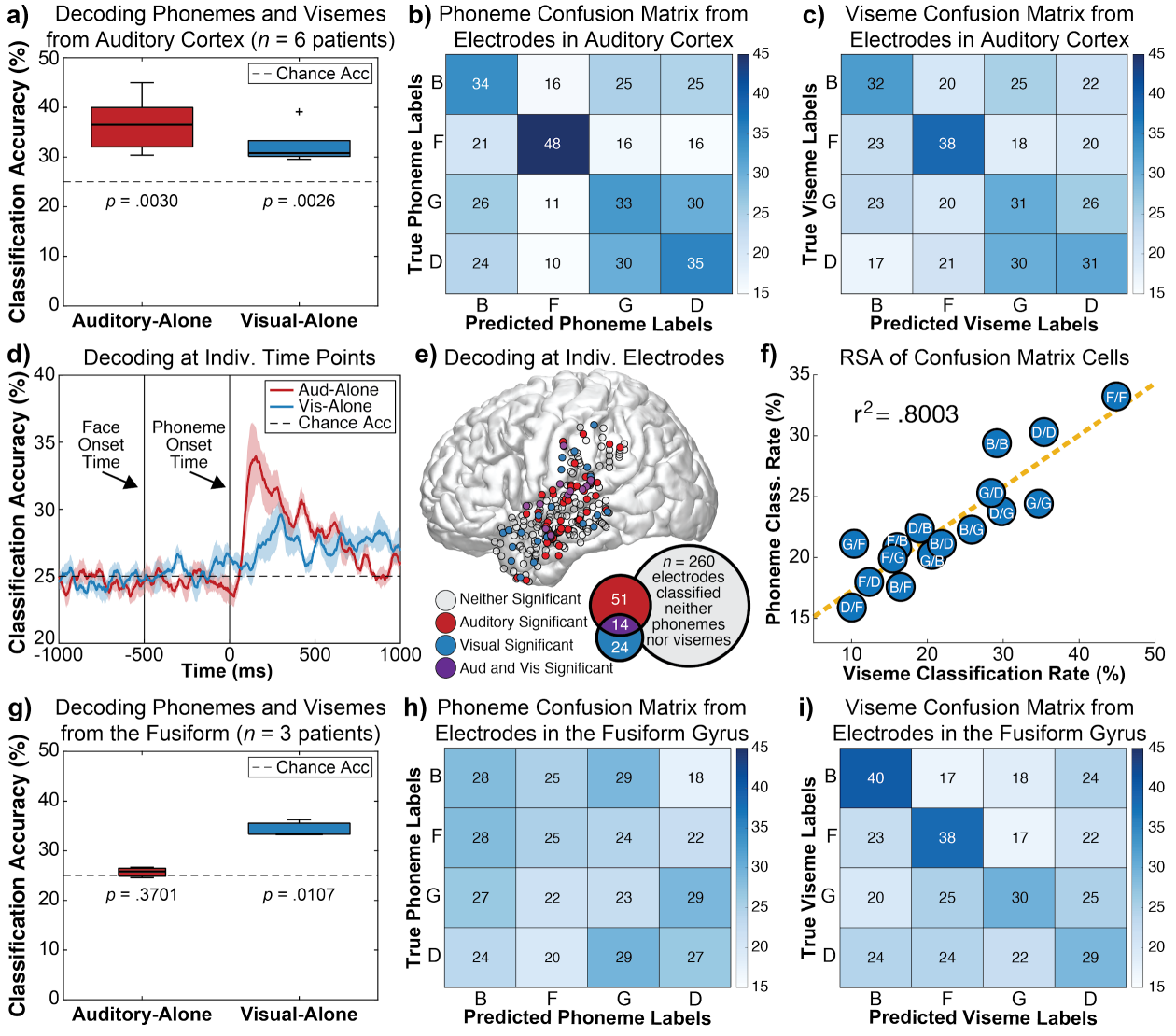

Figure 4. iEEG classification of phoneme and viseme identities from auditory (n = 6 patients) and visual (n = 3 patients) regions. (a) Accuracy of an SVM classifier in identifying the correct initial consonant ('B', 'F', 'G', or 'D') from either auditory-only or visual-only words classified at the individual-subject level from spatial and temporal iEEG information. Both visemes and phonemes were reliably classified from auditory electrodes. Chance accuracy is 25% and plots show group-level boxplots. (b-c) Group-averaged confusion matrices taken from 4-class auditory-only and visual-only SVM classifiers. Cells denote the frequency at which each consonant-initial word was predicted (x-axis) relative to the true labels (y-axis). 'F' initial words were best classified across both auditory-only and visual-only conditions, whereas 'G' and 'D' initial words were more readily confused. (d) Group-averaged classification at individual time-points from auditory electrodes (phoneme-onset at 0 sec) showing significant classification accuracy for both auditory-only and visual-only trials shortly after phoneme onset; in the visual-only condition, this time-point reflected the associated speech onset time even though no auditory stimulus was presented. Shaded region reflects SEM. (e) Spatial distribution of electrodes at which auditory-only (red) or visual-only (blue) trials were reliably classified ($p<.05$ based on binomial statistics); purple electrodes reflect significant classification in both conditions (with 11 out of 14 of these electrodes present in the superior temporal gyrus) and gray electrodes reflect non-significant classification in either condition. Electrodes

