Published in final edited form as:
J Cogn Neurosci. 2009 September ; 21(9): 1790–1805. doi:10.1162/jocn.2009.21118.

A Multisensory Cortical Network for Understanding Speech in Noise

Christopher W. Bishop and **Lee M. Miller**
University of California, Davis

Abstract

In noisy environments, listeners tend to hear a speaker's voice yet struggle to understand what is said. The most effective way to improve intelligibility in such conditions is to watch the speaker's mouth movements. Here we identify the neural networks that distinguish understanding from merely hearing speech, and determine how the brain applies visual information to improve intelligibility. Using functional magnetic resonance imaging, we show that understanding speech-in-noise is supported by a network of brain areas including the left superior parietal lobule, the motor/premotor cortex, and the left anterior superior temporal sulcus (STS), a likely apex of the acoustic processing hierarchy. Multisensory integration likely improves comprehension through improved communication between the left temporal–occipital boundary, the left medial-temporal lobe, and the left STS. This demonstrates how the brain uses information from multiple modalities to improve speech comprehension in naturalistic, acoustically adverse conditions.

INTRODUCTION

Speech, with its singular role in communication, is perhaps the most important everyday stimulus for humans. Unfortunately, speech is typically embedded in a noisy background or "cocktail party" (Cherry, 1953) of competing talkers, reverberations, and other environmental sounds. Under such adverse conditions, listeners tend to hear a speaker's voice but struggle to understand it. Hearing but failing to understand a speaker frustrates listeners with healthy hearing and takes an even greater mental and emotional toll on hearing-impaired listeners (Knutson & Lansing, 1990). Fortunately, watching the speaker's mouth move significantly improves intelligibility. Here we use fMRI to show how the brain distinguishes understanding from merely hearing speech in naturalistic conditions, and how it applies visual information to improve comprehension.

Similar to auditory-object identification areas in non-human primates (Rauschecker & Tian, 2000), human neuroimaging studies of intelligible auditory speech have identified specialized regions along the length of the superior/middle lateral temporal lobe. Although speech is processed extensively in both hemispheres, these speech-related areas show left-hemisphere dominance and multiple hierarchical foci, especially in the superior temporal gyrus (STG) and the superior temporal sulcus (STS) (Obleser, Wise, Alex Dresner, & Scott, 2007; Alain et al., 2005; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Giraud et al., 2004; Crinion, Lambon-Ralph, Warburton, Howard, & Wise, 2003; Davis & Johnsrude, 2003; Narain et al., 2003; Humphries, Willard, Buchsbaum, & Hickok, 2001; Scott, Blank, Rosen, & Wise, 2000; Johnsrude & Milner, 1994). A challenge in these experiments was to separate speech-

Reprint requests should be sent to Christopher W. Bishop, Center for Mind & Brain, University of California, 267 Cousteau Pl., Davis, CA 95618, or via cwbishop@ucdavis.edu.

specific or intelligibility-specific mechanisms from general acoustic processing incidental to the task. For instance, several studies controlled acoustics perfectly and altered perception through experience with stylized sine-wave speech, showing the importance of the left superior temporal lobe in distinguishing speech from nonspeech (Desai, Liebenthal, Waldron, & Binder, 2008; Dehaene-Lambertz et al., 2005; Liebenthal, Binder, Piorkowski, & Remez, 2003). In contrast, experiments focusing on comprehension—where all stimuli are heard as speech—typically alter the stimulus acoustics dramatically to manipulate intelligibility (Binder, Liebenthal, Possing, Medler, & Ward, 2004; Davis & Johnsrude, 2003; Scott et al., 2000) (but see Giraud et al., 2004). As a result, few reports address the key perceptual distinction between hearing speech with versus without understanding, denoted in this report as *understanding versus hearing*, in the absence of substantial acoustical confounds. It is thus unclear how much the left superior temporal pathway is involved with stimulus processing—speech-specific or not—versus the phenomenon of comprehension.

Comprehension has long been known to improve when watching a speaker's mouth movements, especially in noisy environments (Benoit, Mohamadi, & Kandel, 1994; MacLeod & Summerfield, 1987; Sumby & Pollack, 1954). Audiovisual integration of speech has been measured in humans with fMRI, electroencephalography, and magnetoencephalography (Bernstein, Auer, Wagner, & Ponton, 2008; Miller & D'Esposito, 2005; van Wassenhove, Grant, & Poeppel, 2005; Klucharev, Mottonen, & Sams, 2003; Sekiyama, Kanno, Miura, & Sugita, 2003; Callan, Callan, Kroos, & Vatikiotis-Bateson, 2001; Calvert, Campbell, & Brammer, 2000; Calvert et al., 1999). Imaging studies using speech as well as nonlinguistic stimuli consistently implicate the middle or posterior STS as a locus for audiovisual integration (Miller & D'Esposito, 2005; Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Macaluso, George, Dolan, Spence, & Driver, 2004; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003; Calvert, 2001; Raij, Uutela, & Hari, 2000), in conjunction with other high-level association regions such as the posterior parietal and prefrontal cortex (Ojanen et al., 2005; Ahissar et al., 2001). Only a few of these reports explicitly address intelligibility during audiovisual integration of speech (Callan et al., 2003, 2004; Sekiyama et al., 2003). For instance, Callan et al. (2004) used spatial wavelet filtered speech to show that perception based on visual place-of-articulation cues (e.g., "ba" vs. "ga") may recruit the left STS and the middle temporal gyrus (MTG). However, no studies have dissociated speech perception from low-level stimulus processing by controlling the profound acoustic and visual differences among stimuli with unequal intelligibility. We therefore do not yet know how the brain uses visual information to achieve its hallmark improvement in comprehension.

We used fMRI during a speech-identification task to achieve two related aims: (i) determine the neural bases of the pivotal behavioral distinction between *hearing* with versus without *understanding* (i.e., detecting speech with vs. without identification), and (ii) identify how the brain uses visual information to improve intelligibility. In a mixed event-related/blocked fMRI design, subjects with healthy hearing identified utterances embedded in an acoustic background of multi-talker babble. Target auditory stimuli were four vowel–consonant– vowel (VCV) utterances. Three stimulus conditions were presented pseudorandomly in blocks: auditory only (AX), audiovisual synchronous (AVs: video of speaker's mouth, synchronous with target utterance), and audiovisual offset (AVo: video occurring temporally offset from utterance). Subjects indicated with a button press whether they could unambiguously identify a target utterance (*understood*) or whether they could hear the utterance, but could not distinguish the VCV from the others (*heard*). For each utterance in each condition, we adaptively adjusted the target/babble signal-to-noise ratio (SNR) within a narrow range around each subject's threshold of *understanding*. We expected activity in the left anterior STS to reflect increased intelligibility, and increased functional connectivity between the STS and visual motion areas to mediate improved intelligibility through audiovisual integration.

METHODS

Subjects

Thirty-four healthy subjects (18 women, ages 18–35 years, mean age = 22 years) gave written consent according to procedures approved by the University of California and were paid or given class credit for their participation. All participants were right-handed, learned English as a first language, had normal or corrected-to-normal vision, and self-reported good hearing. None had taken psychoactive medication within 3 months prior to participating. Unless otherwise noted, all fMRI and behavioral data are based on 25 subjects (14 women, ages 18–35 years, mean age = 21 years) who met a behavioral selection criterion during the pre-fMRI training session.

Stimuli

All audiovisual stimuli used for the speech identification task, with the exception of background babble, were recorded with a digital camcorder and a remote microphone. An adult female speaker with vocal training produced four nonsense VCV utterances with flat affect and natural prosody. All audiovisual stimuli had a duration of 2 sec, with the utterance temporally centered so the consonant stop occurred 1 sec from the onset. The vowel was always [a] because its formant structure provided a superior SNR relative to the MRI scanner spectrum during functional runs. The four consonants were the stop consonants ([d], [g], [t], [k]), chosen for their balanced articulatory features and visual similarity. The auditory components of each recording were temporally aligned, filtered in Adobe Audition 1.5 (www.adobe.com), and equalized for RMS amplitude. Filtering was designed to control ambiguity among the VCV acoustics (60 dB linear decrease from 250 to 8 kHz) and to compensate for earplugs worn by subjects during the pre-fMRI training and fMRI sessions. The intensity (dB) of each VCV (signal) was adjusted independently relative to constant background 16-person babble (noise) in each of the three experimental conditions to target each subject's 50% identification threshold. The babble was generated by temporally offsetting and combining two instances of eight-speaker babble (4 men, 4 women), graciously provided by Pierre Divenyi. The babble was played continuously except during fixation periods.

A single video of a moving human mouth was played in all visual trials regardless of the utterance. The visual stimulus was made in Adobe Premiere Pro 1.5 by temporally aligning, overlaying, and blurring (40% Fast Blur) all four VCV videos (www.adobe.com). This process preserved gross movement related to the temporal envelope of speech while removing the high spatial-frequency information necessary to distinguish the four utterances based solely on visual cues. The video was exported from Adobe Premiere at 30 frames per second (fps), and each frame was manually duplicated for an effective frame rate of 60 fps. These images were then presented sequentially in Presentation at 60 Hz to reproduce the mouth movement. Only the lower half of the speaker's face was included in the video frame. The first frame of the video remained on the screen between trials during the audiovisual synchronous (AVs) and offset (AVo) conditions and was accompanied by a fixation cross placed near the speaker's mouth. Only the fixation cross was shown during fixation periods and the auditory only (AX) condition. All stimulus presentation was coordinated with Presentation software (Neuro-behavioral Systems; www.neuro-bs.com).

In a task designed to localize visual motion area MT+, subjects were presented with random dot fields moving in one of eight directions with 100% coherency. The stimuli were created in MATLAB using in-house scripts and exported as bitmaps for use in Presentation. Black dots on a white background moved across the entire projected image at an approximate rate of 3 deg/s. Dots were wrapped from one side of the screen to the other to maintain a dot density of

approximately 40% at all times. These dots were accompanied by a fixation cross in the center of the screen that occasionally dimmed from black to gray.

