# A Double Dissociation between Anterior and Posterior Superior Temporal Gyrus for Processing Audiovisual Speech Demonstrated by Electrocorticography

**Muge Ozker[1,2], Inga M. Schepers[3], John F. Magnotti[2], Daniel Yoshor[2], and Michael S. Beauchamp[2]**

## Abstract

■ Human speech can be comprehended using only auditory information from the talker's voice. However, comprehension is improved if the talker's face is visible, especially if the auditory information is degraded as occurs in noisy environments or with hearing loss. We explored the neural substrates of audiovisual speech perception using electrocorticography, direct recording of neural activity using electrodes implanted on the cortical surface. We observed a double dissociation in the responses to audiovisual speech with clear and noisy auditory component within the superior temporal gyrus (STG), a region long known to be important for speech perception. Anterior STG showed greater neural activity to audiovisual speech with clear auditory component, whereas posterior STG showed similar or greater neural activity to audiovisual speech in which the speech was replaced with speech-like noise. A distinct border between the two response patterns was observed, demarcated by a landmark corresponding to the posterior margin of Heschl's gyrus. To further investigate the computational roles of both regions, we considered Bayesian models of multisensory integration, which predict that combining the independent sources of information available from different modalities should reduce variability in the neural responses. We tested this prediction by measuring the variability of the neural responses to single audiovisual words. Posterior STG showed smaller variability than anterior STG during presentation of audiovisual speech with noisy auditory component. Taken together, these results suggest that posterior STG but not anterior STG is important for multisensory integration of noisy auditory and visual speech. ■

## INTRODUCTION

Human speech perception is multisensory, combining auditory information from the talker's voice with visual information from the talker's face. Visual speech information is particularly important in noisy environments in which the auditory speech is difficult to comprehend (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Bernstein, Auer, & Takayanagi, 2004; Sumby & Pollack, 1954). Although visual speech can substantially improve the perception of noisy auditory speech, little is known about the neural mechanisms underlying this perceptual benefit.

Speech varies on a timescale of milliseconds, requiring the brain to accurately integrate auditory and visual speech with high temporal fidelity. However, the most popular technique for measuring human brain activity, BOLD fMRI, is an indirect measure of neural activity with a temporal resolution on the order of seconds, making it difficult to accurately measure rapidly changing neural responses to speech. To overcome this limitation, we recorded from the brains of participants implanted with electrodes for the treatment of epilepsy. This technique, known as electrocorticography, allows for the direct measurement of activity in small populations of neurons with millisecond precision. We measured activity in electrodes implanted over the superior temporal gyrus (STG), a key brain area for speech perception (Mesgarani, Cheung, Johnson, & Chang, 2014; Binder et al., 2000), as participants were presented with audiovisual speech with either clear or noisy auditory or visual components.

The STG is functionally heterogeneous. Regions of anterior STG lateral to Heschl's gyrus are traditionally classified as unisensory auditory association cortex (Rauschecker, 2015). In contrast, regions of posterior STG and STS are known to be multisensory, responding to both auditory and visual stimuli including faces and voices, letters and voices, and recordings and videos of objects (Reale et al., 2007; Miller & D'Esposito, 2005; Beauchamp, Lee, Argall, & Martin, 2004; van Atteveldt, Formisano, Goebel, & Blomert, 2004; Foxe et al., 2002; Calvert, Campbell, & Brammer, 2000).

On the basis of this distinction, we hypothesized that anterior and posterior regions of STG should differ in their electrocorticographic response to clear and noisy audiovisual speech. We expected that auditory association areas in anterior STG should respond strongly to speech with a clear auditory component but show

[1]University of Texas Graduate School of Biomedical Sciences at Houston, [2]Baylor College of Medicine, [3]University of Oldenburg



Journal of Cognitive Neuroscience 29:6, pp. 1044–1060
doi:10.1162/jocn_a_01110

a reduced response to the reduced information available in speech with noisy auditory component. Multisensory areas in posterior STG should be able to use the clear visual speech information to compensate for the noisy auditory speech, resulting in similar responses to speech with clear and noisy auditory components.

A related set of predictions comes from theoretical models of Bayesian integration. In these models, sensory noise and the resulting neural variability is independent in each modality. Combining the modalities through multisensory integration results in a decreased neural variability (and improved perceptual accuracy) relative to unisensory stimulation (Fetsch, Pouget, DeAngelis, & Angelaki, 2012; Knill & Pouget, 2004). Bayesian models predict that unisensory areas, such as those in anterior STG, should have greatly increased variability as the sensory noise in their preferred modality increases. Multisensory areas, like those in posterior STG, should be less influenced by the addition of auditory noise, resulting in similar variability for speech with clear and noisy auditory components.

## METHODS

### Participant Information

All experimental procedures were approved by the institutional review board of Baylor College of Medicine. Five human participants with refractory epilepsy (3 women, mean age = 31 years) were implanted with subdural electrodes guided by clinical requirements. Following surgery, participants were tested while resting comfortably in their hospital bed in the epilepsy monitoring unit.

### Stimuli, Experimental Design, and Task

Visual stimuli were presented on an LCD monitor positioned at 57-cm distance from the participant, and auditory stimuli were played through loudspeakers positioned next to the participant's bed. Two video clips of a female talker pronouncing the single syllable words "rain" and "rock" with clear auditory and visual components (AV) were selected from the Hoosier Audiovisual Multitalker Database (Sheffert, Lachs, & Hernandez, 1996). The duration of each video clip was 1.4 sec, and the duration of the auditory stimulus was 520 msec for "rain" and 580 msec for "rock." The auditory word onsets were 410 msec for "rain" and 450 msec for "rock" after the video onset. The face of the talker subtended approximately 15° horizontally and 15° vertically.

Speech stimuli were consisted of four conditions: Speech with clear auditory and visual components (AV), clear visual but noisy auditory components (AnV), clear auditory but noisy visual components (AVn), and finally noisy auditory and noisy visual components (AnVn).

To create speech stimuli with a noisy auditory component, the auditory component of the speech stimulus was replaced with noise that matched the spectrotemporal power distribution of the original auditory speech. The total power of this speech-specific noise was equated to the total power of the original auditory speech (Schepers, Schneider, Hipp, Engel, & Senkowski, 2013). This process generated speech-like noise.

To create speech stimuli with a noisy visual component, the visual component of the speech stimulus was blurred using a 2-D Gaussian low-pass filter (MATLAB function *fspecial*, filter size = 30 pixels in each direction). Each video frame (image size = 200 × 200 pixels) was filtered separately using 2-D correlation (MATLAB function *imfilter*). Values outside the bounds of the images were assumed to equal the nearest image border. These filter settings resulted in highly blurred videos.

Thirty-two to 56 repetitions of each condition were presented in random sequence. Each 5.4-sec trial consisted of a single 1.4-sec video clip followed by an ISI of 4 sec during which a fixation cross on a gray screen was presented. Participants pressed a mouse button to report which word was presented.

### Electrode Localization and Recording

Before surgery, T1-weighted structural MRI scans were used to create cortical surface models (Figure 1A) with FreeSurfer (Dale, Fischl, & Sereno, 1999; Fischl, Sereno, & Dale, 1999) and visualized using SUMA (Argall, Saad, & Beauchamp, 2006). Participants underwent a whole-head CT after the electrode implantation surgery. The postsurgical CT scan and presurgical MR scan were aligned using AFNI (Cox, 1996), and all electrode positions were marked manually on the structural MR images. Electrode positions were then projected to the nearest node on the cortical surface model using the AFNI program *SurfaceMetrics*. Resulting electrode positions on the cortical surface model were confirmed by comparing them with the photographs taken during the implantation surgery.

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

### Electrophysiological Data Analysis

Data were analyzed in MATLAB 8.5.0 (MathWorks, Inc. Natick, MA) using the FieldTrip toolbox (Oostenveld,

Fries, Maris, & Schoffelen, 2011). To remove common artifacts, the average signal across all electrodes was subtracted from each individual electrode's signal (common average referencing). The continuous data stream was epoched into trials. Line noise at 60, 120, 180 Hz was removed, and the data were transformed to time–frequency space using the multitaper method (three Slepian tapers, frequency window from 10 to 200 Hz, frequency steps of 2 Hz, time steps of 10 msec, temporal smoothing of 200 msec, frequency smoothing of ±10 Hz).

Our primary measure of neural activity was the broadband response in the high-gamma frequency band, ranging from 70 to 110 Hz. This frequency range is thought to reflect the frequency of action potentials in nearby neurons (Jacques et al., 2016; Ray & Maunsell, 2011; Nir et al., 2007; Mukamel et al., 2005). For each trial, the high-gamma response was measured in a window from 0 to 500 msec following auditory stimulus onset (reflecting the ~500 msec duration of the auditory stimulus) and converted to percent signal change measure by comparing the high-gamma response to a within-trial baseline window encompassing −500 to −100 msec before auditory stimulus onset. For instance, a 100% signal change on one trial would mean the power in the high-gamma band doubled from the pre-stimulus to the post-stimulus interval. For each electrode, the mean percent signal change in the high-gamma band across all trials of a given condition was calculated ($\mu$).