beneath the pial surface were projected out to the lateral surface for visualization. (f) Scatterplot quantifying the similarity of classification frequency for auditory-only trials and visual-only trials from auditory electrodes (taken from 8-class classifier). Data reflect pairwise classification values, with the first letter reflecting the real consonant label and the second letter the predicted consonant label. For example, 'F' trials were predicted correctly at high frequency for both auditory and visual trials, whereas 'D' trials were incorrectly labeled as 'F' trials infrequently across both auditory and visual trials. The high correlation ($r^2 = .8003$, $p<.001$ permutation test) is consistent with the hypothesis that visual speech evokes responses targeting similar distributions of neurons to corresponding phoneme responses in the STG. (g) Group-level classification accuracy showing that responses in the fusiform gyrus can distinguish between different visemes but not phonemes. (h-i) Group-level confusion matrices for auditory-only and visual-only trials from fusiform gyrus electrodes.

The successful classification of phonemes and visemes indicated that auditory areas represent information about the consonant initial words for both auditory-only and visual-only speech stimuli. The diagonal of the confusion matrices (Fig. 4b-c) shows that this classification was robust for each of the four auditory-only and visual-only stimuli considered ($p < .05$ for 3 out of 4 phonemes and 3 out of 4 visemes) (statistics shown in Supp. Table 7). Previous research has shown that local auditory responses spatially cluster according to phonetic features[18]; for the stimuli used here, B, G, and D form one cluster and F another. Consistent with these clusters, classification was higher for auditory words with the consonant F compared to words with the consonants B, G, or D ($t(5) = 2.95$, $p = .0319$, $d = 1.20$); a similar trend was observed for visual-alone trials ($t(5) = 1.80$, $p = .1318$, $d = 0.736$), consistent with perceptual ambiguity of these items in phoneme-space.

To examine the time-course of auditory and visual speech representations within the auditory system, we classified the identity of stimuli independently at 10 ms intervals (from -1000 ms to +1000 ms after phoneme onset time). Classification was applied separately for each subject and group-level statistics were calculated across subjects (Fig. 4d). Results showed significant classification accuracy ($p < .05$) for both auditory-only and visual-only trials shortly after phoneme onset, indicating that visemic information is available to the auditory system at the same perceptual stage as is phonemic information.

In a parallel set of analyses, we classified the identity of stimuli independently at each electrode within an auditory region (including the STG, MTG, SMG) to understand the spatial

distribution of phoneme and viseme classification and their overlap. Phonemes were significantly ($p < .05$ using binomial statistics) classified from 65 out of 260 electrodes (25.0%) while visemes were significantly classified from only 38 electrodes (14.6%). This pattern is similar to the overall classification rate observed in our fMRI data, such that phonemes were classified at twice as many vertices compared to visemes within the STG. Restricted to only the STG, 14 electrodes significantly classified visemic information in total, with 11 of these also significantly classified phonemic information, highlighting the spatial overlap of these processes. Again, this is consistent with the pattern observed in our fMRI data, in which few vertices were sensitive to only visemic information.

Because phonemes are represented through population coded responses, misclassification can reveal information about related neural processes. For example, if the rate at which the consonant /d/ is misclassified as /g/ in both auditory-only and visual-only trials is similar, it suggests similar underlying representations. To test whether auditory cortex shows similar representations for phonemic and visemic information, we calculated a correlation between each of the phoneme-pairs across the phoneme and viseme group-averaged confusion matrices. Fig. 4f shows the scatter plot reflecting classification rate for each consonant pair. Across auditory-only and visual-only trials, classification (and misclassification) rates were highly correlated ($r^2 = .8003$, $p < .001$ permutation test). Significance was calculated by randomly permuting the stimulus labels of each trial and repeating the full classification analysis $n = 1000$ times. This is consistent with our hypothesis that the spatiotemporal neural representation of viseme identities in the auditory areas is similar to that of phonemes.

Three of the six subjects had electrodes along the ventral temporal lobe (including the fusiform gyrus). To examine phoneme and viseme representations in this visual region, we

repeated classification on this restricted set of electrodes. Within visual electrodes, group-level classification was significantly above chance for visual-only trials ($t(2) = 8.74$, $p = .0128$, $d = 5.05$) but not auditory-only trials ($t(2) = 1.15$, $p = .3701$, $d = 0.662$). We additionally observed greater classification in visual-only trials relative to auditory-only trials ($t(2) = 5.29$, $p = .0339$, $d = 3.05$). This pattern was seen at the individual-subject level in all three subjects using binomial statistics (all $p < .01$ for visual-only trials and all $p > .24$ for auditory-only trials).

Classification of ERPs revealed robust encoding of phoneme and viseme information in the auditory system, driven by low-frequency oscillatory information (power and phase) that reflects the excitatory/inhibitory balance of local neuronal populations[39]. Higher frequency activity (high-gamma power; HGp), in contrast, is associated with the average rate of action potentials generated by a local population of neurons[40]. Across HGp from all auditory electrodes, we observed significant classification (evaluated using binomial statistics; $p < .05$) in five out of six patients for auditory-only trials and three out of six patients for visual-only trials. Similarly, at the group-level we observed classification accuracy reliably above chance for both auditory-only ($t(5) = 3.74$, $p = .013$, $d = 1.53$), and visual-only trials ($t(5) = 3.56$, $p = .0162$, $d = 1.45$). We additionally observed a trend towards greater classification in auditory-only trials relative to visual-only trials ($t(5) = 2.40$, $p = .0614$, $d = 0.981$). More reliable classification for low-frequency signals evoked by visemes is consistent with the fMRI finding that classification can occur in auditory regions that do not show increased firing rates in response to visual speech.