Pre-fMRI Training

All subjects participated in a pre-fMRI training session 1 to 2 days prior to fMRI scanning. The session was intended to characterize each subject's performance during a speech identification task. Six subjects also familiarized themselves with a MT-localizer task. Auditory stimuli were presented via headphones and attenuated by a pair of E.A.R. TaperFit2 foam earplugs (www.e-a-r.com). Visual stimuli were presented on a Dell 2005FPW monitor placed approximately 67 cm from the subject, resulting in a visual angle of approximately 17° for the face video.

All subjects viewed a PowerPoint presentation explaining the speech identification task. Subjects were allowed to continue only after demonstrating adequate understanding of the task. Four VCV stimuli were presented in continuous 16-person babble in three conditions: AX, AVs, and AVo. Auditory stimuli were temporally aligned with the video in the AVs condition and offset by aminimum of 800 msec in the AVo condition to prevent bimodal fusion (Miller & D'Esposito, 2005). Trials in the AVo condition can be grouped into three categories: audiovisual adjacent (audio leading by 800 msec or lagging by 1120 msec), auditory only, and video only. Approximately 50% to 60% of trials were audio leading. Stimulus onsets in audio only trials never occurred within 2.5 sec of a visual stimulus onset, and vice versa. For analysis purposes, all trials with an auditory stimulus were modeled as a single covariate and the onsets of visual stimuli temporally separated from an auditory stimulus by more than 800 msec were modeled as a separate covariate.

Subjects indicated whether they detected and could identify the VCV (*understood*) or detected the VCV but could not identify it (*heard*) by pressing their middle or index finger, respectively. The behavioral training lasted approximately 15 min and consisted of equal numbers of VCV utterances in each stimulus condition. The VCV/babble SNRs were independently varied in each condition by adjusting the VCV volume in 1- to 2-dB increments for each condition based on the subject's responses. For 21 of the 25 subjects who participated in the fMRI experiment, this followed a tracking algorithm, where SNR was adjusted to maintain approximately a 1:1 cumulative proportion of *understood/heard* responses. A modified one-up–one-down staircase algorithm was used for the remaining four subjects to achieve the same effect. The adaptive algorithms therefore allowed us to quickly target each subject's 50% identification threshold (i.e., threshold for *understanding*) in each condition. Nine of the 34 subjects who participated in the pre-fMRI training (4 women) failed to respond to (i.e., failed to press either button in response to an auditory stimulus) approximately 15% of the speech stimuli and were not asked to participate in the fMRI portion of the study. Unless otherwise noted, all behavioral and fMRI results are based on the 25 (14 women) subjects who satisfied this behavioral criterion.

Six of the scan subjects (1 woman) performed an MT-localizer task. Subjects fixated on a black cross on a white background while a field of black dots remained still or moved in one of eight directions. Subjects responded by pressing a button with their left index finger each time the cross dimmed from black to gray. Responses were monitored and logged to verify that subjects were maintaining fixation.

fMRI Scanning Methods

T2*-weighted EPI data sensitive to blood oxygenation level dependent (BOLD) activity were acquired using a 3-Tesla Siemens Trio MRI scanner and an eight-channel head coil with a one-shot EPI sequence [TR = 1.5 sec; 25 msec echo time; 64 × 64 × 26 acquisition matrix; 4.0 mm slice thickness; and a 220-mm field of view; bandwidth = 2365 Hz/pixel; flip angle = 90°] for

23 of the 25 fMRI subjects. EPI sequence parameters were similar for the first two subjects except for TR, matrix size, and slice thickness [TR = 2.0 sec or 2.18 sec; acquisition matrix = 64 × 64 × 34 or 64 × 64 × 28; slice thickness = 3.0 mm]. Subjects wore E.A.R. TaperFit2 earplugs to attenuate scanner noise. Visual stimuli were projected onto a screen at the subject's feet and a mirror was mounted to the eight-channel head coil to allow subjects to view the screen. Due to the bore size, the visual angle was limited to 12°. Auditory stimuli were delivered using a XITEL Pro HiFi-Link (www.xitel.com) USB-sound card, amplifier, and MR-compatible headphones (http://mr-confon.com) at a comfortable volume.

fMRI Scanning Task

All 25 subjects who participated in the fMRI study performed a speech identification task during six functional scans lasting approximately 7.5 min each. The stimulus presentation was as follows for 23 of the 25 subjects. Five sessions had a 30-sec leading and lagging fixation period and eight 46.6 sec of interleaved AX, AVs, and AVo blocks each. The remaining speech identification session consisted of 30-sec leading fixation, four interleaved, 46.6-sec AX and fixation blocks with one additional 30-sec AX block in place of a lagging fixation period. This gave a total of 17 AX, 16 AVs, and 12 AVo blocks. Except for the 30-sec AX block, each block consisted of 13 pseudorandomly presented auditory stimuli. VCV utterances occurred with a stimulus onset asynchrony (SOA) of 3.0, 4.5, or 6.0 sec (average = 3.6 sec) in exponentially decreasing proportions (67%, 25%, and 8%, respectively), with SOAs and VCV identities balanced across all conditions. The stimulus presentations for the remaining two subjects were closely matched to this paradigm. Subjects indicated whether they could identify the VCV presented in a background of multi-talker babble by pressing their left middle and index fingers. The SNR of each VCV in each condition was adaptively adjusted near each subject's threshold of *understanding* throughout all six functional scans with one of two algorithms, as described in the pre-fMRI training. Responses for this and the MT-localizer were monitored and logged via Presentation (Neurobehavioral Systems). One subject's response times were not logged due to technical failure.

Six subjects (1 woman) performed an MT-localizer task following the speech identification task. The scan consisted of 24-sec leading and lagging fixation periods with 11 moving dot blocks and 11 fixed dot blocks each lasting 18 sec. A fixation cross faded from black to gray, with an average SOA of approximately 4.0 sec. Subjects attended the cross and indicated a change in luminance (black fading to gray) by pressing their left index finger.

Accuracy vs. Perception

Four subjects (3 women, 1 man) were asked to participate in a follow-up psychophysical experiment to determine the correlation between perception (i.e., *understanding* vs. *hearing*) and accuracy of VCV identification. In this experiment, subjects were presented with six to seven approximately 6-min sessions of the speech identification task, each containing 104 VCV stimuli presented in pseudorandom order in the auditory-only condition (see prescan/fMRI scanning task for details). In the interest of time, fixation periods were omitted from this experiment. The task was virtually identical to the prescan and fMRI scanning tasks with an added four alternative forced-choice follow-up question. First, as in the task performed in the fMRI experiment, subjects responded whether or not they subjectively believed they could identify the VCV (i.e., *understood*) by pressing either their left index or middle finger. Regardless of their subjective percept (i.e., *understood* or *heard*), subjects then indicated which VCV they thought was presented by pressing their right index, middle, ring, and little fingers, corresponding to *ada*, *aga*, *aka*, and *ata*, respectively. The same adaptive algorithm was used to adjust the SNR of each VCV in 1- to 2-dB increments (i.e., identification accuracy had no impact on the algorithm). Subjects successfully identified the target VCV (mean ± *SEM*) 66.1 ± 4.7% for the *heard* percept and 86.5 ± 3.7% for the *understood* percept, equating to a 20.4

± 2.9% improvement in accuracy. These behavioral data suggest that subjects conservatively responded *understood*, making any inferences drawn from the fMRI data conservative as well.

fMRI Data Preprocessing

fMRI data were processed on-line with Siemens realignment and distortion correction algorithms before being converted from DICOM 3.0 to Analyze format using XMedCon (xmedcon.sourceforge.net). Analyze images were then corrected for slice acquisition time, spatially realigned, and smoothed with an 8-mm^3 full-width half-maximum kernel in SPM2. Within-session linear trends were removed using an in-house normalization routine that calculated the mean global signal level over all brain voxels for each time point, fitted a line to each session's mean global estimates, and then scaled (divided) these values by the piecewise linear fit. T2, EPI and high-resolution MP-RAGE anatomical images were coregistered and normalized to an MNI template and resampled at 2-mm isotropic voxels before data analysis. Due to the length of our radio-frequency coil, high-resolution anatomical images suffered signal loss at the base of the cerebellum, which could potentially cause erroneous normalization to the standard MNI template. As a result, subject brains were skull-stripped and normalized to a skull-stripped MNI template whose cerebellum faded inferiorly to match the observed signal loss. These modifications were done with MRIcro (www.sph.sc.edu/comd/rorden/mricro.html) and MATLAB (www.mathworks.com), and yielded robust normalizations. Except where noted, all statistical results are from random effects, group tests corrected for multiple comparisons. Also, thresholded SPM *t* maps were smoothed in MRIcro for display purposes.

Data Analysis

Speech Identification—Speech identification data were analyzed using two general linear models (GLM) in SPM2 to explore differences in BOLD activity based on perception and on stimulus condition. Because the distribution of perceptual events were, by design, balanced across all stimulus conditions, no systematic relationships existed between the two. The first was an event-related GLM used to model perception of auditory stimuli: *understood* versus *heard*. Vectors of onset times were created for all *heard* and *understood* auditory stimuli in the AX, AVs, and AVo conditions. Additionally, the onsets of visual stimuli in the AVo condition that were temporally separated from an auditory stimulus by more than 800 msec were included as a regressor in the design matrix. All covariates of interest, including the aforementioned visual stimuli in the AVo condition, were convolved with a canonical hemodynamic response function (HRF) and its time derivative. Siemens and SPM2-generated realignment parameters and session effects were included as confounds. Importantly, SNR was also included as a regressor for each subject. The SNR for each trial was scaled (divided by) the maximum deviation from threshold (i.e., the maximum absolute value of mean centered SNR). The scaled SNR was then used to modulate parametrically the same canonical HRF and included as a regressor in the design matrix. Contrast images for each subject were included in a group-level *t* test. Stimulus conditions were analyzed with a modified block-level GLM. Onset vectors were created for each block in all six sessions, convolved with the canonical HRF, and included in a design matrix with motion confounds and session effects. The same previously described SNR regressor was also included in the block GLM.