Our second analysis focused on neural variability across repeated presentations of identical stimuli. One obvious measure of variability is variance (defined as the square of the standard deviation across all observations). However, the variance of neural responses is known to increase with increasing response amplitude (Ma, Beck, Latham, & Pouget, 2006; Tolhurst, Movshon, & Dean, 1983), and our initial analysis demonstrated differences in response amplitude between speech with clear and noisy auditory components (Table 1). To search for variability differences without the confound of these amplitude differences, we used a different measure of variability known as the coefficient of variation (CV), which normalizes across amplitude differences by dividing the standard deviation of the response across trials by the mean response amplitude ($CV = \sigma/\mu$; Churchland et al., 2010; Gur, Beylin, & Snodderly, 1997). The CV assumes that variance covaries linearly with amplitude. We tested this assumption by calculating the Pearson correlation between the mean and variance of the high-gamma response across all anterior and posterior STG electrodes and found it to be reasonable for the four different stimulus conditions (AV: $r = .96, p = 10^{-16}$; AnV: $r = .86, p = 10^{-8}$; AVn: $r = .97, p = 10^{-16}$; AnVn: $r = .91, p = 10^{-11}$). Although CV has the advantage of accounting for the known correlation between amplitude and variance, it has the disadvantage that it becomes undefined as response amplitude approaches zero. For this reason, response amplitudes of less than 15% were excluded from the CV analysis, affecting 3 of 16 anterior electrodes in Figure 3 and 8 of 216 condition-electrode pairs in Table 2 and Table 7.

### Anatomical Classification and Electrode Selection

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

The cortical surface atlases supplied with FreeSurfer were used to automate ROI creation. The entire segmented STG was obtained from the Destrieux atlas (right hemisphere STG atlas value = 152, left hemisphere = 78; Destrieux, Fischl, Dale, & Halgren, 2010) and the anterior and posterior boundaries of the posterior STG were obtained from the Desikan-Killiany atlas (RH = 44, LH = 79; Desikan et al., 2006).

A total of 527 intracranial electrodes were recorded from. Of these, 55 were located on the STG. These were examined for stimulus-related activity, defined as significant high-gamma responses to audiovisual speech compared with prestimulus baseline ($p < 10^{-3}$, equivalent to ~40% increase in stimulus power from baseline). A total of 27 electrodes met both anatomical and functional criteria and were selected for further analysis. To simplify future meta-analyses and statistical comparisons between experiments, we do not report $p$ values as inequalities but instead report actual values (rounded to the nearest order of magnitude for $p$ values less than .001).

### Response Timing Measurements

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

### Linear Mixed Effects Modeling

We used the *lme4* package (Bates, Mächler, Bolker, & Walker, 2014) available for the R statistical language

(R Core Team, 2015) to perform a linear mixed effect (LME) analysis of the relationship between the neural response and both fixed and random factors that may influence the response. For the main LME analyses (Tables 1–5), the fixed factors were the location of each electrode (Anterior or Posterior), the presence or absence of auditory noise (Clear A or Noisy A), and the presence or absence of visual noise Clear V of Noisy V. The random factors were the mean response of each electrode across all conditions and the stimulus exemplar. The use of stimulus exemplar as a random factor accounts for differences in response to individual stimuli and allows for inference beyond the levels of the factors tested in the particular experiment (i.e., generalization to other stimuli).

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

### Additional Experiment: Varying Levels of Auditory Noise

In an additional control experiment, we recorded responses to audiovisual speech with varying levels of auditory noise. Similar to the main experiment, for each auditory word, noise that matched the spectrotemporal power distribution of the auditory speech was generated, then noise and the original auditory speech were added together with different weights while keeping the total power constant (Schepers et al., 2013). We parametrically increased the amount of auditory noise in 11 steps from 0% to 100% in 10% increments. Forty-two to 44 repetitions were presented for each noise level. The participant's task was to discriminate between four different words: Rain, Rock, Neck, and Mouth.

### Model Creation

A simple Bayesian model was constructed to aid in interpretation of the data (Figure 6) using a recently developed model of human multisensory speech perception (Magnotti & Beauchamp, 2017). Briefly, the high-dimensional neuronal response vector is conceptualized as a point in 2-D space. In this space, the $x$ axis represents auditory feature information and the $y$ axis represents visual feature information. Speech tokens are located at a fixed point in this space (shown in Figure 6 as the black dot at the center of each ellipse). For each presentation of an audiovisual speech stimulus, the brain encodes the auditory and visual information with noise. Over many trials, we characterize the distribution of the encoded speech stimulus as an ellipse. The axes of the ellipse correspond to the relative precision of the representation along each axis. Modalities are encoded separately, but through extensive experience with audiovisual speech, encoding a unisensory speech stimulus provides some information about the other modality. Although the results are robust across a range of parameters, for demonstration purposes, we assume that the variability of the preferred to non-preferred modality for audiovisual speech with a clear auditory component is 2:1 (shown in Figure 6 as the asymmetry of the ellipses in the auditory and visual representations). The integrated representation is formed according to Bayes rule, which combines the two modalities into a single representation that has smaller variance than either of the component modalities: $\Sigma_{AV} = (\Sigma_A^{-1} + \Sigma_V^{-1})^{-1}$ (Ma, Zhou, Ross, Foxe, & Parra, 2009). For audiovisual speech with noisy auditory component, we assume that the variability in the auditory representation increases by 150% while keeping the relative variability at the same ratio of 2:1 (shown in Figure 6 as larger ellipse). We model the visual representation of speech with noisy auditory component as being either identical to the representation of speech with a clear auditory component or with a gain term that reduces variability by 50% (with the relative variability remaining at 2:1). The multisensory representation is calculated in the same fashion with and without gain.

## RESULTS

Across participants, a total of 27 speech-responsive electrodes were identified on the STG. Using the posterior border of Heschl's gyrus as an anatomical landmark, 16 of these electrodes were located over anterior STG and 11 electrodes were located over posterior STG (Figure 1A). We hypothesized that the presence of noise in the speech stimulus (Figure 1B–E) might differentially affect responses in anterior and posterior electrodes. To test this hypothesis, we used the response amplitude in the gamma band as the dependent measure and fit a LME model with electrode location (Anterior vs. Posterior), the presence or absence of auditory noise in the stimulus (Clear A vs. Noisy A), and the presence or absence of visual noise in the stimulus (Clear V vs. Noisy V) as fixed factors. To account for overall differences in response amplitude across electrodes and stimulus exemplars, these were added to the model as random factors.

### Amplitude of the Responses to Clear and Noisy Speech

As shown in Table 1, there were three significant effects in the LME model. There was a small but significant effect of electrode location ($p = .01$) driven by a smaller overall response in posterior electrodes (Anterior vs. Posterior: 136 ± 27% vs. 101 ± 24%, mean signal change from baseline averaged across all stimulus conditions ± *SEM*) and

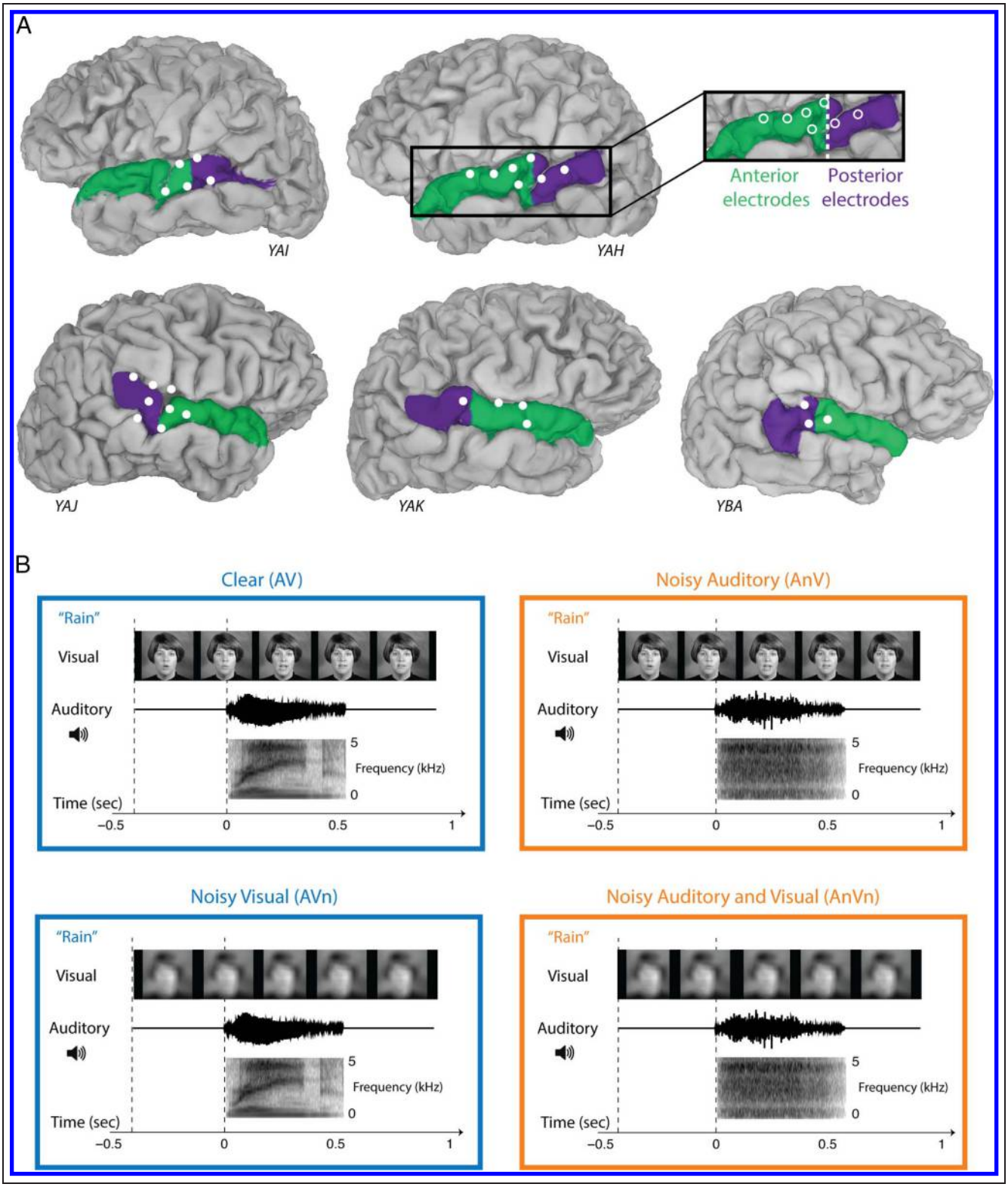


**Figure 1.** Electrode locations and audiovisual speech stimuli. (A) Cortical surface models of the brains of five participants (with anonymized subject ID). White circles show the location of implanted electrodes with a significant response to speech stimuli in the left hemisphere (top row) and right hemisphere (bottom row). In each hemisphere, the STG was parcellated into anterior (green) and posterior (purple) portions, demarcated by the posterior-most portion of Heschl's gyrus. (B) Clear audiovisual speech (AV) consisted of a movie of a talker pronouncing the word "rain" or "rock." Visual stimulus (top row) shows sample frames from the video. Auditory stimulus is shown as sound pressure level (middle row) and spectrogram (bottom row). Black vertical dashed lines indicate visual and auditory stimulus onsets. For noisy auditory speech (AnV), the auditory component was replaced with speech-specific noise of equal power to the original auditory speech. For noisy visual speech (AVn), the visual component was blurred using a low-pass Gaussian filter. For noisy auditory and noisy visual speech (AnVn), the auditory component was replaced with speech-specific noise and the visual component was blurred.