Discussion

Extensive research has shown that silent visual speech can modulate activity within primary auditory regions in humans[13-15,23,41,42]. However, multiple sources of information could be

372 contained in these visual-driven auditory responses including visual motion timing information[43],
373 visual parsing of speech rate[11], visual-derived spectral information[1], general effects on attention or
374 arousal[44], or viseme-to-phoneme transformations[15]. Here we tested the hypothesis that the
375 identities of individual visemes are represented in the auditory system through distributed patterns
376 of activation, and these spatial distributions match corresponding phoneme representations. Using
377 fMRI and intracranial electrodes implanted in auditory regions we found that the auditory system
378 reliably encodes the identity of visemes using spatially distributed activity in a similar manner to
379 heard words. Moreover, visemes evoked spatially similar activity to matching phonemes,
380 consistent with the hypothesis that visual speech targets corresponding phoneme representations

381 Data from both fMRI and iEEG showed reliable classification of visemes from auditory
382 regions, maximal in the left pSTS and STG bilaterally. Visemic information is likely first encoded
383 in the visual cortex[45] and then relayed to the auditory cortex. Consistent with this view we observed
384 significant classification of visemes throughout visual regions (including early visual areas using
385 fMRI and the fusiform gyrus using iEEG). Future functional connectivity analyses can be used to
386 examine the paths of transmission of lipreading information from visual to auditory regions and
387 computational analyses to examine how viseme information is transformed into phoneme or
388 phonetically tuned features. For example, dynamic causal modelling (DCM) has previously shown
389 that visual speech modulates auditory processing through ventral and dorsal routes[46]. Because
390 auditory populations show opposing effects to visual speech (increased activity in posterior
391 STG/STS and suppression in mid- to anterior STG) DCM may reveal discrete pathways of
392 information transmission, such as fusiform to pSTS/pSTG connections and alternative pathways
393 to the mid- to anterior STG.

Classification of iEEG data enables inferences about *when* phoneme and viseme information is available to the auditory system. This temporal resolution is necessary to understand whether visemes are used by the auditory system at the same time that auditory phonemes are processed (at the perceptual level), or if viseme representations emerge only after auditory processes are completed to support categorical decisions about what was heard. The present data showed significant classification accuracy for both auditory-only and visual-only trials shortly after phoneme onset, indicating that visemic information is available to the auditory system at the same perceptual stage as is phonemic information. It remains possible that silent visual speech can encode visemic information in the auditory system before phoneme onset in cases that visual speech precedes auditory onset[25].

Mechanistically, we show that categorical visual speech information is likely encoded through the suppression of neural activity in mid- to anterior STG and increased activity in the posterior STG and STS. This is supported by converging evidence from iEEG and fMRI that silent visual speech evoked decreased BOLD and HGp in mid- to anterior STG regions (including primary auditory cortex), and increased BOLD and HGp in the posterior STG and STS. Despite these differences in activation, classification was observed throughout the STG and pSTS suggesting two distinct mechanisms through which visual information is used to modulate phoneme populations. In posterior activations, we suggest that silent visual speech selectively activates matching phoneme-tuned neurons in a categorical manner. Conversely, we suggest that visemes suppress activity in the STG in a targeted manner to inhibit incorrect representations in phonetically tuned neuronal populations[25] to indirectly refine the representation of correct phonetic features. While speculative, one possible explanation for why visemes avoid directly activating

matching phonetically tuned neurons is to limit the potential for crossmodal hallucinations[47]; i.e., hearing speech during silent lipreading.

A limitation of the present work is that the small set of phonemes and visemes presented provide a limited account of the full distribution of phonemes and visemes present in English. This limitation was necessary to ensure adequate signal-to-noise ratios to enable classification of the individual phonemes and visemes, but future research can examine the full distribution of phoneme and viseme representations using more natural speech stimuli[48] in auditory-visual contexts using high-density intracranial electrodes. Data from such experiments would be predicted to show that phoneme tuning functions (the spatial selectivity of responses to a specific phoneme) will be more precise (narrower and more distinct from other phonemes) during auditory-visual speech compared to auditory-only speech. Moreover, we predict that phoneme and viseme spatial maps will imperfectly overlap (as the same viseme could denote ‘pet’ or ‘bet’) and that the dissimilarity in phoneme and viseme maps explain categorical shifts in perception during the McGurk effect (a perceptual illusion in which visual speech alters which phoneme is heard[49]).

In sum, the present studies support the hypothesis that silent visual speech information is represented in the auditory system for the purpose of refining phonetic and phonemic population responses, to in turn support speech perception fluency. This crucial form of information shared between auditory and visual regions likely reflects only one type of signal shared, and leaves open the possibility that other visual features (e.g., visual motion timing, visual parsing of speech rate, visual-derived spectral information) modulate auditory neurons in complementary ways to support speech perception in the natural environment.