MT-localizer—MT-localizer data were analyzed using a block GLM in SPM2. Boxcar vectors modeling the moving dot blocks were convolved with the canonical HRF and included, with motion confounds, in the design matrix. To increase power, group tests were performed by treating each hemisphere as a separate "subject" (right hemispheres flipped to the left), thereby doubling the number of data points. Also, we calculated the center of mass for area MT+ for each subject using the following procedure. First, MT+ was functionally defined by identifying voxels significant (FDR $p < .05$, data not shown) in a group-level *t* test of the

"moving dots" boxcar covariate. Second, MT+ was anatomically restricted to the posterior/lateral temporal lobe. Finally, the center of mass was computed for each subject. This was done by summing the product of all positive *t* values and their [*x y z*] coordinates (mm) and dividing this by the sum of all positive *t* values within area MT+. A similar calculation was carried out for the left lateral temporal–occipital boundary (LTO), as defined in the *understood* > *heard* contrast (FDR $p < .05$; Figure 2). The Euclidean distance between the two centers of mass was then computed for each subject and a paired *t* test was performed in MATLAB to determine whether this distance was significantly greater than zero across subjects (see Results).

Functional Connectivity—We used the method described by Rissman, Gazzaley, and D'Esposito (2004) to compare functional connectivity among brain regions when subjects *heard* and *understood* speech. In a massive event-related GLM in SPM2, the onset of every perceived auditory stimulus (*heard* or *understood*), and any visual stimulus temporally distant from an auditory stimulus (>800 msec) in the AVo condition, was identified. Each onset was convolved with a canonical HRF and included as a separate covariate in the design matrix. Motion confounds were included and each session was estimated independently.

The activity estimates for all trials (parameter estimates, or betas) were sorted by percept and used to construct connectivity maps. If one percept had more trials than the other for a particular subject, a random sample of trials equal to the number of trials in the other percept was used. A seed or reference region was defined for each subject by finding a region consisting of no more than 10 of the most significant, contiguous voxels within a functionally defined region of interest (ROI). Significance was based on an *F* test for all covariates-of-interest in the event-related speech identification GLM. Trial beta estimates for these seed voxels were averaged and combined to create an average beta time series for the reference region (i.e., reference-beta-series) for each percept. Pearson's correlation coefficients between the reference-beta-series and the other voxels' beta-series were then calculated using in-house scripts in MATLAB. The resulting Pearson's correlation coefficients were *z*-transformed by calculating the hyperbolic arctangent of each coefficient and dividing it by the known standard deviation of the hyperbolic arctangent function [1/sqrt(N − 3), where *N* is the number of trials used in the correlation]. Correlation maps for *understood* and *heard* percepts were constructed for each subject and subtracted to create the equivalent of an SPM2 contrast map for functional connectivity for each seed region, as in similar analyses using coherence (Sun, Miller, & D'Esposito, 2004). These "contrast maps" were included in a group *t* test to determine if the *z*-transformed correlation coefficient between a voxel and the seed was significantly greater when subjects *understood* versus *heard* the target speech. We tested for increased connectivity (i.e., correlation) between regions of the left temporal lobe, including the left medial-temporal lobe (MTL), LTO, and the left anterior STS. MTL and LTO ROIs were functionally and anatomically defined in a conjunction between *understood* > *heard* across all conditions ($p < .005$) and AVs > AVo ($p < .005$). Although these regions were both significant at FDR $p < .05$ in both contrasts, a relaxed threshold was used to achieve robust seed selection across subjects. The left anterior STS ROI (Figure 2F) was anatomically and functionally defined in the *understood* > *heard* contrast in the audioonly condition at a relaxed threshold ($p < .05$), but was significant at $p < .005$ (see Supplementary Figure S1 for average beta estimates for each of these seed regions).

SNR Sensitivity—In order to identify brain regions sensitive to SNR, we included the subject-level beta estimates for the event-related SNR regressor (see Data Analysis for details) in a group-level *t* test.

RESULTS

Behavior

Perception as a function of SNR is shown in Figure 1A and B. *Understanding* thresholds are normalized to 0 dB, corresponding to a mean of −4.71 dB relative to background babble. Across all subjects, conditions, and utterances, greater SNR gave a smooth, steep function of increasing *understood* responses. The vast majority of utterances (90%) were presented within a narrow range around each subject's threshold (±4 dB). This low SNR variance between trials and percepts, with mean SNR anchored at threshold, guaranteed that the SNR probability distributions for heard and understood percepts be highly overlapping (Figure 1B). Any given SNR value therefore had a substantial probability of being *heard* or *understood*. Consequently, for virtually identical speech stimuli, subjects reported the two distinct percepts in equal proportions (mean *understood* proportion ± *SEM* = 50.8 ± 0.9% for 21 subjects using a 50% tracking algorithm, and 52.9 ± 1.6%, for all 25 subjects, 4 of whom used a one-up-one-down staircase algorithm; see Methods).

Performance was also essentially identical across stimulus conditions AX, AVs, and AVo. The proportions *of understood* responses were indistinguishable [Figure 1C: mean ± *SEM* of 52.4 ± 1.6% for AX, 52.2 ± 1.4% for AVs, and 53.9 ± 1.9% for AVo, $p = .77$, ANOVA F(2, 72)], as were reaction times [Figure 1D: mean ± *SEM* from stimulus onset: AX, 1309 ± 28 msec; AVs, 1310 ±31 msec; AVo, 1326 ± 32 msec; $p = .90$, ANOVA $F(2, 69)$], suggesting that the task was equally difficult. This approach also highlighted the expected improvement in intelligibility associated with cross-modal integration in the AVs condition. Improved intelligibility was demonstrated by a significant decrease in the SNR of *understood* speech in the AVs condition relative to *understood* speech in the AVo and AX conditions [mean ± *SEM* −3.40 ±0.71 dB for AX, −4.59 ± 0.71 dB for AVs, −3.26 ± 0.65 dB for AVo, $p = .03$, ANOVA $F(2, 72)$]. That is, subjects *understood* noisier acoustic speech when it occurred with synchronous visual cues.

Understanding versus Hearing

We evaluated the BOLD signal based on whether subjects could detect and identify (*understood*) or could detect but not identify (*heard*) the target VCV. Across all stimulus conditions, a broad network of brain regions showed greater activity with *understanding* (FDR$p < .05$; Figure 2, Table 1). Temporal regions included the left MTL, the bilateral temporal–occipital boundary (LTO), and the right hippocampus. Parietal and frontal regions included the left superior parietal lobule (SPL) and the posterior intraparietal sulcus (IPS), the left superior frontal sulcus (SFS), the bilateral postcentral gyrus (PostCG) and the central sulcus (CS), and the right cingulate sulcus. Subcortical structures included the left body of the caudate nucleus, the bilateral putamen and tail of the caudate nucleus. Despite previous reports linking the left STS to improved speech intelligibility, our analysis did not show a significant increase in mean BOLD activity for the *understood* > *heard* contrast across all conditions (FDR $p < .05$). However, as in numerous other studies (Uppenkamp, Johnsrude, Norris, Marslen-Wilson, & Patterson, 2006;Liebenthal et al., 2005;Davis & Johnsrude, 2003;Scott & Johnsrude, 2003;Scott et al., 2000), the left anterior STS showed greater mean BOLD activity in the *understood* > *heard* contrast in the AX condition (Figure 2F; significant at $p < .005$, but displayed at $p < .05$). We confirmed that LTO is spatially distinct from visual motion area MT + by comparing their centers of mass (see Methods). The left LTO's center of mass (mean [−42.35, −70.40, 10.85]) was spatially separated from the left MT+'s center of mass (mean [−41.11, −75.70, −2.22]) by a Euclidean distance of 14.49 ± 1.32 mm (mean ± *SEM*), with the LTO lying dorsal and anterior to MT+. This distance was significantly greater than zero (paired *t* test, $p < .002$).

Audiovisual Contribution to Understanding

We identified brain areas that use visual information for *understanding* through a strict requirement: A region must show *both* greater activity with cross-modal integration *and* greater activity with *understanding*. First, we found all areas sensitive to cross-modal integration of speech, having larger BOLD for synchronous (AVs) versus temporally offset (AVo) audiovisual speech. These included the left bilateral LTO, the left MTL, the inferior frontal gyrus, the bilateral IPS, the left precentral sulcus, the bilateral occipito-temporal sulcus, and the right PostCG (FDR $p < .05$; Figure 3, Table 1). We then conjoined those areas with our *understanding* network described above (also significant at FDR $p < .05$). Importantly, only a subset of our network for *understanding* satisfied both criteria. These regions included the left MTL, the bilateral LTO, the left posterior IPS, the bilateral putamen, and the right PostCG (Figure 4). A whole-brain repeated measures ANOVA in SPM2 revealed no significant interactions between condition and perception (FDR $p < .05$). Additionally, we performed repeated measures ANOVAs on average beta estimates in our key temporal lobe regions: left STS, left MTL, and left LTO (see Supplementary Figure S1). Each region was functionally defined by the same criteria used for functional connectivity seed selection (see Methods). None of these regions showed a significant interaction between percept and condition [$F(2, 48)$, $p > .05$].

The conjunction analysis distinguishes which regions contribute visual information toward *understanding*, but cannot identify how or in what network context this occurs. We therefore performed functional connectivity analysis on trial-by-trial BOLD estimates (Rissman et al., 2004) of regions sensitive to *understanding* in the left temporal lobe. We hypothesized that the perceptual difference between *understanding* and *hearing* would be reflected by increased functional connectivity between temporal regions sensitive to *understanding*, including the left LTO, the anterior STS, and the MTL. The left LTO showed increased connectivity with the left MTL when subjects *understood* versus *heard*, but not with the left STS. Interestingly, the left MTL showed increased connectivity with the left STS, suggesting that the left MTL may act as a relay or hub for cross-modal information within our *understanding* network ($p < .05$; Figure 5).