**Table 1.** Linear Mixed-effects Model of the Response Amplitude

| *Fixed Effects* | *Estimate* | *SE* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| Baseline | 183.1 | 24.8 | 33.7 | *7.4* | $10^{-8}$ |
| **Auditory noise (An)** | **−109.6** | **13.5** | **188** | **−8.1** | $\mathbf{10^{-13}}$ |
| **Posterior location × An** | **140.6** | **21.2** | **188** | **6.6** | $\mathbf{10^{-10}}$ |
| **Posterior location** | **−101** | **38.7** | **34.2** | **−2.6** | **.01** |
| Visual noise (Vn) | 21.6 | 13.5 | 188 | 1.6 | .11 |
| An × Vn | −13.3 | 19.1 | 188 | −0.7 | .49 |
| Posterior location × Vn | −8.9 | 21.2 | 188 | −0.4 | .67 |
| Posterior location × An × Vn | 3.6 | 29.9 | 188 | 0.1 | .91 |

Results of an LME model of the response amplitude. The fixed effects were the location of each electrode (Anterior vs. Posterior), the presence or absence of auditory noise (An) in the stimulus and the presence or absence of visual noise (Vn) in the stimulus. Electrodes and stimulus exemplar were included in the model as random factors. For each effect, the model estimates (in units of percent signal change) for that factor are shown relative to baseline, the response in anterior electrodes to clear audiovisual speech (AV stimulus condition). The "*SE*" column shows the standard error of the estimate. The degrees of freedom ("*df*"), *t* value, and *p* value derived from the model were calculated according to the Satterthwaite approximation, as provided by the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2015). The baseline is shown first; all other effects are ranked by absolute *t* value. Significant effects are shown in **bold**. The significance of the baseline fixed effect is grayed-out because it was prespecified: only electrodes with significant amplitudes were included in the analysis.

two larger effects: the main effect of auditory noise ($p = 10^{-14}$) and the interaction between auditory noise and the location of the electrode ($p = 10^{-10}$). Speech with clear auditory components evoked a larger response than speech with noisy auditory components (Clear A, consisting of the average of the AV and AVn conditions, 151 ± 27% vs. Noisy A, consisting of the average of the AnV and AnVn conditions, 93 ± 14%, mean ± *SEM* across electrodes) driving the main effect of auditory noise. However, the response patterns were very different in anterior and posterior electrodes, leading to the significant interaction in the LME model (Figure 2A). Speech with clear auditory components evoked a larger response than speech with noisy auditory component in anterior electrodes (Clear A vs. Noisy A: 194 ± 39% vs. 78 ± 16%, mean ± *SEM* across electrodes) but speech with clear auditory components evoked a smaller response than speech with noisy auditory component in posterior electrodes (88 ± 23% vs. 115 ± 25%).

To determine if the interaction between electrode location and the response to auditory noise was consistent, we plotted the amplitude of the response to Clear A versus Noisy A for all electrodes using one symbol per electrode (Figure 2B). All of the anterior electrodes lay above the line of equality, indicating uniformly larger responses for Clear A, and all of the posterior electrodes lay on or below the line of equality, indicating similar responses for Clear A and Noisy A.

To examine the interaction between location and auditory noise in a single participant, we examined two electrodes: an anterior electrode located just anterior to the A–P boundary and an adjacent electrode located 10 mm more posterior, just across the anterior–posterior boundary (Figure 2C and D). In the anterior electrode, the response to Clear A speech was much larger than the response to Noisy A speech (Clear A vs. Noisy A: 461 ± 35% vs. 273 ± 21%, mean across trials ± *SEM*; unpaired *t* test across trials: $t(147) = 4.6$, $p = 10^{-6}$), whereas in the adjacent posterior electrode, the response to Clear A speech was similar to the response to Noisy A speech (Clear A vs. Noisy A: 313 ± 21% vs. 349 ± 18%, $t(147) = 1.3$, $p = .2$). Hence, two electrodes located on either side of the anterior–posterior boundary showed very different patterns of responses to Clear A and Noisy A speech.

To examine the effect of anatomical location on the response to Clear A and Noisy A speech in more detail, we calculated each electrode's location in a reference frame defined by the STG (Figure 2E) and the difference in the electrode's response amplitude to Clear A and Noisy A speech (Clear A − Noisy A). First, we examined electrodes sorted by their medial-to-lateral position on the STG and observed no discernible pattern (Figure 2F). Second, we examined electrodes sorted by their anterior-to-posterior position on the STG (Figure 2G). Anterior electrodes showed uniformly positive values for Clear A – Noisy A (Clear A) whereas posterior electrodes showed zero or negative values for Clear A – Noisy A. However, we did not observe a gradient of responses between more anterior and more posterior electrodes, suggesting a sharp transition across the A–P boundary rather than a gradual shift in response properties along the entire extent of the STG. To quantify this observation, we tested two simple models. In the discrete model, there was a sharp transition between response properties on either side of the A–P boundary; in the continuous model, there was a gradual change in response properties across the entire extent of the STG.

For the discrete model, we fit the amplitude versus location points with two constants ($y = b$; horizontal lines with a fixed mean and zero slope, one mean for the anterior electrodes and one for the posterior electrodes;