Methods

fMRI Experiment

Planned analyses and sample size stopping justification for the fMRI study were pre-registered at OSF (https://osf.io/6fzwd/?view_only=60484583a2bb4dcdb8e27788c7c4c373). Minor deviations from the pre-registered protocol are noted throughout the methods section. The study was approved by the Institutional Review Board (IRB) of the University of Michigan.

Subjects

FMRI data was acquired from n = 64 subjects (F = 47, M = 17) recruited from the University of Michigan's Psychology paid-subject pool (individuals who had previously expressed interest in research studies) and through word of mouth. Subjects' ages ranged from 18-32 (Mean: 22.87, SD = 3.29) and included 56 right-handed, 7 left-handed, and 1 ambidextrous individual. Written consent was obtained from each subject. subjects were paid USD $20 per hour for their time. Data was collected from each subject in a single session lasting approximately 1 hour and 15 minutes. Because power analyses using multivariate pattern analyses (MVPA) remain a challenge, we determined our sample size based on univariate power analysis (on the assumption that this would yield a minimum acceptable sample size). Sample size to detect visual-only effects was determined using data from the auditory-only condition in a preliminary sample using NeuroPower (using random field theory, cluster threshold p=.05, alpha=.05, n=27). Estimated sample sizes ranged from n=62 to 64 across the pairwise phoneme comparisons (/fafa/ vs /mama/, /fafa/ vs /kaka/, and /kaka/ vs /mama/) and n=64 was selected to ensure adequate power. No data from the visual-only condition was analyzed prior to submission of the pre-registration.

462 ***Tasks, Stimuli and Experimental Design***

463 We used an auditory and visual speech paradigm optimized for an event-related fMRI
464 design. On each trial, subjects were presented with a three-alternative forced-choice task that
465 consisted of either an auditory-only stimulus or a visual-only stimulus. Three types of phonemes;
466 /fafa/, /kaka/ and /mama/ and three types of visemes; /fafa/, /kaka/ and /mama/ were used for this
467 task. These specific phonemes were chosen to maximize the differentiability between the
468 individual phonemic representations in the neuronal populations of the STG[30,42]. Fig. 1a shows the
469 timing and structure of the task. Each trial for both the auditory-only and visual-only conditions
470 lasted for 2 seconds. The auditory-only trials began with a fixation cross against a black screen,
471 with the phonemes presented 250 ms after the appearance of the fixation cross. The visual-only
472 trial began with the appearance of a female actor's face on the screen, with lip movements
473 beginning 250 ms after face onset. After the presentation of each auditory-only or visual-only trial,
474 subjects were presented with 3 options (/fa/, /ka/, and /ma/) and were instructed to press one of
475 three associated buttons on an MRI-safe button response box.

476 The first 24 subjects were shown response choices that always appeared in the same order
477 (/fa/, /ka/, or /ma/) with a stable mapping between response choice and button (the index finger
478 was always used to make the response for /fa/, the middle finger for /ka/ and the ring finger for
479 /ma/). While performing the sample size estimates for our power analysis, we saw that the stable
480 mapping between response choices and button presses resulted in response type differentiability
481 in the motor cortex consistent with prior evidence for motor regions encoding information about
482 finger movements[50]. Hence, to counteract this effect and to negate the confounds of motor region
483 responses during speech perception[51], we altered the pre-registered protocol for the remaining 40
484 subjects, who were shown response choices that were randomized after each trial.

Subjects had 1.25 seconds to respond to the answer choices. If the subject failed to register a response within 1.25 seconds, the trial was recorded as a missed response. Every trial was followed by a 5-6 second jitter period (sampled from a uniform random distribution) which acted as the intertrial interval (ITI)[52]. In each run, subjects completed 60 trials that were split between 30 auditory-only and 30 visual-only trials, with 10 trials each for every phoneme and viseme; trial types and stimuli were randomly intermixed in each run.

In total, subjects completed five runs, resulting in 300 trials in total (150 phonemes, 150 visemes) during the task, with each run lasting 8 minutes and 30 seconds. Psychtoolbox was used for stimulus delivery and recording timing information and subject responses. Auditory stimuli were presented using fMRI compatible Avotec headphones that had integrated earmuffs in order to achieve maximum reduction of scanner noise. The sound level of stimuli was held constant for all subjects. While presenting auditory speech stimuli in an MRI scanner can be challenging, the undegraded nature of the auditory stimuli enabled near perfect accuracy throughout the task. A mirror system reflected the visual stimuli from an LCD projector onto a mirror (width of the mirror: 12cm, approximate viewing distance between eye and mirror: 15cm; width and height of the face on screen: 9cm x 12cm) located inside the magnet bore of the scanner.

Data Exclusion Criteria

To ensure that subjects included in analyses demonstrated persistent attention throughout the task, we pre-registered exclusion criteria to remove subjects with behavioral accuracy rates less than 75% for either auditory-only or visual-only conditions: no subjects were excluded based on this cutoff.

fMRI Data Collection

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

Data Processing

fMRI data was reconstructed with realignment and fieldmap correction applied using SPM12 to each of the five T2* runs for inhomogeneity recovery of signal in the B0 field. Physiological noise was removed using RETROICOR[53]. For both the univariate and multivariate analysis, preprocessing steps were completed using SPM12 (Wellcome Department of Cognitive Neurology, London, UK; https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). We utilized The Decoding Toolbox (https://sites.google.com/site/tdtdecodingtoolbox/; version 3.997) for the whole-brain multivariate analyses.