Quantifying Acoustical Effects

In order to identify brain regions sensitive to SNR, we performed a group-level *t* test on the SNR regressor included in the event-related GLM (see Methods). As expected, the BOLD time course for our *understanding* network was not adequately explained by SNR variance. Instead, regions sensitive to SNR were limited to the bilateral Heschl's gyri and STG, the middle STS, the left MTG, and a small number of occipital regions, including the bilateral parieto-occipital sulcus and the bilateral calcarine sulcus (Table 1, Figure 6; FDR $p < .05$).

DISCUSSION

Here we present a unified account of how the brain understands speech in acoustically adverse conditions and how it applies visual information to improve intelligibility. In contrast to previous studies (Dehaene-Lambertz et al., 2005; Binder et al., 2004; Liebenthal et al., 2003), all our stimuli were heard as speech and all acoustic variations (i.e., SNR variance) were accounted for in our models. A broad network of areas distinguishes *understanding* from *hearing*. As in other studies, the left anterior STS, a putative apex of the temporal lobe's speech-processing hierarchy (Uppenkamp et al., 2006; Liebenthal et al., 2005; Davis & Johnsrude, 2003; Scott & Johnsrude, 2003; Scott et al., 2000), was sensitive to intelligibility in the audio-only condition. Only a subset of the *understanding* areas also showed greater activity with audiovisual integration: left MTL, bilateral LTO, left posterior IPS, bilateral putamen, bilateral

tail of the caudate nucleus, and right PostCG. Interestingly, regions located in the temporal lobe communicate more effectively when speech is *understood* versus *heard*.

Although members of the same functional network, the LTO and the MTL likely play very different roles in speech comprehension. MTL structures are perhaps best known for their importance in declarative memory, including episodic memory for verbal material (Strange, Otten, Josephs, Rugg, & Dolan, 2002; Martin, Wiggs, & Weisberg, 1997). However, MTL structures have also been reported in speech and language tasks without manifestly mnemonic demands (Hoenig & Scheef, 2005; Meyer et al., 2005; Davis & Johnsrude, 2003). Our results support the view that the MTL is involved in perception as well as memory (Murray, Bussey, & Saksida, 2007; Buckley, 2005; Martin et al., 1997), particularly in perceiving ambiguous stimuli including sentences (MacKay, Stewart, & Burke, 1998; Rotenberg & Muller, 1997) (but see (Rodd, Davis, & Johnsrude, 2005)). The hippocampus, for instance, may further improve comprehension of audiovisual speech by evaluating cross-modal congruence at an abstract representational level (Gottfried & Dolan, 2003), possibly supported by strong reciprocal projections from the auditory and multisensory STS (Van Hoesen, 1995).

The LTO, on the other hand, lies at a relatively high level of sensory processing, ventral/posterior to intelligibility loci identified in auditory-only studies (Davis & Johnsrude, 2003; Narain et al., 2003). It could easily communicate with the adjacent language-related cortex, including “Wernicke’s area” (Wise et al., 2001). For instance, posterior temporal areas have been shown to reflect lexico-phonological and lexico-semantic processing (Vandenbulcke, Peeters, Dupont, Van Hecke, & Vandenberghe, 2007; Demonet, Thierry, & Cardebat, 2005; Binder et al., 2000). However, because our utterances were recognizable yet meaningless nonwords, we believe the LTO instead supports a more general function, such as object recognition (Griffiths & Warren, 2004; Lewis et al., 2004), based on features from both auditory and visual modalities (Beauchamp, Lee, Haxby, & Martin, 2003; Wise et al., 2001). One attractive hypothesis is that the LTO uses visual biological motion to constrain noisy, ambiguous auditory representations (Hall, Fussell, & Summerfield, 2005; Callan et al., 2003; Santi, Servos, Vatikiotis-Bateson, Kuratate, & Munhall, 2003; Olson, Gatenby, & Gore, 2002; Campbell et al., 2001; Zatorre, 2001; MacSweeney et al., 2000) that are not necessarily speech-specific (Beauchamp et al., 2003; Allison, Puce, & McCarthy, 2000).

We have therefore identified the importance of the left MTL and the bilateral LTO for improving intelligibility through vision. However, the univariate nature of these analyses could not characterize the cooperative networks whereby these areas exert their influence. Functional connectivity analysis showed that the MTL communicates differentially with both the STS and the LTO during understanding, whereas the LTO lacks strong or differential connectivity with the STS. Thus, the MTL may play an integrative role in comprehension, associating information from the LTO and the STS. The STS and the LTO, in contrast, may operate more independently of each other, contributing complementary information relevant to the task. For instance, the STS and the LTO may provide auditory speech decoding and supramodal object/lexical information toward *understanding*, respectively.

Although the present study did not attempt to isolate which visual features contribute to intelligibility, our choice of visual stimulus was intended to provide an interpretive constraint. It was created by averaging all four VCV videos so vision alone could not uniquely distinguish the auditory utterances. Therefore, information about place-of-articulation (POA), or where the vocal tract constricts to create a given sound, was unavailable. Consequently, our subjects’ behavioral improvement by adding vision is due primarily to the temporal envelope of speech, which is robustly represented in the visual signal (Grant & Seitz, 2000). This, and the fact that we strictly controlled intelligibility across conditions, might explain why our results differ from prior studies, one of which attempted to isolate the unique contribution of POA (Callan et al.,

2003, 2004). We chose our visual signal for three reasons. First, we wanted the available information to be highly naturalistic. POA is not always apparent in real environments with variable lighting and head orientation (restaurants, meetings, etc.), where the listener cannot always see the speaker's tongue. Second, we wanted the task to be challenging in a naturalistic way, namely, by requiring that the subject use both modalities whenever they offered complementary information. Finally, we wanted to highlight the most beneficial attributes of audiovisual speech. POA, surprisingly, may be the least informative part of visual speech: Low-pass spatial-filtered visual speech similar to ours can show an audiovisual improvement equivalent to unfiltered speech, with high-visual-frequency information, including POA, being largely redundant (Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004).

One essential control in our design, noteworthy in its own right, is explicitly modeling acoustic variability (SNR). The only areas sensitive to SNR were the bilateral Heschl's gyri, the STG, the STS, the left MTG, and a few small regions in the occipital lobe. Crucially, no areas that reflected *understanding* also showed sensitivity to SNR, thereby confirming their independence from low-level signal attributes. Recall that the stimulus variance about each subject's threshold was typically an extremely narrow ±1 to 4 dB. This result refines what we know from large acoustic intensity changes, which affect BOLD spatial extent and intensity in primary and adjacent nonprimary auditory cortices (Binder et al., 2004; Jancke, Shah, Posse, Grosse-Ryuken, & Muller-Gartner, 1998). Also relevant is a study on foreground/ background decomposition that identified the rostrolateral Heschl's gyrus as maintaining acoustic target-related activity against a frequency-modulated tone background (Scheich et al., 1998). Finally, in the one other report that parametrically distorted speech while controlling for intelligibility, similar regions in the left STG were shown to have increased activity for speech-in-noise versus normal speech or modulated noise alone (Davis & Johnsrude, 2003).

Although the current study cannot specify the roles of the remaining members of the *understanding* network, it suggests a number of testable hypotheses. Outside the temporal lobe, the *understanding* regions may be loosely grouped as: (i) high-level cortices for supramodal association, attention, and control; or (ii) areas with motor or sensorimotor representations of speech. The high-level cortices include the left posterior IPS, the anterior cingulate sulcus, the left SFS, and the left SPL. The posterior IPS likely contributes to *understanding* through multisensory integration and supramodal spatio-temporal transformations, as shown with nonspeech and speech stimuli (Molholm et al., 2006; Miller & D'Esposito, 2005; Calvert, 2001; Andersen, 1997). The anterior cingulate, in contrast, tends to be more involved with conflict monitoring, error detection, and effortful control (Botvinick, Cohen, & Carter, 2004). For realistic speech-in-noise tasks, it may be important for monitoring and rejecting irrelevant cross-modal information to improve auditory comprehension (Weissman, Warner, & Woldorff, 2004), as well as resolving difficult perceptual decisions such as phonemic category (Blumstein, Myers, & Rissman, 2005; Giraud et al., 2004). The left SFS could work in concert with the anterior cingulate, consistent with the dorsolateral prefrontal cortex's role in controlled processing while inhibiting multisensory distraction (Weissman et al., 2004). A fourth high-level cortex reflecting *understanding* was the left SPL. This area probably contributes to comprehension in noisy environments through its control of auditory and supramodal attention (Wu, Weissman, Roberts, & Woldorff, 2007; Shomstein & Yantis, 2004, 2006). Utterances to which subjects successfully allocate their attention would be more consistently *understood*.

Finally, two areas sensitive to *understanding* deserve special mention for their role in speech production: the CS with the adjacent precentral gyrus and PostCG and the bilateral striatum. Our CS loci overlap precisely with those observed for both speech perception and production (Wilson, Saygin, Sereno, & Iacoboni, 2004). This suggests a "mirror" system (Rizzolatti & Arbib, 1998) in which motor or sensorimotor representations could be recruited for the purpose of perception, particularly during multisensory integration (Skipper, van Wassenhove,

Nusbaum, & Small, 2007; Skipper, Nusbaum, & Small, 2005). The caudate and the putamen have furthermore been shown to affect sequencing in both speech perception and production (Giraud et al., 2007; Pickett, Kuniholm, Protopapas, Friedman, & Lieberman, 1998). Their involvement has been noted in explicit linguistic processing (Friederici, Ruschemeyer, Hahne, & Fiebach, 2003; Binder et al., 1997), as well as in identifying degraded versus clean speech (Meyer, Steinhauer, Alter, Friederici, & von Cramon, 2004) or environmental sounds (Lewis et al., 2004). Considering its strong auditory and multisensory inputs (Nagy, Eordegh, Paroczy, Markus, & Benedek, 2006; Yeterian & Pandya, 1998) and its outputs to speech motor areas (Henry, Berman, Nagarajan, Mukherjee, & Berger, 2004), the striatum may serve a function supportive of a mirror system in the frontal cortex.