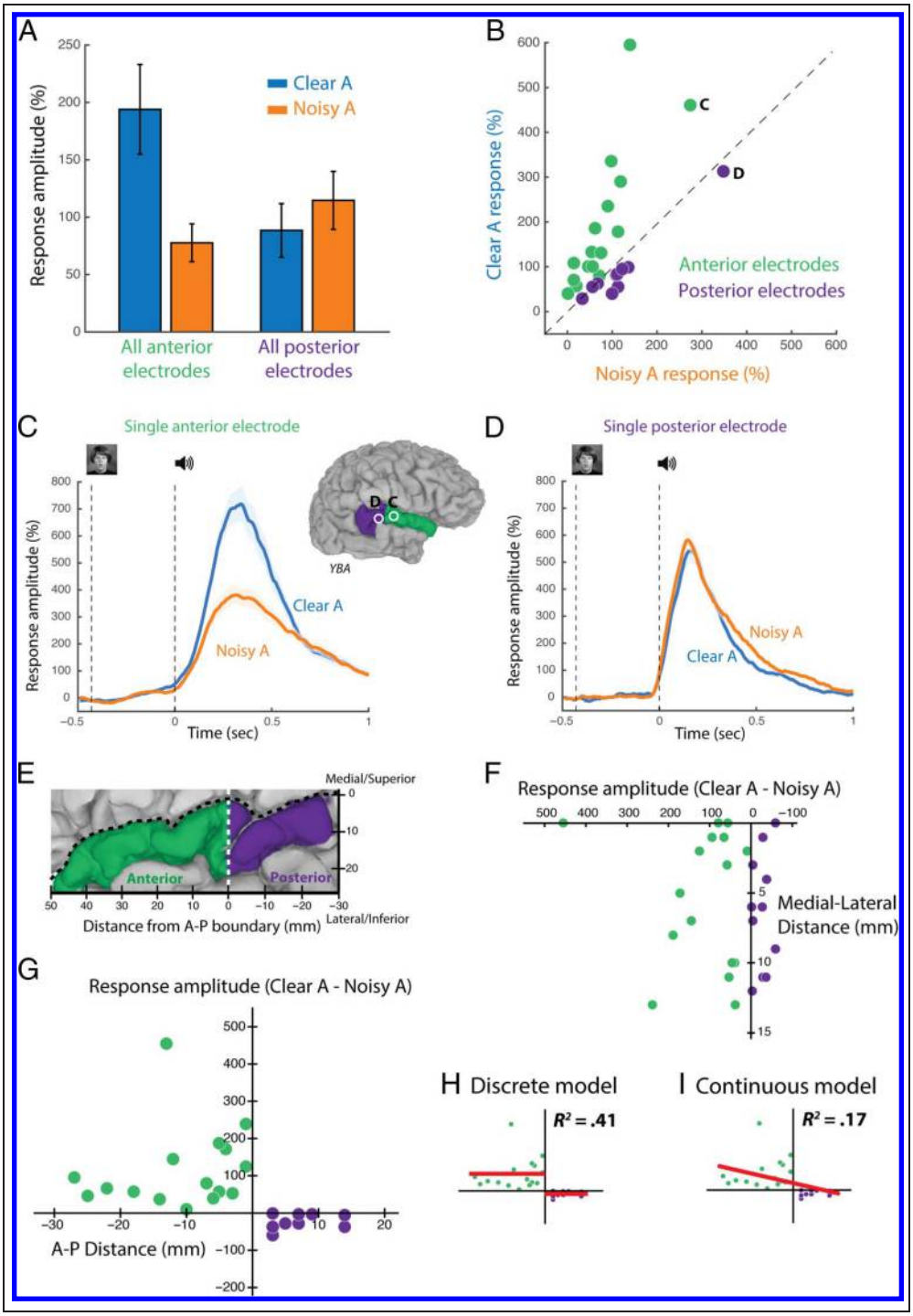


**Figure 2.** Response amplitudes. (A) The response to speech with clear auditory component (Clear A, combination of AV and AVn stimulus conditions) and noisy auditory component (Noisy A, combination of AnV and AnVn conditions) collapsed across electrodes (error bars show *SEM*). The response amplitude is the mean percent change in high-gamma power (70–110 Hz) in the 0–500 msec time window relative to prestimulus baseline (−500 to −100 msec). (B) The response to Clear A versus Noisy A speech for each individual electrode, with each anterior electrode shown as a green circle and each posterior electrode shown as a purple circle. The black dashed line represents the line of equality. Electrodes shown in C and D are labeled. (C) High-gamma response to Clear A speech (blue trace) and Noisy A speech (orange trace) for a single anterior electrode (labeled "C" in inset brain). Shaded regions indicate the *SEM* across trials. Black vertical dashed lines indicate visual and auditory stimulus onsets, respectively. (D) High-gamma response to Clear A and Noisy A speech in a single posterior electrode (labeled "D" in inset brain). (E) Coordinate system for STG measurements. *y* Axis indicates distance from medial/superior border of STG (black dashed line); *x* axis shows distance from the A–P boundary (white dashed line). (F) The response amplitude to Clear A speech minus the response amplitude to Noisy A speech as a function of distance from the medial/superior border, one symbol per electrode (anterior electrodes in green, posterior electrodes in purple). (G) The response amplitude to Clear A minus Noisy A speech as a function of distance from the A–P boundary. (H) Discrete model: Constant values were fit separately to the anterior and posterior electrode data in G ($y$ = a and $y$ = b), and the correlation with the data was calculated. (I) Continuous model: A linear model with two parameters was fit to both anterior and posterior electrodes ($y = mx + b$).

Figure 2H). For the continuous model, we fit the amplitude versus location points with a single line ($y = mx + b$; Figure 2I). Both models fit the data using an equal number of parameters (2). The two models were compared using $R^2$ as a measure of the explained variance and Akaike Information Criterion (AIC) as a measure of likelihood. The discrete model fit the amplitude versus location points much better than the continuous model ($R^2$ = .41 vs. .17), and the AIC revealed that the discrete model was more than 100 times more likely to explain the observed data ($e^{(\text{AIC continuous} - \text{AIC discrete})/2} = 102$).

To allow easier comparison of the A–P boundary with the functional neuroimaging literature, we converted each participant's brain into standard space and measured

the coordinates of each electrode. The average location in standard space of the Heschl's gyrus landmark, the boundary between the anterior and posterior STG ROIs, was $y = -27 \pm 2$ (mean across participants $\pm$ *SD*). The mean position in standard space of all anterior electrodes was ($x = \pm 66$, $y = -18$, $z = 6$), whereas for posterior electrodes the mean position was ($x = \pm 67$, $y = -34$, $z = 12$).

### Variability of the Responses to Clear and Noisy Speech

Theoretical models predict that combining the information available about speech content from the auditory and visual modalities should reduce neural variability (Fetsch et al., 2012; Knill & Pouget, 2004); see discussion and Figure 6 for more details. We hypothesized that the presence of noise in the speech stimulus might differentially affect the response variability in anterior and posterior electrodes. To test this hypothesis, we fit the same LME model used to examine response amplitude, except that response variability (CV) was used as the dependent measure. As shown in Table 2, there were three significant effects in the LME model, including an effect of electrode location ($p = .02$) driven by a larger overall response variability in posterior electrodes than in anterior electrodes (Anterior vs. Posterior: 0.85 $\pm$ 24% vs. 0.99 $\pm$ 0.1, mean CV averaged across all stimulus conditions $\pm$ *SEM*). The other two effects showed a larger effect size: the main effect of auditory noise ($p = 10^{-6}$) and the interaction between auditory noise and the location of the electrode ($p = 10^{-8}$).

Speech with Noisy A resulted in larger response variability than speech with Clear A (Clear A vs. Noisy A: 0.89 $\pm$ 0.06 vs. 0.93 $\pm$ 0.06, mean $\pm$ *SEM* across electrodes) driving the main effect of auditory noise in the model. However, the response patterns were very different in anterior and posterior electrodes, leading to the significant interaction (Figure 3A). Speech with noisy auditory component resulted in a larger response variability than speech with clear auditory component in anterior electrodes (Clear A vs. Noisy A: 0.73 $\pm$ 0.05 vs. 0.96 $\pm$ 0.1, mean $\pm$ *SEM* across electrodes) but speech with a noisy auditory component resulted in a smaller response variability than speech with a clear auditory component in posterior electrodes (Clear A vs. Noisy A: 1.1 $\pm$ 0.1 vs. 0.9 $\pm$ 0.1).

To determine if the interaction between electrode location and the response variability for auditory noise was consistent, we plotted the variability of the response to Clear A versus Noisy A for all electrodes using one symbol per electrode (Figure 3B). Most of the anterior electrodes lay below the line of equality, indicating larger variability for Noisy A, whereas most of the posterior electrodes lay above the line of equality, indicating smaller variability for noisy A.

To demonstrate the effect at the single electrode level, we examined the interaction between location and auditory noise in a single participant, we examined two electrodes: an anterior electrode and a posterior electrode (Figure 3C and D). Figure 3C shows the normalized responses for a single anterior electrode for single trials of speech with clear and noisy auditory components. In this anterior electrode, there was variability across trials in both conditions, but the variability was much greater for speech with a noisy auditory component than for speech with a clear auditory component (Clear A vs. Noisy A: 1.1 vs. 1.7, unpaired *t* test across normalized trial amplitudes: $t(221) = 5.4$, $p = 10^{-7}$). In a posterior electrode from the same participant (Figure 3B), the opposite pattern was observed: The variability was much greater for speech with a clear auditory component than for speech with a noisy auditory component (Clear A vs. Noisy A: 1.4 vs. 0.9, $t(221) = 5$, $p = 10^{-6}$). Hence, two electrodes located on either side of the anterior–posterior boundary showed very different patterns of response variability.

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

**Table 2.** Linear Mixed-effects Model of the Response Variability

| *Fixed Effects* | *Estimate* | *SE* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| **Baseline** | **0.76** | **0.1** | **29.8** | **8** | **$10^{-8}$** |
| **Posterior location × An** | **−0.59** | **0.1** | **179.9** | **−5.7** | **$10^{-7}$** |
| **Auditory noise (An)** | **0.31** | **0.07** | **180.4** | **4.6** | **$10^{-5}$** |
| **Posterior location** | **0.35** | **0.14** | **39.8** | **2.5** | **.02** |
| Posterior location × Vn | −0.13 | 0.1 | 179.5 | −1.3 | .2 |
| Posterior location × An × Vn | 0.15 | 0.15 | 179.5 | 1 | .31 |
| An × Vn | 0.03 | 0.09 | 179.6 | 0.3 | .77 |
| Visual noise (Vn) | 0.01 | 0.06 | 179.5 | 0.1 | .89 |