Preprocessing

Before preprocessing the functional images, SPM's display tool was used to set the origin of the anatomical volumes for each subject manually by picking the location of the anterior commissure. After this, functional volumes were reconstructed and realigned, physiological noise

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

Univariate Analysis

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

In the first level analysis, whole brain beta maps were generated individually for all seven conditions for each of the 64 subjects. These maps also included information from regressors for motion correction (six head movement parameters). Freesurfer's group-analysis pipeline was used for second level analyses[54]. Specifically, each subject's data was projected onto the cortical surface of the fsaverage subject (using the command mris_preproc) and smoothed using a FWHM of

10mm (using the command mri_surf2surf). General linear models were estimated with the command mri_glmfit separately for each hemisphere and condition, excluding motor and frontal areas due to the initial n = 24 subjects with consistent phoneme-motor mappings. Significant vertices were identified at the group level using the command mri_glmfit-sim using a vertex level threshold of $p < .001$ and cluster-level threshold of $p < .05$ (estimated with 10000 permutations) to control for multiple comparisons; p-values were adjusted for separate tests of the two hemispheres.

Multivariate Analysis

To identify regions that reliably differentiated classes of phonemes and classes of visemes, we performed searchlight based MVPA analyses. Preprocessing steps for univariate and multivariate analyses were matched except for the normalization and smoothing, such that for the multivariate analysis, these two steps were performed after the first level analysis was completed. For the decoding analysis, we utilized The Decoding Toolbox[55] with a LIBSVM[56] based support vector machine (SVM) implementation. For each of the individual subjects, we built a SVM classifier with a cross-validation scheme for the five runs. We used these classifiers to build two separate models: one to classify between the three phonemes and the other to classify between the three visemes. The phoneme models were constructed to identify voxels that reliably decoded the identity of each of the three phonemes while the viseme models were built to identify voxels that reliably decoded the identity for each of the three visemes. These models were implemented as independent whole-brain searchlight analysis in the first level of the MVPA model. For each of the models, beta estimates were calculated and extracted from a 3-voxel radius sphere. 4 fMRI runs were used for training and 1 run for testing in an iterative manner. The searchlight center was

shifted through voxel-wise patterns throughout the brain to extract whole-brain accuracy maps for auditory-only and visual-only conditions. Chance-level accuracy (33.3%) was subtracted from individual subjects and conditions so that null-hypothesis values could be set to zero. Group-level analyses and multiple comparison corrections were performed using Freesurfer and matched those in the Univariate Analyses.

ROI-Based Decoding Analyses

Following the whole-brain searchlight analysis, we selected five regions of interests from each hemisphere (ROI) based on results from literature[23,41,42]. Four ROIs (STG, pSTS, fusiform, and hMT+) were pre-registered. The fifth ROI (V1/V2) was included in the classification analyses given the strong univariate response in the visual-only condition. ROIs were identified at the individual subject level based on Freesurfer aparc-aseg labeling[57]. Selected labels included 'superiortemporal', 'bankssts','MT_exvivo.thresh', the combinated labels 'FG1.mpm.vpnl' to 'FG4.mpm.vpnl', and the combined labels 'V1_exvivo.thresh' and 'V2_exvivo.thresh'. Contrast beta estimates (condition vs fixation) were extracted for each subject, stimulus (6 phonemes and visemes), block, and ROI. SVM analyses were performed at the individual subject level with models trained on n-1 blocks (leave-one-out classification) using the 'fitcecoc' function in MATLAB.

Multivariate Similarity Analysis

We used the same data from the ROI-based decoding analyses to examine the correlation of spatial activity across the conditions within the STG. We restricted vertices to those with significant classification in both auditory-only and visual-only conditions (purple vertices in Fig.

2c) within the STG and calculated the correlation between vertex-wise beta estimates for phoneme and viseme pairs. At the single subject level, we then correlated the spatial distribution of STG activity across each of the 6 stimuli (3 phonemes and 3 visemes) in a pairwise manner. We averaged correlations across matching pairs (e.g., the phoneme /ma/ and the viseme /ma/) and separately mismatching pairs (e.g., the phoneme /ma/ and the visemes /ka/ and /fa/), to yield a pair of values for each subject, and then compared these values at the group level to examine whether visemes evoke similar spatial distributions of activity to the matching phoneme (e.g., that the viseme MA evokes a more similar spatial layout to the phoneme MA compared to the phoneme KA).

IEEG Experiment

The study was approved by the Institutional Review Boards (IRB) at the University of Michigan and Henry Ford Hospitals.

Subjects and Recordings

N = 6 patients (2 female, 4 male) undergoing clinical evaluation using iEEG for intractable epilepsy consented to participate in this study under an institutional review board (IRB) approved protocol at the University of Michigan or Henry Ford hospital. Patients' ages ranged from 12-39 years of age (mean = 29.7, std = 9.8) and 5 were right-handed (one patient self-reported to be ambidextrous). All patients were native English speakers. Clinically implanted depth electrodes (5 mm center-to-center spacing) and/or subdural electrodes (10 mm center-to-center spacing) were used to acquire iEEG data from subjects. IEEG data from a total of 459 electrodes were recorded from the six subjects. The type and location of electrodes implanted were based on the clinical

needs of the patients. Electrodes were implanted within left auditory areas for 2 patients and right auditory areas for 4 patients. IEEG recordings were acquired at either 4096 Hz (n = 4 patients) or 1000 Hz (n = 2 patients) due to differences in clinical amplifiers.