Countless listeners, including half a billion worldwide with hearing loss, struggle daily to understand speech in adverse acoustic environments. Virtually all of them can benefit from watching the speaker's mouth move. Here we identify the neural networks that distinguish *understanding* from merely hearing speech in noise, and show how the brain uses visual information to improve intelligibility. This report serves as a basis for several further directions in speech research and speaks to the broader topic of ecological object/event perception. Speech represents a paradigm for how the brain recognizes highly overlearned objects under degraded conditions by harnessing contextual cues and integrating sensory modalities.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the National Institutes of Health (NIH)/National Institute on Deafness and Other Communication Disorders (NIDCD). We thank Kristina Backer for her help collecting behavioral data, Pierre Divenyi for the multi-talker babble stimuli, and David Whitney for helpful suggestions regarding the MT-localizer.

REFERENCES

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proceedings of the National Academy of Sciences, U.S.A 2001;98:13367–13372.

Alain C, Reinke K, McDonald KL, Chau W, Tam F, Pacurar A, et al. Left thalamo-cortical network implicated in successful speech separation and identification. Neuroimage 2005;26:592–599. [PubMed: 15907316]

Allison T, Puce A, McCarthy G. Social perception from visual cues: Role of the STS region. Trends in Cognitive Sciences 2000;4:267–278. [PubMed: 10859571]

Andersen RA. Multimodal integration for the representation of space in the posterior parietal cortex. Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences 1997;352:1421–1428.

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A. Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. Nature Neuroscience 2004;7:1190–1192.

Beauchamp MS, Lee KE, Haxby JV, Martin A. FMRI responses to video and point-light displays of moving humans and manipulable objects. Journal of Cognitive Neuroscience 2003;15:991–1001. [PubMed: 14614810]

Benoit C, Mohamadi T, Kandel S. Effects of phonetic context on audio-visual intelligibility of French. Journal of Speech and Hearing Research 1994;37:1195–1203. [PubMed: 7823561]

Bernstein LE, Auer ET Jr, Wagner M, Ponton CW. Spatiotemporal dynamics of audiovisual speech processing. Neuroimage 2008;39:423–435. [PubMed: 17920933]

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, et al. Human temporal lobe activation by speech and nonspeech sounds. Cerebral Cortex 2000;10:512–528. [PubMed: 10847601]

Binder JR, Frost JA, Hammeke TA, Cox RW, Rao SM, Prieto T. Human brain language areas identified by functional magnetic resonance imaging. Journal of Neuroscience 1997;17:353–362. [PubMed: 8987760]

Binder JR, Liebenthal E, Possing ET, Medler DA, Ward BD. Neural correlates of sensory and decision processes in auditory object identification. Nature Neuroscience 2004;7:295–301.

Blumstein SE, Myers EB, Rissman J. The perception of voice onset time: An fMRI investigation of phonetic category structure. Journal of Cognitive Neuroscience 2005;17:1353–1366. [PubMed: 16197689]

Botvinick MM, Cohen JD, Carter CS. Conflict monitoring and anterior cingulate cortex: An update. Trends in Cognitive Sciences 2004;8:539–546. [PubMed: 15556023]

Buckley MJ. The role of the perirhinal cortex and hippocampus in learning, memory, and perception. Quarterly Journal of Experimental Psychology B 2005;58:246–268.

Callan DE, Callan AM, Kroos C, Vatikiotis-Bateson E. Multimodal contribution to speech perception revealed by independent component analysis: A single-sweep EEG case study. Brain Research, Cognitive Brain Research 2001;10:349–353. [PubMed: 11167060]

Callan DE, Jones JA, Munhall K, Callan AM, Kroos C, Vatikiotis-Bateson E. Neural processes underlying perceptual enhancement by visual speech gestures. NeuroReport 2003;14:2213–2218. [PubMed: 14625450]

Callan DE, Jones JA, Munhall K, Kroos C, Callan AM, Vatikiotis-Bateson E. Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. Journal of Cognitive Neuroscience 2004;16:805–816. [PubMed: 15200708]

Calvert GA. Crossmodal processing in the human brain: Insights from functional neuroimaging studies. Cerebral Cortex 2001;11:1110–1123. [PubMed: 11709482]

Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, David AS. Response amplification in sensory-specific cortices during crossmodal binding. NeuroReport 1999;10:2619–2623. [PubMed: 10574380]

Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Biology 2000;10:649–657. [PubMed: 10837246]

Campbell R, MacSweeney M, Surguladze S, Calvert G, McGuire P, Suckling J, et al. Cortical substrates for the perception of face actions: An fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). Brain Research, Cognitive Brain Research 2001;12:233–243. [PubMed: 11587893]

Cherry EC. Some experiments on the recognition of speech with one and with two ears. Journal of the Acoustical Society of America 1953;25:975–979.

Crinion JT, Lambon-Ralph MA, Warburton EA, Howard D, Wise RJ. Temporal lobe regions engaged during normal speech comprehension. Brain 2003;126:1193–1201. [PubMed: 12690058]

Davis MH, Johnsrude IS. Hierarchical processing in spoken language comprehension. Journal of Neuroscience 2003;23:3423–3431. [PubMed: 12716950]

Dehaene-Lambertz G, Pallier C, Serniclaes W, Sprenger-Charolles L, Jobert A, Dehaene S. Neural correlates of switching from auditory to speech perception. Neuroimage 2005;24:21–33. [PubMed: 15588593]

Demonet JF, Thierry G, Cardebat D. Renewal of the neurophysiology of language: Functional neuroimaging. Physiological Reviews 2005;85:49–95. [PubMed: 15618478]

Desai R, Liebenthal E, Waldron E, Binder JR. Left posterior temporal regions are sensitive to auditory categorization. Journal of Cognitive Neuroscience 2008;20:1174–1188. [PubMed: 18284339]

Friederici AD, Ruschemeyer SA, Hahne A, Fiebach CJ. The role of left inferior frontal and superior temporal cortex in sentence comprehension: Localizing syntactic and semantic processes. Cerebral Cortex 2003;13:170–177. [PubMed: 12507948]

Giraud AL, Kell C, Thierfelder C, Sterzer P, Russ MO, Preibisch C, et al. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cerebral Cortex 2004;14:247–255. [PubMed: 14754865]

Giraud AL, Neumann K, Bachoud-Levi AC, von Gudenberg AW, Euler HA, Lanfermann H, et al. Severity of dysfluency correlates with basal ganglia activity in persistent developmental stuttering. Brain and Language 2007;104:190–199. [PubMed: 17531310]

Gottfried JA, Dolan RJ. The nose smells what the eye sees: Crossmodal visual facilitation of human olfactory perception. Neuron 2003;39:375–386. [PubMed: 12873392]

Grant KW, Seitz PF. The use of visible speech cues for improving auditory detection of spoken sentences. Journal of the Acoustical Society of America 2000;108:1197–1208. [PubMed: 11008820]

Griffiths TD, Warren JD. What is an auditory object. Nature Reviews Neuroscience 2004;5:887–892.

Hall DA, Fussell C, Summerfield AQ. Reading fluent speech from talking faces: Typical brain networks and individual differences. Journal of Cognitive Neuroscience 2005;17:939–953. [PubMed: 15969911]

Henry RG, Berman JI, Nagarajan SS, Mukherjee P, Berger MS. Subcortical pathways serving cortical language sites: Initial experience with diffusion tensor imaging fiber tracking combined with intraoperative language mapping. Neuroimage 2004;21:616–622. [PubMed: 14980564]

Hoenig K, Scheef L. Mediotemporal contributions to semantic processing: fMRI evidence from ambiguity processing during semantic context verification. Hippocampus 2005;15:597–609. [PubMed: 15884095]

Humphries C, Willard K, Buchsbaum B, Hickok G. Role of anterior temporal cortex in auditory sentence comprehension: An fMRI study. NeuroReport 2001;12:1749–1752. [PubMed: 11409752]

Jancke L, Shah NJ, Posse S, Grosse-Ryuken M, Muller-Gartner HW. Intensity coding of auditory stimuli: An fMRI study. Neuropsychologia 1998;36:875–883. [PubMed: 9740361]

Johnsrude I, Milner B. The effect of presentation rate on the comprehension and recall of speech after anterior temporal-lobe resection. Neuropsychologia 1994;32:77–84. [PubMed: 8818156]

Klucharev V, Mottonen R, Sams M. Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Brain Research, Cognitive Brain Research 2003;18:65–75. [PubMed: 14659498]

Knutson JF, Lansing CR. The relationship between communication problems and psychological difficulties in persons with profound acquired hearing loss. Journal of Speech and Hearing Disorders 1990;55:656–664. [PubMed: 2232746]

Lewis JW, Wightman FL, Brefczynski JA, Phinney RE, Binder JR, DeYoe EA. Human brain regions involved in recognizing environmental sounds. Cerebral Cortex 2004;14:1008–1021. [PubMed: 15166097]

Liebenthal E, Binder JR, Piorkowski RL, Remez RE. Short-term reorganization of auditory analysis induced by phonetic experience. Journal of Cognitive Neuroscience 2003;15:549–558. [PubMed: 12803966]

Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA. Neural substrates of phonemic perception. Cerebral Cortex 2005;15:1621–1631. [PubMed: 15703256]

Macaluso E, George N, Dolan R, Spence C, Driver J. Spatial and temporal factors during processing of audiovisual speech: A PET study. Neuroimage 2004;21:725–732. [PubMed: 14980575]

MacKay DG, Stewart R, Burke DM. H.M. revisited: Relations between language comprehension, memory, and the hippocampal system. Journal of Cognitive Neuroscience 1998;10:377–394. [PubMed: 9869711]

MacLeod A, Summerfield Q. Quantifying the contribution of vision to speech perception in noise. British Journal of Audiology 1987;21:131–141. [PubMed: 3594015]

MacSweeney M, Amaro E, Calvert GA, Campbell R, David AS, McGuire P, et al. Silent speechreading in the absence of scanner noise: An event-related fMRI study. NeuroReport 2000;11:1729–1733. [PubMed: 10852233]

Martin A, Wiggs CL, Weisberg J. Modulation of human medial temporal lobe activity by form, meaning, and experience. Hippocampus 1997;7:587–593. [PubMed: 9443055]

Meyer M, Steinhauer K, Alter K, Friederici AD, von Cramon DY. Brain activity varies with modulation of dynamic pitch variance in sentence melody. Brain and Language 2004;89:277–289. [PubMed: 15068910]

Meyer P, Mecklinger A, Grunwald T, Fell J, Elger CE, Friederici AD. Language processing within the human medial temporal lobe. Hippocampus 2005;15:451–459. [PubMed: 15714509]

Miller LM, D'Esposito M. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. Journal of Neuroscience 2005;25:5884–5893. [PubMed: 15976077]

Molholm S, Sehatpour P, Mehta AD, Shpaner M, Gomez-Ramirez M, Ortigue S, et al. Audio-visual multisensory integration in superior parietal lobule revealed by human intracranial recordings. Journal of Neurophysiology 2006;96:721–729. [PubMed: 16687619]

Munhall KG, Kroos C, Jozan G, Vatikiotis-Bateson E. Spatial frequency requirements for audiovisual speech perception. Perception & Psychophysics 2004;66:574–583. [PubMed: 15311657]

Murray EA, Bussey TJ, Saksida LM. Visual perception and memory: A new view of medial temporal lobe function in primates and rodents. Annual Review of Neuroscience 2007;30:99–122.