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

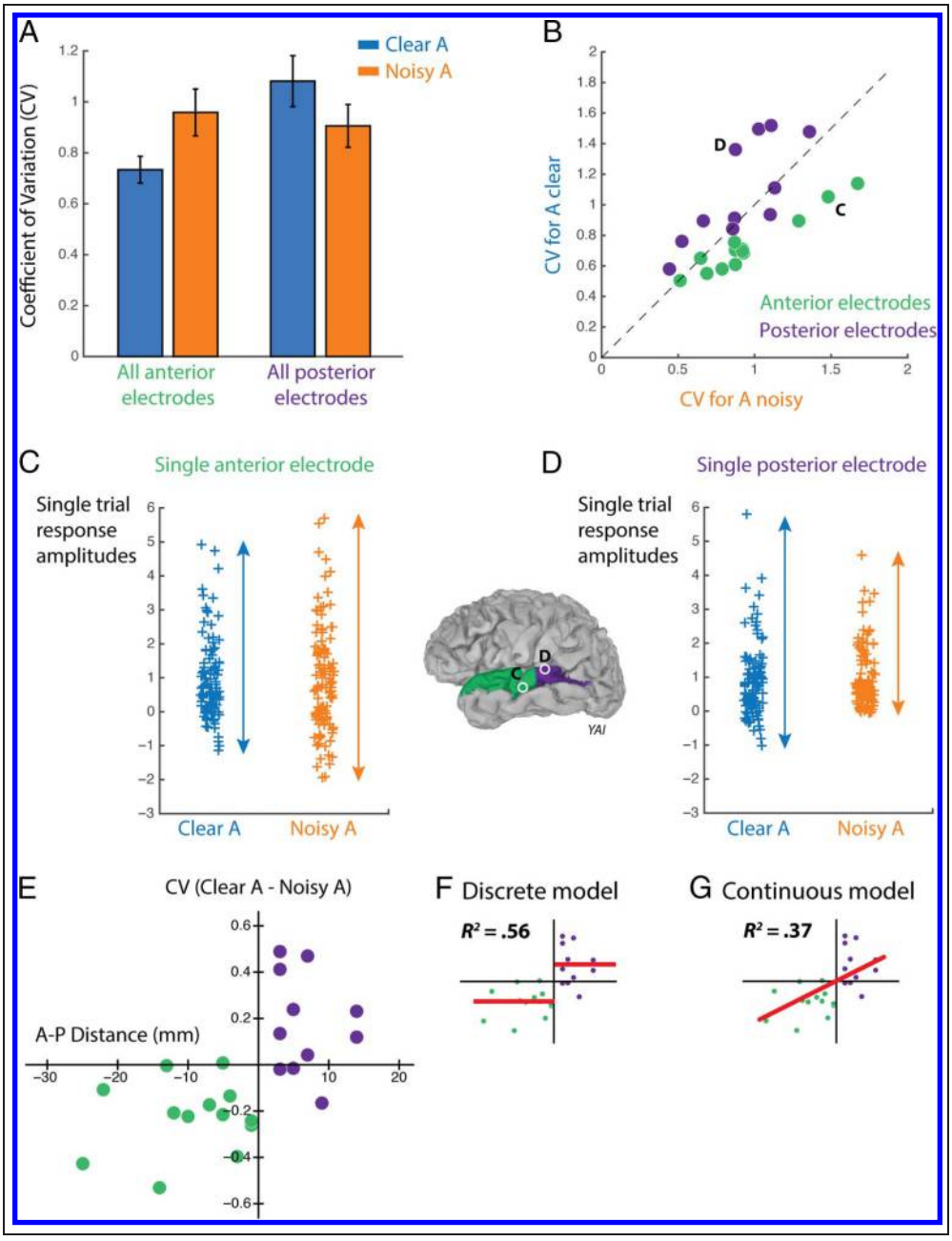


**Figure 3.** Response variability. (A) Response variability to speech with clear auditory component (Clear A, combination of AV and AVn stimulus conditions) and noisy auditory component (Noisy A, combination of AnV and AnVn conditions) collapsed across electrodes (error bars show *SEM*). The response variability was measured as the CV, defined as the standard deviation of the high-gamma response divided by the mean of the high-gamma response; this measure accounts for the differences in the mean response between conditions shown in Figure 2. (B) The response variability to Clear A versus Noisy A speech for each individual electrode, with each anterior electrode shown as a green circle and each posterior electrode shown as a purple circle. The black dashed line represents the line of equality. Electrodes shown in C and D are labeled. (C) High-gamma response amplitudes to single presentations of Clear A speech (blue symbols) and Noisy A speech (orange symbols) for a single anterior electrode (labeled "C" in inset brain), normalized by the mean response across trials (value of one indicates a single trial response equal to the mean response across trials). Arrows illustrate CV, a measure of variability. (D) High-gamma response amplitudes to single presentations of speech for a single posterior electrode (labeled "D" in inset brain).

Clear A − CV for Noisy A) and plotted it against that electrode's A–P location on the STG (Figure 3E). Paralleling the analysis performed on response amplitude, discrete and continuous models were fit to the data (Figure 3F and G). The discrete model fit the amplitude versus location points much better than the continuous model ($R^2$ = .56 vs. .37) and the AIC revealed that the discrete model was more likely to explain the observed data ($e^{(\text{AIC continuous} - \text{AIC discrete})/2} = 74$). Hence, the difference in response variability between electrodes is more accurately described as arising from two groups (Anterior and Posterior) with categorically different variability rather than as a continuous change in variability from anterior to posterior.

## Timing of the Responses to Clear and Noisy Speech

The high temporal resolution of electrocorticography allows for examination of the detailed timing of the neuronal responses. Figure 4 (A and B) show the average response of anterior and posterior electrodes to Clear A and Noisy A speech. In anterior electrodes, the high-gamma response to Clear A speech started at 77 msec after auditory stimulus onset, reached half-maximum amplitude at 110 msec, peaked at 210 msec, and returned to the half-maximum value at 290 msec, resulting in a total response duration (measured as the FWHM) of 190 msec.

To determine the effects of auditory noise and electrode location on the timing of the neuronal response,

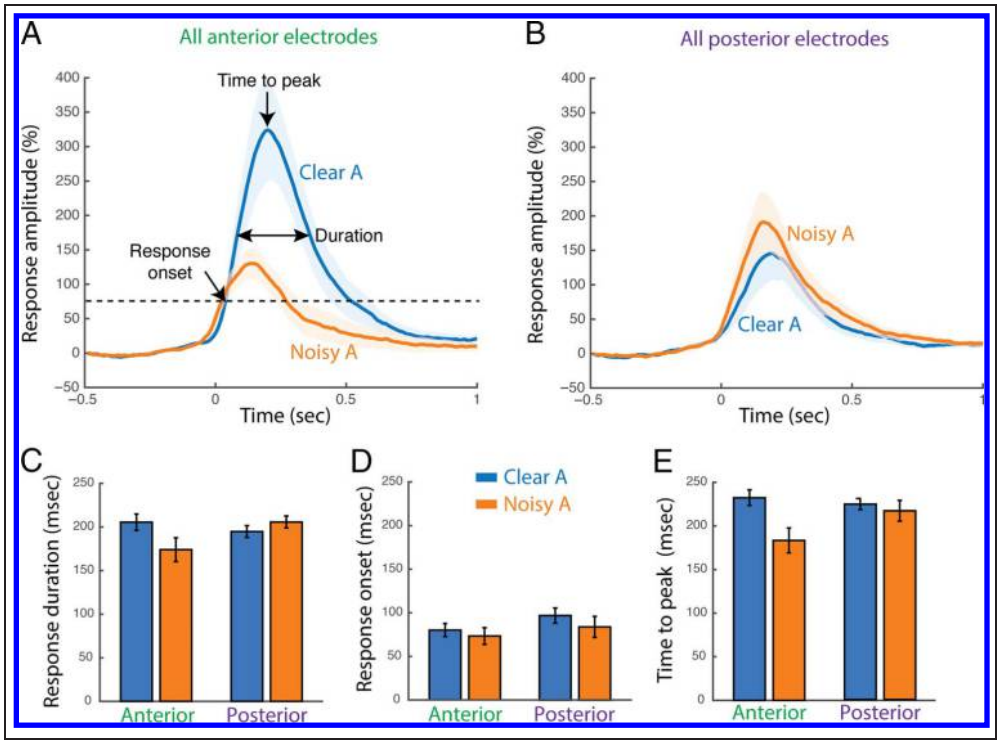


**Figure 4.** Response timing. (A) High-gamma response amplitudes to Clear A and Noisy A speech averaged across all anterior electrodes, shown as percent signal change from baseline relative to time from auditory stimulus onset (error bars show *SEM*). Three measures of the response were calculated. Response onset time is the first time point at which the signal deviates three standard deviations from baseline. Time to peak is the time point of maximal response amplitude. Duration indicates the time between the first and last time points at which the response is equal to half of its maximum value (FWHM). (B) High-gamma response amplitudes to Clear A and Noisy A speech averaged across all posterior electrodes. (C) The response duration for Clear A versus Noisy A speech in anterior electrodes (left) and posterior electrodes (right). Error bars show *SEM*. (D) The response onset in anterior and posterior electrodes. (E) The time to peak in anterior and posterior electrodes.

for each electrode we estimated response duration, onset time, and time-to-peak and separately fit three LME models with each temporal variable as the dependent measure. For the LME model with response duration as the dependent measure (Table 3 and Figure 4C) the only significant effects were the main effect of auditory noise ($p = 10^{-5}$) and the interaction between auditory noise and electrode location ($p = 10^{-5}$). These effects were driven by an overall longer response duration for Clear A speech than for Noisy A speech (Clear A vs. Noisy A: 194 ± 6 msec vs. 187 ± 9 msec, mean across electrodes ± *SEM*), with anterior electrodes showing longer responses for Clear A speech (Clear A vs. Noisy A: 205 ± 9 msec vs. 174 ± 14 msec) and posterior electrodes showing shorter responses for Clear A speech (Clear A vs. Noisy A: 195 ± 7 msec vs. 206 ± 7 msec).

For the LME model with response onset as the dependent measure, there were no significant main effects or

**Table 3.** Linear Mixed-effects Model of the Response Duration

| *Fixed Effects* | *Estimate* | *SE* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| **Baseline** | **206.2** | **9.6** | **41.4** | **21.4** | **$10^{-16}$** |
| **Posterior location × An** | **48.6** | **10.9** | **189** | **4.4** | **$10^{-5}$** |
| **Auditory noise (An)** | **−30.9** | **7** | **189** | **−4.4** | **$10^{-5}$** |
| Posterior location | −15.1 | 15.1 | 41.4 | −1 | .32 |
| Posterior location × Vn | 8.9 | 10.9 | 189 | 0.8 | .42 |
| Posterior location × An × Vn | −12.2 | 15.5 | 189 | −0.8 | .43 |
| Visual noise (Vn) | −1.4 | 7 | 189 | −0.2 | .84 |
| An × Vn | −1.3 | 9.9 | 189 | −0.1 | .89 |

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

**Table 4.** Linear Mixed-effects Model of the Response Onset

| *Fixed Effects* | *Estimate* | *SE* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| **Baseline** | **81.5** | **9.2** | **27.6** | **8.8** | **$10^{-9}$** |
| Posterior location | 17.6 | 13.6 | 41.3 | 1.3 | .2 |
| Posterior location × An | −9.1 | 9.8 | 187.9 | −0.9 | .35 |
| An × Vn | −7.1 | 8.8 | 187.9 | −0.8 | .42 |
| Auditory noise (An) | −2.6 | 6.3 | 187.9 | −0.4 | .68 |
| Visual noise (Vn) | −2.6 | 6.3 | 187.9 | −0.4 | .68 |
| Posterior location × An × Vn | 5 | 13.9 | 187.9 | 0.4 | .72 |
| Posterior location × Vn | −1.3 | 9.8 | 187.9 | −0.1 | .9 |

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

interactions (Table 4 and Figure 4D). For the LME model with time-to-peak as the dependent measure (Table 5 and Figure 4E), there was a significant main effect of auditory noise ($p = 10^{-9}$) and an interaction between auditory noise and electrode location ($p = 10^{-4}$) driven by a longer time-to-peak for Clear A speech (Clear A vs. Noisy A: 229 ± 6 msec vs. 197 ± 10 msec, mean across electrodes ± *SEM*), more so in anterior electrodes (Clear A vs. Noisy A: 232 ± 9 msec vs. 183 ± 14 msec) than posterior electrodes (Clear A vs. Noisy A: 224 ± 6 msec vs. 216 ± 12 msec).