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[58]. The CT was not resliced or resampled. The localization of each electrode was performed using custom software[59]. 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/[60,61]).

Task and Stimuli

Subjects were tested at their bedside in an Epilepsy Monitoring Unit using a laptop running Psychtoolbox[62]. The task paradigm was adapted from a prior study [63] which was designed to behaviorally study multiple aspects of auditory-visual speech integration. The stimuli consisted of a female speaker who produced 40 commonly used 1-2 syllable words that each started with one of the four consonants: *'b', 'f', 'g', 'd'* (10 of each). The phoneme in the second position of each of these words was generally balanced across each of the four groups. Each stimulus was recorded at a frame rate of 29.97 frames per second, and trimmed to 1100 ms in length. Further adjustments

were made such that the first consonantal burst of each word occurred at 500 ms during the video playback by removing leading video frames.

Each subject underwent two task variants using the same stimuli and task design to increase trial numbers, and to reduce classifier overfitting. Supp. Fig. 3 shows the task schematic for both variants of the task. In variant one, subjects were presented with words one at a time, in one of two main conditions: auditory-only or visual-only. Subjects then identified the initial speech sound of the presented stimulus using a button press to select one of four options shown on the computer screen. For example, on a trial with the word *"bag"*, the options presented to the subject were *'b', 'g', 'd', 'th'*. The paradigm included 40 trials per consonant in each main condition, such that each of the 40 words were presented 4 times in the visual-only condition and another 4 times in the auditory-only condition. This resulted in a total of 320 trials for each subject using task variant 1. The words used in our task are presented in Supplemental Table 5.

In task variant 2, subjects were presented with trials in one of four main conditions: auditory-only, visual-only, congruent audiovisual, or incongruent audiovisual. Task stimuli and instructions were the same as in variant 1. Variant 2 included 20 trials per consonant in each main condition. A second factor that was manipulated in this variant was the background noise level of the stimuli such that half of the words used in each condition were presented in either a low noise or a high noise context. In the low noise context, the auditory stimuli were presented as they were recorded (SNR = 32.2 dB SPL). In the high noise context, pink noise was added to reduce the signal-to-noise (SNR) ratio of the signals to -6 db SPL. In this task variant, only data from the auditory-only and visual-only conditions were included in analyses because they matched the main conditions obtained from Task variant 1. This resulted in a total of 80 auditory-only and 80 visual-only trials for each subject using task variant 2.

A total of 480 trials (Task variant 1: 320 trials, task variant 2: 160 trials) with 60 trials for each consonant (*'b', 'g', 'd', 'f'*) per condition was obtained from the combined data of both task variants. Each subject received a randomized trial order. For the auditory-only condition, a gray rectangle was presented 500 ms before sound onset. Stimuli offset occurred 600 ms after sound onset time. In the visual-only condition, face onset occurred 500 ms before the time when phoneme onset would naturally occur. A wait time of 1.25 seconds was provided for the subjects to respond to each of the stimuli.

IEEG Data Preprocessing

Data were preprocessed using bipolar referencing, such that signals from adjacent electrodes were subtracted in a pairwise manner. This ensured that the final signals of interest were obtained from neuronal populations that provided maximal localized responses[64]. Analyses in auditory regions were restricted to electrodes (registered in MNI space) that were within 10 mm of the Freesurfer anatomical labels 'superiotemporal', 'middletemporal' or 'supramarginal'. Excessively noisy electrodes were removed either manually or statistically by identifying electrodes with raw signals that were 5 SD greater in comparison to all other electrodes. For complementary analyses in visual regions, electrode locations were anatomically restricted to the 'inferiortemporal' and 'fusiform' labels.

Drift was removed from each channel (using residuals from fits to a 3rd order polynomial and high-pass filtering at 0.1 Hz). Power-line interference was removed by notch-filtering at 60 Hz and its harmonics. ERPs were extracted from this minimally processed signal. HGp activity was extracted from the continuous time-series after wavelet convolution and power transformation (70-150 Hz in 5 Hz intervals, wavelet cycles = 20 at 70 Hz, and increased linearly to maintain the

same wavelet duration across frequencies). ERP and HGp data were segmented into 2 second epochs centered around speech onset time for a specific stimulus: trial onset was defined as the point when the initial consonant burst occurred. All data were then resampled to 1000 Hz.

Electrodes from both the left and right hemispheres were projected into the left hemisphere for analyses and visualization. This projection was performed by registering each subject's skull-stripped brain to the Freesurfer cvs_avg35_inMN152 template image through affine registration using the Freesurfer function 'mri_robust_register'[65]. Right hemisphere electrode coordinates were then reflected onto the left hemisphere across the sagittal axis.