Nagy A, Eordegh G, Paroczy Z, Markus Z, Benedek G. Multisensory integration in the basal ganglia. European Journal of Neuroscience 2006;24:917–924. [PubMed: 16930419]

Narain C, Scott SK, Wise RJ, Rosen S, Leff A, Iversen SD, et al. Defining a left-lateralized response specific to intelligible speech using fMRI. Cerebral Cortex 2003;13:1362–1368. [PubMed: 14615301]

Obleser J, Wise RJ, Alex Dresner M, Scott SK. Functional integration across brain regions improves speech perception under adverse listening conditions. Journal of Neuroscience 2007;27:2283–2289. [PubMed: 17329425]

Ojanen V, Mottonen R, Pekkola J, Jaaskelainen IP, Joensuu R, Autti T, et al. Processing of audiovisual speech in Broca's area. Neuroimage 2005;25:333–338. [PubMed: 15784412]

Olson IR, Gatenby JC, Gore JC. A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. Brain Research, Cognitive Brain Research 2002;14:129–138. [PubMed: 12063136]

Pickett ER, Kuniholm E, Protopapas A, Friedman J, Lieberman P. Selective speech motor, syntax and cognitive deficits associated with bilateral damage to the putamen and the head of the caudate nucleus: A case study. Neuropsychologia 1998;36:173–188. [PubMed: 9539237]

Raij T, Uutela K, Hari R. Audiovisual integration of letters in the human brain. Neuron 2000;28:617–625. [PubMed: 11144369]

Rauschecker JP, Tian B. Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proceedings of the National Academy of Sciences, U.S.A 2000;97:11800–11806.

Rissman J, Gazzaley A, D'Esposito M. Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage 2004;23:752–763. [PubMed: 15488425]

Rizzolatti G, Arbib MA. Language within our grasp. Trends in Neurosciences 1998;21:188–194. [PubMed: 9610880]

Rodd JM, Davis MH, Johnsrude IS. The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. Cerebral Cortex 2005;15:1261–1269. [PubMed: 15635062]

Rotenberg A, Muller RU. Variable place–cell coupling to a continuously viewed stimulus: Evidence that the hippocampus acts as a perceptual system. Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences 1997;352:1505–1513.

Santi A, Servos P, Vatikiotis-Bateson E, Kuratate T, Munhall K. Perceiving biological motion: Dissociating visible speech from walking. Journal of Cognitive Neuroscience 2003;15:800–809. [PubMed: 14511533]

Scheich H, Baumgart F, Gaschler-Markefski B, Tegeler C, Tempelmann C, Heinze HJ, et al. Functional magnetic resonance imaging of a human auditory cortex area involved in foreground–background decomposition. European Journal of Neuroscience 1998;10:803–809. [PubMed: 9749748]

Scott SK, Blank CC, Rosen S, Wise RJ. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 2000;123:2400–2406. [PubMed: 11099443]

Scott SK, Johnsrude IS. The neuroanatomical and functional organization of speech perception. Trends in Neurosciences 2003;26:100–107. [PubMed: 12536133]

Sekiyama K, Kanno I, Miura S, Sugita Y. Auditory-visual speech perception examined by fMRI and PET. Neuroscience Research 2003;47:277–287. [PubMed: 14568109]

Shomstein S, Yantis S. Control of attention shifts between vision and audition in human cortex. Journal of Neuroscience 2004;24:10702–10706. [PubMed: 15564587]

Shomstein S, Yantis S. Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. Journal of Neuroscience 2006;26:435–439. [PubMed: 16407540]

Skipper JI, Nusbaum HC, Small SL. Listening to talking faces: Motor cortical activation during speech perception. Neuroimage 2005;25:76–89. [PubMed: 15734345]

Skipper JI, van Wassenhove V, Nusbaum HC, Small SL. Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. Cerebral Cortex 2007;17:2387–2399. [PubMed: 17218482]

Strange BA, Otten LJ, Josephs O, Rugg MD, Dolan RJ. Dissociable human perirhinal, hippocampal, and parahippocampal roles during verbal encoding. Journal of Neuroscience 2002;22:523–528. [PubMed: 11784798]

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 1954;26:212–215.

Sun FT, Miller LM, D'Esposito M. Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data. Neuroimage 2004;21:647–658. [PubMed: 14980567]

Uppenkamp S, Johnsrude IS, Norris D, Marslen-Wilson W, Patterson RD. Locating the initial stages of speech–sound processing in human temporal cortex. Neuroimage 2006;31:1284–1296. [PubMed: 16504540]

Van Hoesen GW. Anatomy of the medial temporal lobe. Magnetic Resonance Imaging 1995;13:1047–1055. [PubMed: 8750316]

van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences, U.S.A 2005;102:1181–1186.

Vandenbulcke M, Peeters R, Dupont P, Van Hecke P, Vandenberghe R. Word reading and posterior temporal dysfunction in amnestic mild cognitive impairment. Cerebral Cortex 2007;17:542–551. [PubMed: 16603712]

Weissman DH, Warner LM, Woldorff MG. The neural mechanisms for minimizing cross-modal distraction. Journal of Neuroscience 2004;24:10941–10949. [PubMed: 15574744]

Wilson SM, Saygin AP, Sereno MI, Iacoboni M. Listening to speech activates motor areas involved in speech production. Nature Neuroscience 2004;7:701–702.

Wise RJ, Scott SK, Blank SC, Mummery CJ, Murphy K, Warburton EA. Separate neural subsystems within "Wernicke's area". Brain 2001;124:83–95. [PubMed: 11133789]

Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cerebral Cortex 2003;13:1034–1043. [PubMed: 12967920]

Wu CT, Weissman DH, Roberts KC, Woldorff MG. The neural circuitry underlying the executive control of auditory spatial attention. Brain Research 2007;1134:187–198. [PubMed: 17204249]

Yeterian EH, Pandya DN. Corticostriatal connections of the superior temporal region in rhesus monkeys. Journal of Comparative Neurology 1998;399:384–402. [PubMed: 9733085]

Zatorre RJ. Do you see what I'm saying? Interactions between auditory and visual cortices in cochlear implant users. Neuron 2001;31:13–14. [PubMed: 11498046]

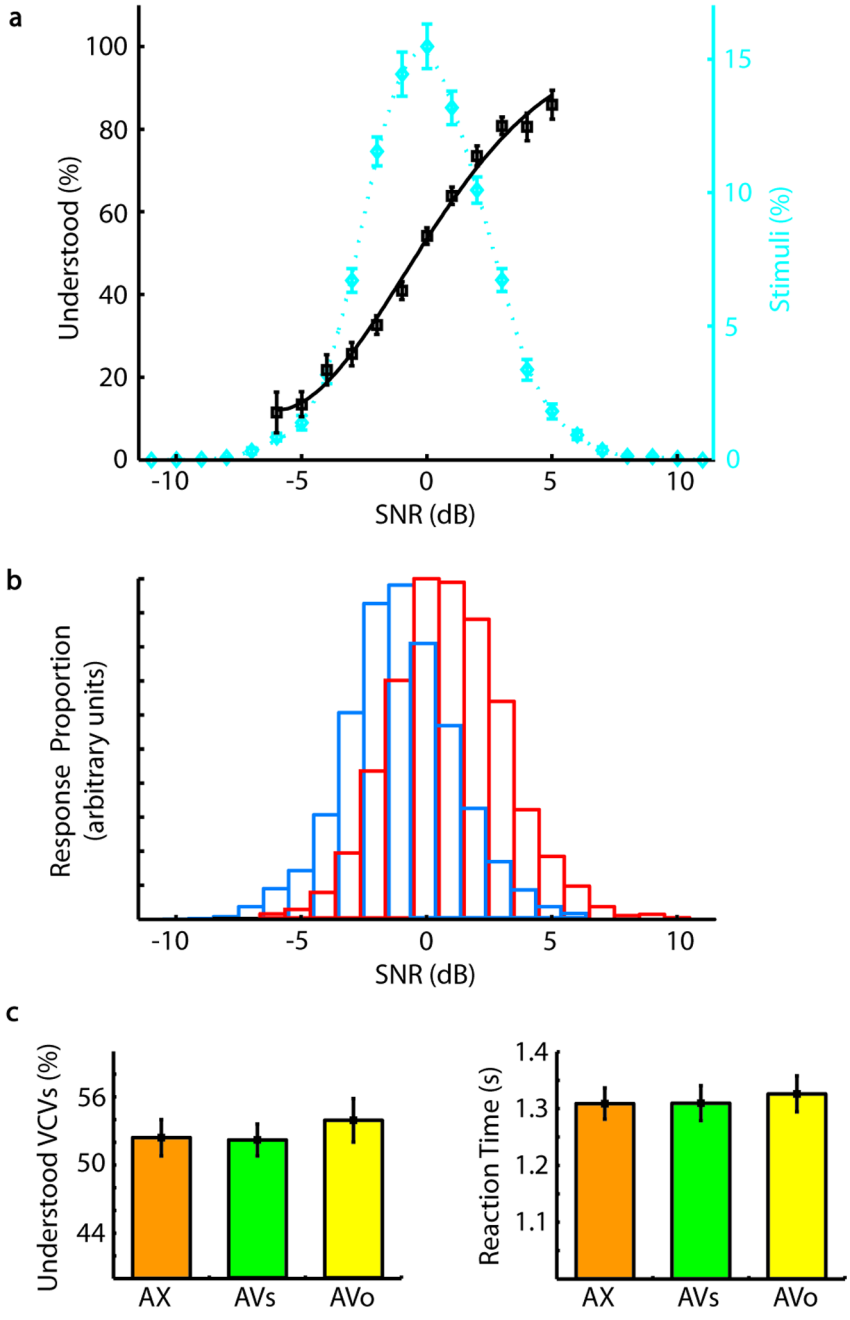

Figure 1.