## Relationship between Neuronal Responses and Perceptual Accuracy

Participants performed a task that required them to respond to the identity of the word present in each trial. Across participants, only AnVn trials consistently generated enough errors to compare correct and incorrect trials (AV: 99 ± 3%, AVn: 98 ± 3%, AnV: 81 ± 20%, AnVn: 63 ± 15%; % correct, mean across participants ± *SD*). To determine the relationship between neuronal response amplitude and behavioral accuracy within AnVn trials, an LME model was constructed with response amplitude as the dependent measure, electrode location (Anterior vs. Posterior), and behavioral accuracy (Correct vs. Incorrect) as fixed factors and stimulus exemplar, participant, and electrode (nested within participant) as random factors (Table 6). In the LME model, the only significant effect was an interaction between electrode location and behavioral accuracy ($p = .01$) driven by smaller amplitudes in correct trials for anterior electrodes (Correct vs. Incorrect: 84 ± 15% vs. 93 ± 20%, mean gamma power signal change relative to baseline across electrodes ± *SEM*) but larger amplitudes in correct trials for posterior electrodes (Correct vs. Incorrect: 122 ± 27% vs. 106 ± 26%). A similar model with CV as the dependent measure did not show any significant effects (Table 7).

## Potential Confound: Intelligibility

We observed very different neuronal responses to audiovisual speech with noisy auditory component in anterior compared with posterior electrodes, attributing this difference to the differential contributions of anterior and posterior STG to multisensory integration. However, we used only high levels of auditory noise in our audiovisual speech stimuli. To determine how the level of auditory noise influenced the effect, in one patient we presented audiovisual speech with 11 different levels of auditory noise and examined the neural response in two electrodes located on either side of the anterior–posterior boundary (Figure 5A). First, we examined how these data compared with our previous results by collapsing the 11 different levels of noise into just two categories "low noise" (0–40% noise levels) and "high noise" (50–100% noise levels)

**Table 5.** Linear Mixed-effects Model of the Response Peak Time

| *Fixed Effects* | *Estimate* | *SE* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| **Baseline** | **234.3** | **10.4** | **36** | **22.6** | **$10^{-16}$** |
| **Auditory noise (An)** | **−46.5** | **7.4** | **187.9** | **−6.3** | **$10^{-9}$** |
| **Posterior location × An** | **45.5** | **11.5** | **187.9** | **3.9** | **$10^{-4}$** |
| Posterior location | −12.5 | 15.8 | 41.6 | −0.8 | .44 |
| Posterior location × Vn | 8.7 | 11.5 | 187.9 | 0.8 | .45 |
| Visual noise (Vn) | −3.9 | 7.4 | 187.9 | −0.5 | .6 |
| Posterior location × An × Vn | −8.4 | 16.3 | 187.9 | −0.5 | .61 |
| An × Vn | −4.9 | 10.4 | 187.9 | −0.5 | .64 |

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

**Table 6.** Linear Mixed-effects Model of the Effect of Accuracy on Response Amplitude

| *Fixed Effects* | *Estimate* | *SE* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| **Baseline** | **105.2** | **36.1** | **4.2** | **2.9** | **.04** |
| **Incorrect responses × Posterior location** | **−25.6** | **10.2** | **65.8** | **−2.5** | **.01** |
| Incorrect responses | 11.3 | 6.6 | 66.1 | 1.7 | .09 |
| Posterior location | 19.6 | 21.8 | 22.8 | 0.9 | .38 |

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

similar to our initial analysis of Clear A and Noisy A audiovisual speech (Figure 5B). The responses were similar to that observed with just two levels of noise (compare Figure 5B and Figure 2A). An LME model fit to the data across the different noise levels (Table 8) showed significant effects of noise level ($p = 10^{-16}$), electrode location ($p = 10^{-16}$), and an interaction between noise level and location ($p = 10^{-16}$), driven by significantly greater response in anterior electrodes to low noise stimuli (Low vs. High: 248 ± 13% vs. 124 ± 8%, mean across trials ± *SEM*) and similar responses in posterior electrodes to low and high noise conditions (Low vs. High: 95 ± 5% vs. 115 ± 5%). Next, we examined the response to each different level of auditory noise. In the anterior electrode, increasing levels of auditory noise led to smaller responses, whereas in the posterior electrode, increasing levels of auditory noise led to similar or slightly larger gamma band responses (Figure 5C). We quantified this by fitting a line to the anterior and posterior electrode responses at 11 different auditory noise levels. The anterior electrode fit was significant ($R^2 = .9, p = 10^{-6}$) with a negative slope ($m = -24$), whereas the posterior electrode fit was not significant ($R^2 = .07, p = .4$) with a slightly positive slope ($m = 1.32$).

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

## DISCUSSION

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

These data are best understood within the framework of Bayes optimal models of multisensory integration (Alais & Burr, 2004; Ernst & Banks, 2002) and speech perception (Bejjanki, Clayards, Knill, & Aslin, 2011; Ma et al., 2009). In these models, different sensory modalities are posited to contain independent sources of environmental and sensory noise. Because of the independence of noise sources across modality, Bayesian integration results in a multisensory representation that has smaller variance than either of the unisensory variances (Fetsch et al., 2012; Knill & Pouget, 2004).

Recently, a Bayesian model of causal inference in audiovisual speech perception was proposed (Magnotti & Beauchamp, 2017). Figure 6 shows an application of this model to our data. We assume that anterior STG contains

**Table 7.** Linear Mixed-effects Model of the Effect of Accuracy on Response Variability

| *Fixed Effects* | *Estimate* | *SE* | *df* | *t* | *p* |
|---|---|---|---|---|---|
| **Baseline** | **1** | **0.19** | **7** | **5.4** | **$10^{-3}$** |
| Posterior location | −0.13 | 0.2 | 31.3 | 0.6 | .53 |
| Incorrect responses | 0.02 | 0.11 | 67.1 | 0.2 | .86 |
| Incorrect responses × Posterior location | 0.01 | 0.17 | 66.5 | 0.1 | .95 |

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

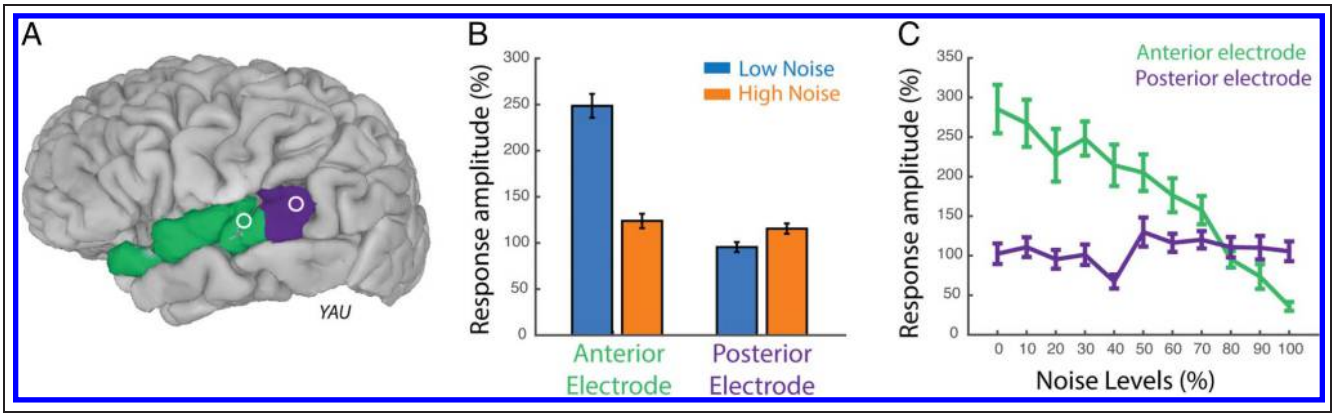


**Figure 5.** Response amplitude with varying levels of auditory noise. (A) The location of an anterior and a posterior electrode in a single participant. (B) The response amplitude in the anterior electrode (left bars) and posterior electrode (right bars) to audiovisual speech with low levels of auditory noise (Low Noise: 0% to 40%) and high levels of auditory noise (High Noise: 50% to 100%) averaged across trials (error bars show *SEM*). (C) Response amplitude for the anterior and posterior electrodes at each of 11 different auditory noise levels (0–100%) averaged across trials (error bars show *SEM*).

a unisensory representation of auditory speech, that extrastriate visual areas contain a representation of visual speech, and that posterior STG contains a representation of multisensory speech formed by integrating inputs from unisensory auditory and visual areas (Bernstein & Liebenthal, 2014; Nath & Beauchamp, 2011). The neural implementation of Bayes optimal integration is thought to rely on probabilistic population codes (Angelaki, Gu, & DeAngelis, 2009; Ma et al., 2006) in which pools of neurons encode individual stimuli in a probabilistic fashion. These population codes are modeled as Gaussians in which amplitude and variability are inversely related. A smaller, more focal Gaussian indicates larger amplitude and less variability in the population code, whereas a larger Gaussian indicates smaller amplitude and more variability.

For audiovisual speech with a clear auditory component (Clear A), the neural population code in anterior STG has a given amplitude and variability. When auditory noise is added (Noisy A), the population code amplitude decreases and the variability increases (Ma et al., 2006), an accurate description of the response in anterior STG for noisy compared with clear auditory speech.

For the visual representation in lateral extrastriate cortex, the visual information is the same in the Clear A and Noisy A conditions, predicting similar population codes for both conditions (Figure 6B). For the multisensory representation in posterior STG, the population code is calculated as the optimal integration of the response in auditory and visual representations. The visual information serves to compensate for the increased auditory noise in the Noisy A condition, so that the population code for the integrated representation is only slightly broader for Noisy A than Clear A speech, a match to the observation that the amplitude and variability of the response to Noisy A and Clear A speech are much more similar in posterior STG than they are in anterior STG.