Classifiers for Calculating Decoding Accuracy

A support vector machine[66] classifier was utilized for calculating decoding accuracy. Classifiers for stimulus trials were built for individual subjects and group-level analyses were performed by combining results from individual subjects (subject as a random effect). Classification was performed on downsampled data (10 Hz except where stated otherwise) to reduce dimensional complexity. Phonemes and visemes were classified using a 4-fold multiclass classifier from 0 to 500 ms following sound onset time (or the corresponding point in the visual movie); electrodes and time-points were treated as dimensions in the classification of individual trials. Time-series analyses were performed independently at each time-point (500 Hz) and accuracies were smoothed at the individual subject-level across 20 time points using the Matlab function 'movmean'. Electrode-level analyses were performed on individual electrodes located in auditory regions.

Similarity Analysis

To test whether auditory cortex showed a similar representation for phonemic and visemic information, we examined the similarity of the phoneme and viseme confusion matrices. Specifically, we paired each of the 16 cells in the two confusion matrices and used Pearson correlation to examine their relationship. Significance was calculated by randomly permuting the stimulus labels of each trial and repeating the full classification analysis $n = 1000$ times.

Calculating Individual Subject Classification Significance

The four classes tested within each condition yielded chance levels of classification at 25%. To calculate significance above this chance level, we used binomial statistics for within-subject significance testing[67,68]. We used the 'binocdf' function in MATLAB for this, by considering two parameters: the number of trials, and probability of success at each instance (25%). This gives rise to a binomial chance-level probability that varies depending on the number of data points used for classification in each of the models that were built. This resulted in a chance probability of 29.58% ($p = 0.05$) for a 4-class classifier with 240 trials.

References

1 Plass, J., Brang, D., Suzuki, S. & Grabowecky, M. Vision perceptually restores auditory spectral dynamics in speech. *Proc Natl Acad Sci U S A* **117**, 16920-16927 (2020). https://doi.org:10.1073/pnas.2002887117

2 Micheli, C. *et al.* Electrocorticography reveals continuous auditory and visual speech tracking in temporal and occipital cortex. *Eur J Neurosci* **51**, 1364-1376 (2020). https://doi.org:10.1111/ejn.13992

3 Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C. & Foxe, J. J. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral cortex* **17**, 1147-1153 (2007).

4 Rosemann, S. & Thiel, C. M. Audio-visual speech processing in age-related hearing loss: Stronger integration and increased frontal lobe recruitment. *Neuroimage* **175**, 425-437 (2018).

5 Aabedi, A. A. *et al.* Convergence of heteromodal lexical retrieval in the lateral prefrontal cortex. *Sci Rep* **11**, 6305 (2021). https://doi.org:10.1038/s41598-021-85802-5

6 Caplan, D., Gow, D. & Makris, N. Analysis of lesions by MRI in stroke patients with acoustic-phonetic processing deficits. *Neurology* **45**, 293-298 (1995).

7 Crinion, J. T., Warburton, E. A., Lambon-Ralph, M. A., Howard, D. & Wise, R. J. Listening to narrative speech after aphasic stroke: the role of the left anterior temporal lobe. *Cereb Cortex* **16**, 1116-1125 (2006). https://doi.org:10.1093/cercor/bhj053

8 Kraus, N. *et al.* The neural legacy of a single concussion. *Neurosci Lett* **646**, 21-23 (2017). https://doi.org:10.1016/j.neulet.2017.03.008

9 Thompson, E. C. *et al.* Difficulty hearing in noise: a sequela of concussion in children. *Brain Inj* **32**, 763-769 (2018). https://doi.org:10.1080/02699052.2018.1447686

10 Anderson, C. A., Wiggins, I. M., Kitterick, P. T. & Hartley, D. E. Adaptive benefit of cross-modal plasticity following cochlear implantation in deaf adults. *Proceedings of the National Academy of Sciences* **114**, 10256-10261 (2017).

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

12 Fisher, C. G. Confusions among visually perceived consonants. *Journal of speech and hearing research* **11**, 796-804 (1968).

13 Calvert, G. A. *et al.* Activation of auditory cortex during silent lipreading. *science* **276**, 593-596 (1997).

14 Pekkola, J. *et al.* Primary auditory cortex activation by visual speech: an fMRI study at 3 T. *Neuroreport* **16**, 125 (2005).

15 Karthik, G. *et al.* Visual speech differentially modulates beta, theta, and high gamma bands in auditory cortex. *Eur J Neurosci* **54**, 7301-7317 (2021). https://doi.org:10.1111/ejn.15482

16 Thézé, R., Giraud, A.-L. & Mégevand, P. The phase of cortical oscillations determines the perceptual fate of visual cues in naturalistic audiovisual speech. *Science advances* **6**, eabc6348 (2020).

17 Mégevand, P. *et al.* Crossmodal phase reset and evoked responses provide complementary mechanisms for the influence of visual speech in auditory cortex. *Journal of Neuroscience* **40**, 8530-8542 (2020).

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

19 Formisano, E., De Martino, F., Bonte, M. & Goebel, R. " Who" is saying" what"? Brain-based decoding of human voice and speech. *Science* **322**, 970-973 (2008).

20 Chang, E. F. *et al.* Categorical speech representation in human superior temporal gyrus. *Nat Neurosci* **13**, 1428-1432 (2010). https://doi.org:10.1038/nn.2641

21 Raizada, R. D., Tsao, F. M., Liu, H. M. & Kuhl, P. K. Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. *Cereb Cortex* **20**, 1-12 (2010). https://doi.org:10.1093/cercor/bhp076

22 Leonard, M. K., Baud, M. O., Sjerps, M. J. & Chang, E. F. Perceptual restoration of masked speech in human cortex. *Nat Commun* **7**, 13619 (2016). https://doi.org:10.1038/ncomms13619

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

24 Plata Bello, J. *et al.* Visual inputs decrease brain activity in frontal areas during silent lipreading. *PloS one* **14**, e0223782 (2019).