Behavioral measures. (A) The mean percentage of *understood* responses (black squares) versus signal-to-noise ratio (SNR) is fit with a Weibull function (black, solid line). The likelihood of a subject *understanding* speech increases with increasing SNR. Additionally, a plot of the mean percentage of total stimuli versus SNR (blue diamonds) is fit with a smoothing spline function (blue, dotted line), revealing that most stimuli were presented within ±4 dB of threshold. SNR values are mean centered within subject, condition, and VCV. Axis label 0 dB corresponds to −4.71 dB relative to background babble. (B) Distribution of *understood* (red) and *heard* (blue) responses versus SNR. Overlapping regions are marked with blue and red edged boxes. Although sharply distinct perceptually, the SNRs of *heard* and *understood* speech show a high

degree of overlap. (C) The mean percentage of *understood* responses and mean reaction times are plotted for the audio only (AX, orange), audiovisual synchronous (AVs, green), and audiovisual offset (AVo, yellow) stimulus conditions. Reaction times are relative to trial onset and are based on 24 subjects. All other panels are based on 25 subjects. Stimuli that did not elicit a response (i.e., misses) were not included in the stimulus count. Error bars reflect standard error.

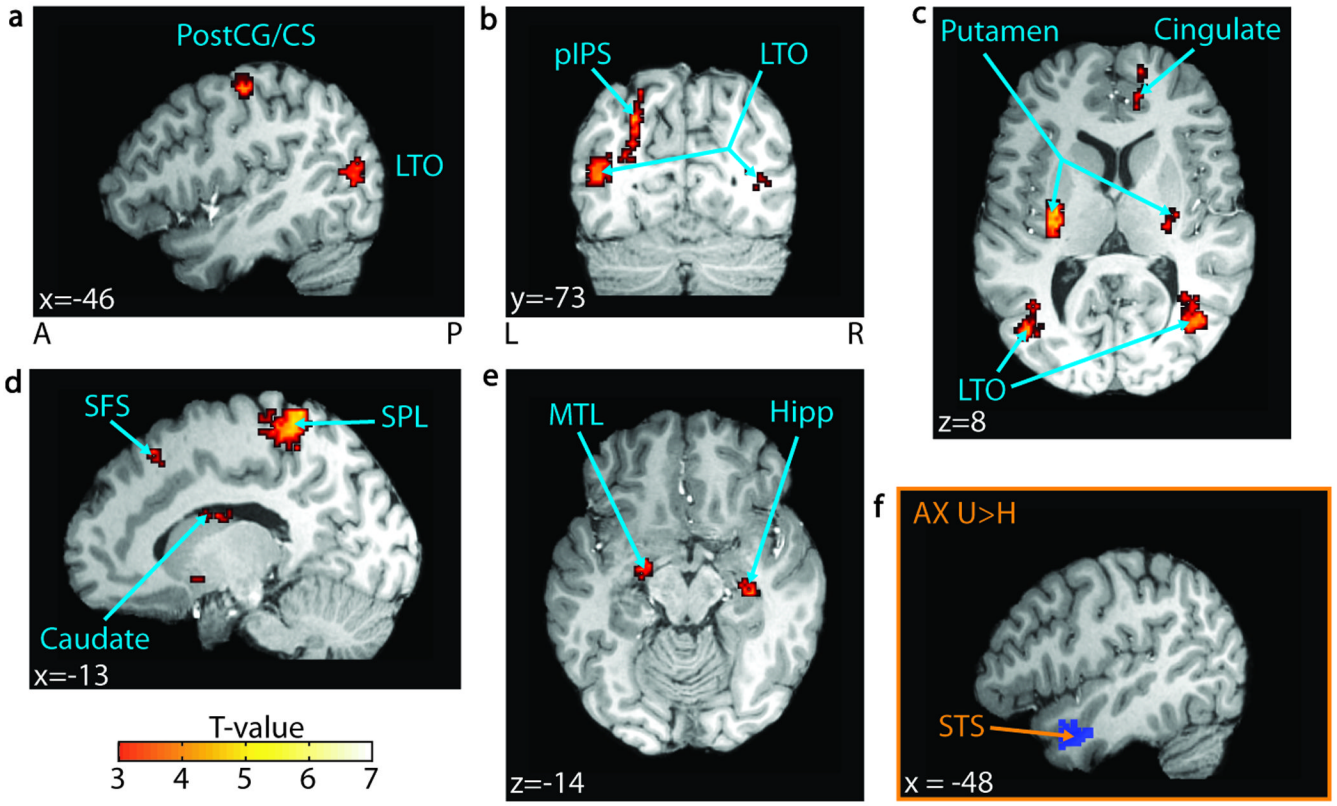

Figure 2.

Understanding versus *hearing*. (A–E) *t* Values for regions significant in a group-level *t* test for the *understood* > *heard* contrast across all conditions (FDR $p < .05$, $n = 25$) are overlaid on a representative brain. The significant regions include (A) the left postcentral gyrus (PostCG) and the central sulcus (CS); (B) the left posterior intraparietal sulcus (pIPS) and the bilateral lateral temporal–occipital boundary (LTO); (C) the bilateral putamen and the right anterior cingulate sulcus (Cingulate); (D) the left superior frontal sulcus (SFS), the left superior parietal lobule (SPL), and the left body of the caudate nucleus; and (E) the left medial temporal lobe (MTL) and the right hippocampus (Hipp). Additionally, the left superior temporal sulcus (STS) showed increased BOLD activity in the *understood* > *heard* contrast in the audio-only condition ($p < .005$, $n = 25$). (F) The left STS seed region used in the connectivity analysis, defined by the *understood* > *heard* contrast in the audio-only condition ($p < .05$, $n = 25$), is overlaid on a single subject's brain. MNI coordinates are reported in white text in each panel. A = anterior; P = posterior; L = left; R = right.

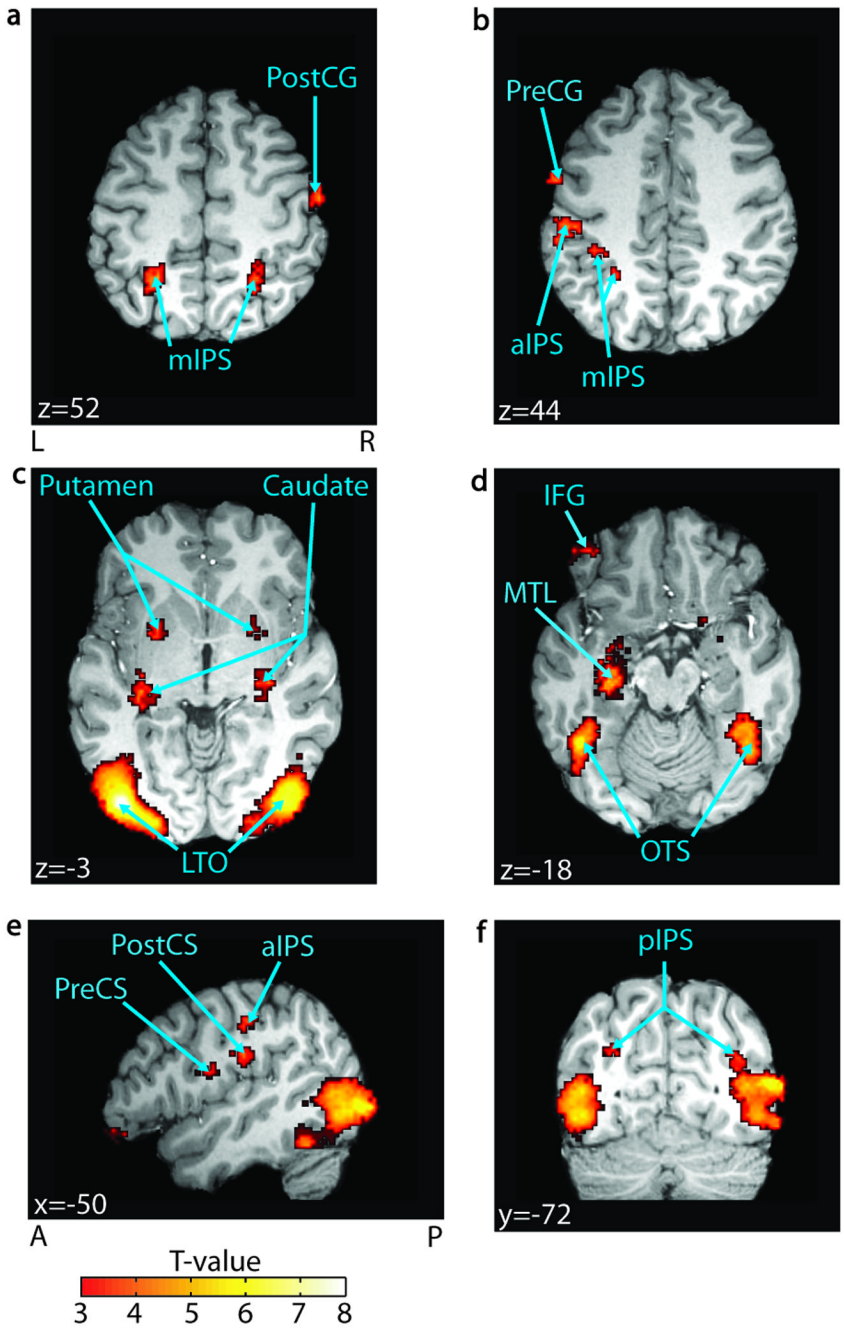

Figure 3.