A close inspection of the data shows that, contrary to Bayesian models, the response in posterior STG was slightly more focal (30% greater amplitude and 16% reduced variability) for Noisy A compared with Clear A conditions. Although counterintuitive, this result is consistent with evidence that visual cortex responds more to noisy than clear audiovisual speech (Schepers, Yoshor, & Beauchamp, 2015). This enhancement may be attributable to top–down modulation from higher-level areas

**Table 8.** Linear Model of the Effect of Varying Auditory Noise Levels on Response Amplitude

| *Fixed Effects* | *Estimate* | *SE* | *t* | *p* |
|---|---|---|---|---|
| **Baseline** | **248.5** | **8.6** | *28.9* | $10^{-16}$ |
| **Posterior location** | **−153.1** | **12.2** | **−12.6** | $\mathbf{10^{-16}}$ |
| **High auditory noise** | **−124.8** | **11.6** | **−10.7** | $\mathbf{10^{-16}}$ |
| **High auditory noise × Posterior location** | **144.8** | **16.5** | **8.8** | $\mathbf{10^{-16}}$ |

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

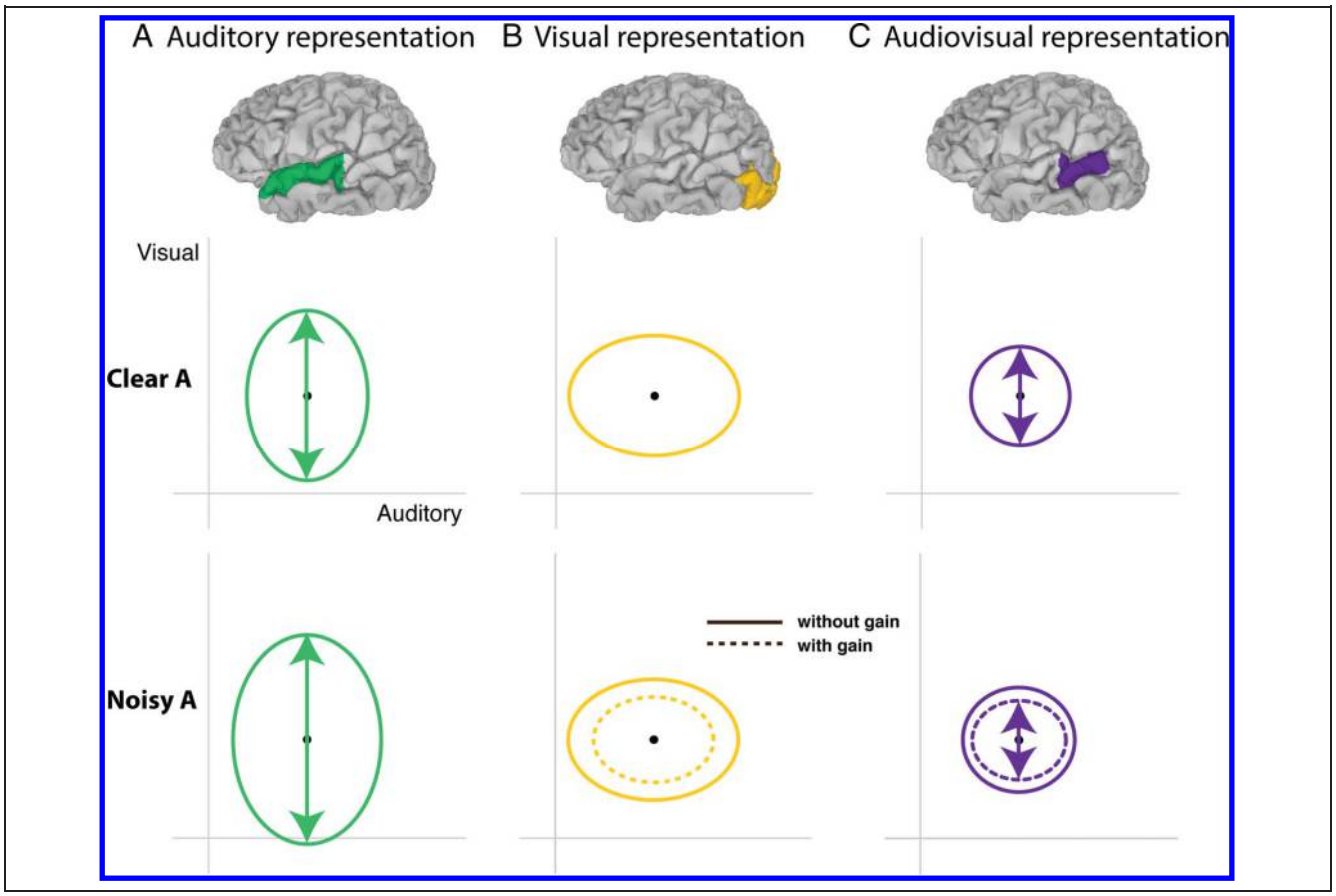


**Figure 6.** Bayesian model of audiovisual speech with auditory noise. (A) The model assumes that a neural representation of the auditory component of audiovisual speech exists in anterior STG (top row: brain region colored green). The high-dimensional neural representation is projected onto a 2-D space (middle and bottom rows) in which the *x* axis represents auditory feature information and the *y* axis represents visual feature information. The stimulus representation is shown as an ellipse indicating the cross-trial variability in representation of an identical physical stimulus due to sensory noise. For audiovisual speech with clear auditory component (Clear A) in anterior STG (green ellipse in middle row), there is less variability along the auditory axis and more variability along the visual axis, indicated by the shape of the ellipse. For audiovisual speech with noisy auditory component (Noisy A) in anterior STG (green ellipse in bottom row), there is greater variability along both axes due to the added stimulus noise (see Methods for details). (B) The model assumes that a neural representation of the visual component of audiovisual speech exists in lateral extrastriate visual cortex (top row: brain region colored yellow). In the visual representation, there is less variability along the visual axis and more variability along the auditory axis, indicated by the shape of the ellipse. For audiovisual speech with noisy auditory component (Noisy A), the visual component of the speech is identical, so the representation should be identical (yellow ellipse in bottom row). However, evidence from Schepers et al. (2015) demonstrates that response in visual cortex to Noisy A speech is actually greater than to Clear A speech, suggesting an increase in gain due to attentional modulation or other top–down factors. The representation with gain modulation is shown with the dashed yellow ellipse. (C) The model assumes that a neural representation that integrates both auditory and visual components of audiovisual speech exists in posterior STG (top row: brain region colored purple). Because of the principles of Bayesian integration, this representation has smaller variability than either the auditory representation or the visual representation (compare size of purple ellipse in each row with green and yellow ellipses). Assuming gain modulation, the integrated representation of Noisy A speech (dashed purple ellipse in bottom row) has smaller variability than the representation of Clear A speech (purple ellipse in middle row).

that increase the gain in visual cortex, similar to attentional modulation in which representations in visual cortex are heightened and/or sharpened by spatial or featural attention (Maunsell & Treue, 2006; Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999). This gain increase would be adaptive because it would increase the likelihood of decoding speech from visual cortex under conditions of low or no auditory information, at the cost of additional deployment of attentional and neural resources. We implemented this gain modulation in our Bayesian model as reduced variance in the visual representation for Noisy A compared with Clear A speech. When this visual representation with reduced variance is integrated with the noisy auditory representation, the resulting multisensory representation becomes more focal for Noisy A than Clear A speech, a fit to the observed increased amplitude and reduced variability for Noisy A compared with Clear A speech in posterior STG.

Although the Bayesian model provides a conceptual framework for understanding how multisensory integration could affect the amplitude and variance of neuronal population responses, it is agnostic about the actual stimulus features important for integration. We did not

observe a main effect of visual noise (or an interaction between visual noise and auditory noise) LME on amplitude and variance (Tables 1 and 2). Most of the relevant information provided by the visual signal during auditory–visual speech perception is related to the timing of mouth opening and closing relative to auditory speech. The blurring procedure used to generate the noisy visual speech may leave this timing information intact.

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

An interaction between electrode location and response amplitude was also observed in an analysis of perceptual accuracy (only speech with both noisy auditory and noisy visual component generates enough errors for this analysis). In anterior electrodes, responses were larger for incorrect trials, whereas in posterior electrode responses were larger for correct trials. This supports the idea that posterior electrodes are particularly important in the perception of noisy speech, with larger amplitude indicating a more focal peak of activity in the population code and less uncertainty about the presented stimulus.

### Anterior versus Posterior Anatomical Specialization

There was a strikingly sharp boundary between the anterior and posterior response patterns, suggesting that anterior and posterior STG are functionally distinct. We divided STG at the posterior border of Heschl's gyrus (mean $y = -27$), a landmark that also has been used in previous neuroimaging studies of speech processing (Okada et al., 2010; Specht & Reul, 2003; Hickok & Poeppel, 2000). A functional division in STG near Heschl's gyrus is consistent with the division of the auditory system into two processing streams, one of which runs anterior-ventral from Heschl's gyrus and one of which runs posterior-dorsal (Pickles, 2015; Rauschecker, 2015). These two streams are often characterized as specialized for processing "what" or object identity features (anterior-ventral) and "where" or object location features (posterior-dorsal) by analogy with the different streams of visual processing (Mishkin & Ungerleider, 1982). However, these labels do not neatly map onto an anterior preference for clear speech and a posterior preference for noisy speech (Leonard & Chang, 2014) and may reflect preferences for different rates of spectrotemporal modulation (Hullett, Hamilton, Mesgarani, Schreiner, & Chang, 2016).