25 Karas, P. J. *et al.* The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. *Elife* **8**, e48116 (2019). https://doi.org:10.7554/eLife.48116

26 Blank, H. & Davis, M. H. Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS biology* **14**, e1002577 (2016).

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

28 Kilian-Hütten, N., Valente, G., Vroomen, J. & Formisano, E. Auditory cortex encodes the perceptual interpretation of ambiguous sound. *Journal of Neuroscience* **31**, 1715-1720 (2011).

29 Bonte, M., Hausfeld, L., Scharke, W., Valente, G. & Formisano, E. Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *Journal of Neuroscience* **34**, 4548-4557 (2014).

30 Mesgarani, N. & Chang, E. F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**, 233-236 (2012). https://doi.org:10.1038/nature11020

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

32 Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc Natl Acad Sci U S A* **103**, 3863-3868 (2006). https://doi.org:10.1073/pnas.0600244103

33 Arsenault, J. S. & Buchsbaum, B. R. Distributed neural representations of phonological features during speech perception. *Journal of Neuroscience* **35**, 634-642 (2015).

34 Schwartz, J.-L. & Savariaux, C. No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology* **10**, e1003743 (2014).

35 Arnal, L. H., Morillon, B., Kell, C. A. & Giraud, A. L. Dual neural routing of visual facilitation in speech processing. *J Neurosci* **29**, 13445-13453 (2009). https://doi.org:10.1523/JNEUROSCI.3194-09.2009

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

37 Bernstein, L. E., Jiang, J., Pantazis, D., Lu, Z.-L. & Joshi, A. Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. *Human Brain Mapping* **32**, 1660-1676 (2011). https://doi.org:https://doi.org/10.1002/hbm.21139

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

39 Gao, R., Peterson, E. J. & Voytek, B. Inferring synaptic excitation/inhibition balance from field potentials. *Neuroimage* **158**, 70-78 (2017).

40 Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J. & Hsiao, S. S. Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *Journal of Neuroscience* **28**, 11526-11536 (2008).

41 Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H. & Martin, A. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature neuroscience* **7**, 1190-1192 (2004).

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

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

44 Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S. & Puce, A. Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci* **12**, 106-113 (2008). https://doi.org:10.1016/j.tics.2008.01.002

45 Nidiffer, A. R., Cao, C. Z., O'Sullivan, A. & Lalor, E. C. A linguistic representation in the visual system underlies successful lipreading. *bioRxiv* (2021).

46 Zhang, L. & Du, Y. Lip movements enhance speech representations and effective connectivity in auditory dorsal stream. *NeuroImage*, 119311 (2022).

47 Nair, A. & Brang, D. Inducing synesthesia in non-synesthetes: Short-term visual deprivation facilitates auditory-evoked visual percepts. *Consciousness and cognition* **70**, 70-79 (2019).

48 Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006-1010 (2014). https://doi.org:10.1126/science.1245994

49 McGurk, H. & MacDonald, J. Hearing lips and seeing voices. *Nature* **264**, 746-748 (1976).

50 Shen, G. *et al.* Decoding the individual finger movements from single-trial functional magnetic resonance imaging recordings of human brain activity. *European Journal of Neuroscience* **39**, 2071-2082 (2014).

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

52 Zeithamova, D., de Araujo Sanchez, M.-A. & Adke, A. Trial timing and pattern-information analyses of fMRI data. *Neuroimage* **153**, 221-231 (2017).

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

54 Hagler Jr, D. J., Saygin, A. P. & Sereno, M. I. Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. *Neuroimage* **33**, 1093-1103 (2006).

55 Hebart, M. N., Görgen, K. & Haynes, J.-D. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in neuroinformatics* **8**, 88 (2015).

56 Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* **2**, 1-27 (2011).

57 Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968-980 (2006).

58 Viola, P. & Wells III, W. M. Alignment by maximization of mutual information. *International journal of computer vision* **24**, 137-154 (1997).

59 Brang, D., Dai, Z., Zheng, W. & Towle, V. L. Registering imaged ECoG electrodes to human cortex: A geometry-based technique. *J Neurosci Methods* **273**, 64-73 (2016). https://doi.org:10.1016/j.jneumeth.2016.08.007

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

61 Fischl, B., Sereno, M. I. & Dale, A. M. Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage* **9**, 195-207 (1999).

62 Kleiner, M., Brainard, D. & Pelli, D. What's new in Psychtoolbox-3? (2007).

63 Ross, L. A. *et al.* Impaired multisensory processing in schizophrenia: deficits in the visual enhancement of speech comprehension under noisy environmental conditions. *Schizophr Res* **97**, 173-183 (2007). https://doi.org:10.1016/j.schres.2007.08.008

64 Yao, D. *et al.* Which reference should we use for EEG and ERP practice? *Brain topography* **32**, 530-549 (2019).

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

66 Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20**, 273-297 (1995).

67 Demandt, E. *et al.* Reaching movement onset-and end-related characteristics of EEG spectral power modulations. *Frontiers in neuroscience* **6**, 65 (2012).

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