Although we are not aware of previous studies examining changes in the neural variability to Clear A and Noisy A audiovisual speech, a number of neuroimaging studies have reported A–P differences in the amplitude of the neural response to Clear A and Noisy A audiovisual speech. Stevenson and James (2009) presented clear audiovisual speech and audiovisual speech with noise added to both modalities (noisy auditory + noisy visual), contrasting both against a standard baseline condition consisting of simple visual fixation. Anterior regions of STG/STS showed greater responses to clear than noisy audiovisual speech (Figure 5C and Table 3 in their paper, $y = -20$ compared with $y = -18$ in this study, mean across left and right hemispheres). Their results slightly differ from ours because they showed that posterior STG/STS (Figure 5D and Table 1 in their paper, $y = -37$, compared with $y = -34$ in this study) displays relatively weak responses to moderately noisy speech. This could be explained by their use of noisy auditory + noisy visual speech versus our use of noisy auditory + clear visual speech: If posterior regions respond to both auditory and visual speech information, degraded visual information might be expected to reduce response amplitudes in posterior regions.

Consistent with these results, Lee and Noppeney (2011) found that anterior STG/STS ($y = -16$, their Table 2) showed significant audiovisual interactions only for clear speech, whereas posterior STG/STS (mean $y = -36$, Table 2 in their paper) showed interactions for both clear and noisy audiovisual speech.

Bishop and Miller (2009) reported greater responses to clear versus noisy audiovisual speech in anterior regions of STG (Table 1 in their paper, $y = -13$ mean across left and right hemispheres), whereas McGettigan and colleagues (2012) reported greater responses for clear than noisy audiovisual speech in both anterior STG ($y = -12$, Table 1 in their paper) and posterior STG ($y = -42$).

Although most neuroimaging studies have reported greater responses to clear than noisy audiovisual speech, two studies have reported the opposite result of greater responses to noisy speech in the STG (Callan et al., 2003; Sekiyama, Kanno, Miura, & Sugita, 2003). However, the interpretation of these studies is complex. Sekiyama and colleagues tested clear and noisy speech consisting of incongruent audiovisual speech (including McGurk syllables) that are known to evoke responses in STS that are both different from congruent syllables and vary markedly from participant to participant (Erickson et al., 2014; Nath & Beauchamp, 2012). Callan and colleagues performed an analysis in which they first subtracted the response to auditory-only clear speech from the response to audiovisual clear speech, then subtracted

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

The idea of an A–P double dissociation is also generally supported by the neuroimaging literature examining brain responses to clear and noisy auditory-only speech, although the many differences in the stimulus materials, task manipulations, and data analysis strategies makes direct comparisons difficult. Obleser, Zimmermann, Van Meter, and Rauschecker (2007) reported a double dissociation, with posterior regions ($y = -26$, Table 1 in their paper) preferring noisy speech to clear speech, whereas anterior regions ($y = -18$) preferred clear speech to noisy speech. A double dissociation was also reported by Du, Buchsbaum, Grady, and Alain (2014): Anterior regions of STG ($y = -15$, Table S2 in their paper) showed greater BOLD amplitude with less auditory noise, whereas posterior regions ($y = -32$) showed greater BOLD amplitude with more auditory noise. Similarly, Wild, Davis, and Johnsrude (2012) found that anterior regions of STG ($y = -12$, Table 1 in their paper) preferred clear to noisy speech, whereas posterior regions ($y = -30$) preferred noisy speech to clear speech.

Single dissociations consistent with an anterior preference for clear speech are also common in the literature. Scott, Blank, Rosen, and Wise (2000) found that anterior regions ($y = -12$) showed greater response amplitudes for clear speech, whereas posterior regions ($y = -38$, Figure 2A in their paper) showed similar response amplitudes. Giraud and colleagues (2004) also reported greater response amplitudes for clear than noisy speech in anterior STG (Table 1 in their paper, $y = -4$ mean across left and right hemispheres) but not posterior STG.

Reprint requests should be sent to Michael S. Beauchamp, Department of Neurosurgery and Core for Advanced MRI, Baylor College of Medicine, 1 Baylor Plaza, S104, Houston, TX 77030, or via e-mail: michael.beauchamp@bcm.edu.

## REFERENCES

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14,* 257–262.

Angelaki, D. E., Gu, Y., & DeAngelis, G. C. (2009). Multisensory integration: Psychophysics, neurophysiology, and computation. *Current Opinion in Neurobiology, 19,* 452–458.

Argall, B. D., Saad, Z. S., & Beauchamp, M. S. (2006). Simplified intersubject averaging on the cortical surface using SUMA. *Human Brain Mapping, 27,* 14–27.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv preprint arXiv:1406.5823.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41,* 809–823.

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audiovisual speech perception. *PLoS One, 6,* e19812.

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44,* 5–18.

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

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., et al. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10,* 512–528.

Bishop, C. W., & Miller, L. M. (2009). A multisensory cortical network for understanding speech in noise. *Journal of Cognitive Neuroscience, 21,* 1790–1805.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport, 14,* 2213–2218.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649–657.

Churchland, M. M., Yu, B. M., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., et al. (2010). Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nature Neuroscience, 13,* 369–378.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers in Biomedical Research, 29,* 162–173.

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

Desikan, R. S., Segonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage, 31,* 968–980.

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage, 53,* 1–15.

Du, Y., Buchsbaum, B. R., Grady, C. L., & Alain, C. (2014). Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proceedings of the National Academy of Sciences, U.S.A., 111,* 7126–7131.

Erickson, L. C., Zielinski, B. A., Zielinski, J. E., Liu, G., Turkeltaub, P. E., Leaver, A. M., et al. (2014). Distinct cortical locations for integration of audiovisual speech and the McGurk effect. *Frontiers in Psychology, 5,* 534.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415,* 429–433.

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience, 15,* 146–154.

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

Foxe, J. J., Wylie, G. R., Martinez, A., Schroeder, C. E., Javitt, D. C., Guilfoyle, D., et al. (2002). Auditory-somatosensory multisensory processing in auditory association cortex: An fMRI study. *Journal of Neurophysiology, 88,* 540–543.

Giraud, A. L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M. O., Preibisch, C., et al. (2004). Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cerebral Cortex, 14,* 247–255.

Gur, M., Beylin, A., & Snodderly, D. M. (1997). Response variability of neurons in primary visual cortex (V1) of alert monkeys. *Journal of Neuroscience, 17,* 2914–2920.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4,* 131–138.

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. *Journal of Neuroscience, 36,* 2014–2026.

Jacques, C., Witthoft, N., Weiner, K. S., Foster, B. L., Rangarajan, V., Hermes, D., et al. (2016). Corresponding ECoG and fMRI category-selective signals in human ventral temporal cortex. *Neuropsychologia, 83,* 14–28.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron, 22,* 751–761.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27,* 712–719.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). *Package "lmerTest"*. R package version, 2.0-29.

Lee, H., & Noppeney, U. (2011). Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension. *The Journal of Neuroscience, 31,* 11338–11350.

Leonard, M. K., & Chang, E. F. (2014). Dynamic speech representations in the human temporal lobe. *Trends in Cognitive Sciences, 18,* 472–479.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9,* 1432–1438.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PLoS One, 4,* e4638.

Magnotti, J. F., & Beauchamp, M. S. (2017). Causal inference explains perception of the McGurk effect and other incongruent audiovisual speech. *PLoS Computational Biology, 13,* e1005229.

Maunsell, J. H. R., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences, 29,* 317–322.

McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: Behavioural and neural interactions. *Neuropsychologia, 50,* 762–776.

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

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience, 25,* 5884–5893.

Mishkin, M., & Ungerleider, L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural Brain Research, 6,* 57–77.

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, 309,* 951–954.

Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience, 31,* 1704–1714.

Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage, 59,* 781–787.

Nir, Y., Fisch, L., Mukamel, R., Gelbard-Sagiv, H., Arieli, A., Fried, I., et al. (2007). Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. *Current Biology, 17,* 1275–1285.

Obleser, J., Zimmermann, J., Van Meter, J., & Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related fMRI. *Cerebral Cortex, 17,* 2251–2257.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I. H., Saberi, K., et al. (2010). Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex, 20,* 2486–2495.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011,* 156869.

Pickles, J. O. (2015). Auditory pathways: Anatomy and physiology. *Handb Clin Neurol, 129,* 3–25.

Rauschecker, J. P. (2015). Auditory and visual cortex of primates: A comparison of two sensory systems. *European Journal of Neuroscience, 41,* 579–585.

Ray, S., & Maunsell, J. H. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biology, 9,* e1000610.

Reale, R. A., Calvert, G. A., Thesen, T., Jenison, R. L., Kawasaki, H., Oya, H., et al. (2007). Auditory-visual processing represented in the human superior temporal gyrus. *Neuroscience, 145,* 162–184.

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

Schepers, I. M., Schneider, T. R., Hipp, J. F., Engel, A. K., & Senkowski, D. (2013). Noise alters beta-band activity in superior temporal cortex during audiovisual speech processing. *Neuroimage, 70,* 101–112.

Schepers, I. M., Yoshor, D., & Beauchamp, M. S. (2015). Electrocorticography reveals enhanced visual cortex responses to visual speech. *Cerebral Cortex, 25,* 4103–4110.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123,* 2400–2406.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research, 47,* 277–287.

Sheffert, S. M., Lachs, L., & Hernandez, L. R. (1996). *Research on spoken language processing progress report no. 21* (pp. 578–583). Bloomington, IN: Speech Research Laboratory, Indiana University.

Specht, K., & Reul, J. (2003). Functional segregation of the temporal lobes into highly differentiated subsystems for auditory perception: An auditory rapid event-related fMRI-task. *Neuroimage, 20,* 1944–1954.

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage, 44,* 1210–1223.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Tolhurst, D. J., Movshon, J. A., & Dean, A. F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research, 23,* 775–785.

van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron, 43,* 271–282.

Wild, C. J., Davis, M. H., & Johnsrude, I. S. (2012). Human auditory cortex is sensitive to the perceived clarity of speech. *Neuroimage, 60,* 1490–1502.