Cross-modal integration. Regions significant in a block-level contrast of AVs > AVo (FDR $p < .05$, $n = 25$) are overlaid on a representative brain. Significant regions include (A) the right postcentral gyrus (PostCG) and the bilateral middle intraparietal sulcus (mIPS); (B) the precentral gyrus (PreCG) and the anterior/ middle IPS (aIPS/mIPS); (C) the bilateral putamen, the tail of the caudate nucleus (caudate), and the lateral temporal–occipital boundary (LTO); (D) the left inferior frontal gyrus (IFG), the left medial temporal lobe (MTL), and the bilateral occipito-temporal sulcus (OTS); (E) the left precentral sulcus (PreCS) and the left postcentral sulcus (PostCS); and finally, (F) the bilateral posterior IPS (pIPS). MNI coordinates are reported in white text in each panel. A = anterior; P = posterior; L = left; R = right.

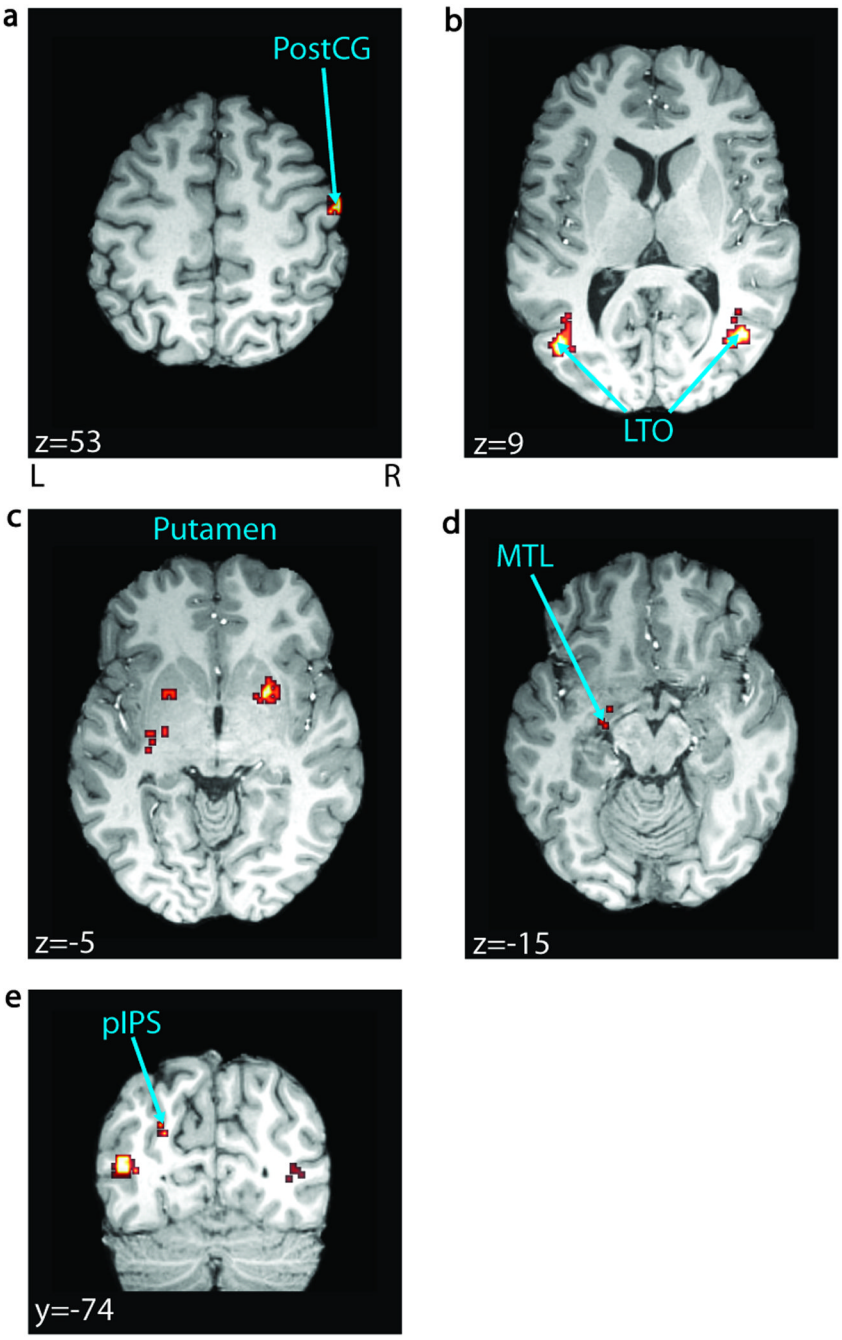

Figure 4.

Audiovisual contribution to *understanding*. Regions involved in applying synchronous visual information to improve intelligibility, identified by conjoining significant voxels in the AVs > AVo (FDR $p < .05$, $n = 25$) and *understood* > *heard* contrasts across all conditions (FDR $p < .05$, $n = 25$) are overlaid on a representative brain. Notably, this role was restricted to a subset of our general network for *understanding* including (A) the right postcentral gyrus (PostCG), (B) bilateral lateral temporal–occipital boundary (LTO), (C) bilateral putamen, (D) left medial temporal lobe (MTL), and (E) left posterior intraparietal sulcus (pIPS). MNI coordinates are reported in white text in each panel. L = left; R = right.

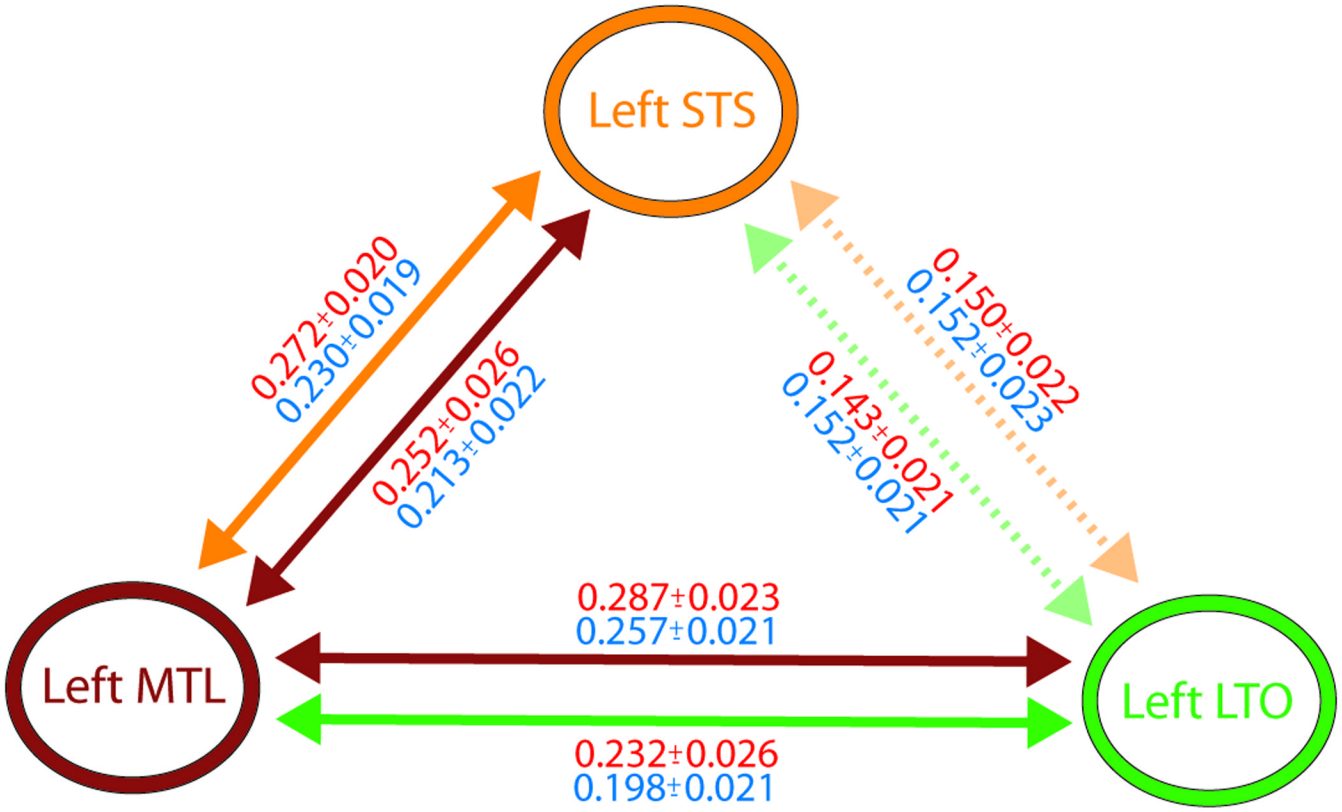

Figure 5.

Functional connectivity in the left temporal cortex. A schematic of connectivity between the left superior temporal sulcus (STS), the left medial temporal lobe (MTL), and the left temporal–occipital boundary (LTO), including correlation coefficients between regions, is shown here. Solid colored arrows indicate increased connectivity for *understood* relative to *heard* between two regions using the color-coded region as a seed. For example, the orange arrow between the left STS and the left MTL indicates that there were voxels in the left MTL that exhibited increased connectivity with the left STS seed for *understanding* relative to *heard*. Correlation coefficients (mean ± *SEM*) between regions for the *understood* (red) and *heard* (blue) percepts are reported proximal to the corresponding arrow. Correlation coefficients were averaged across voxels showing a significant ($p < .05$, $n = 25$) increase in connectivity with the seed region in SPM (see Methods). Broken and faded arrows indicate that no voxels showed increased connectivity with the seed region. In such cases, correlation coefficients were averaged across the entire ROI used for connectivity seed selection (see Methods for details on how ROIs were defined).

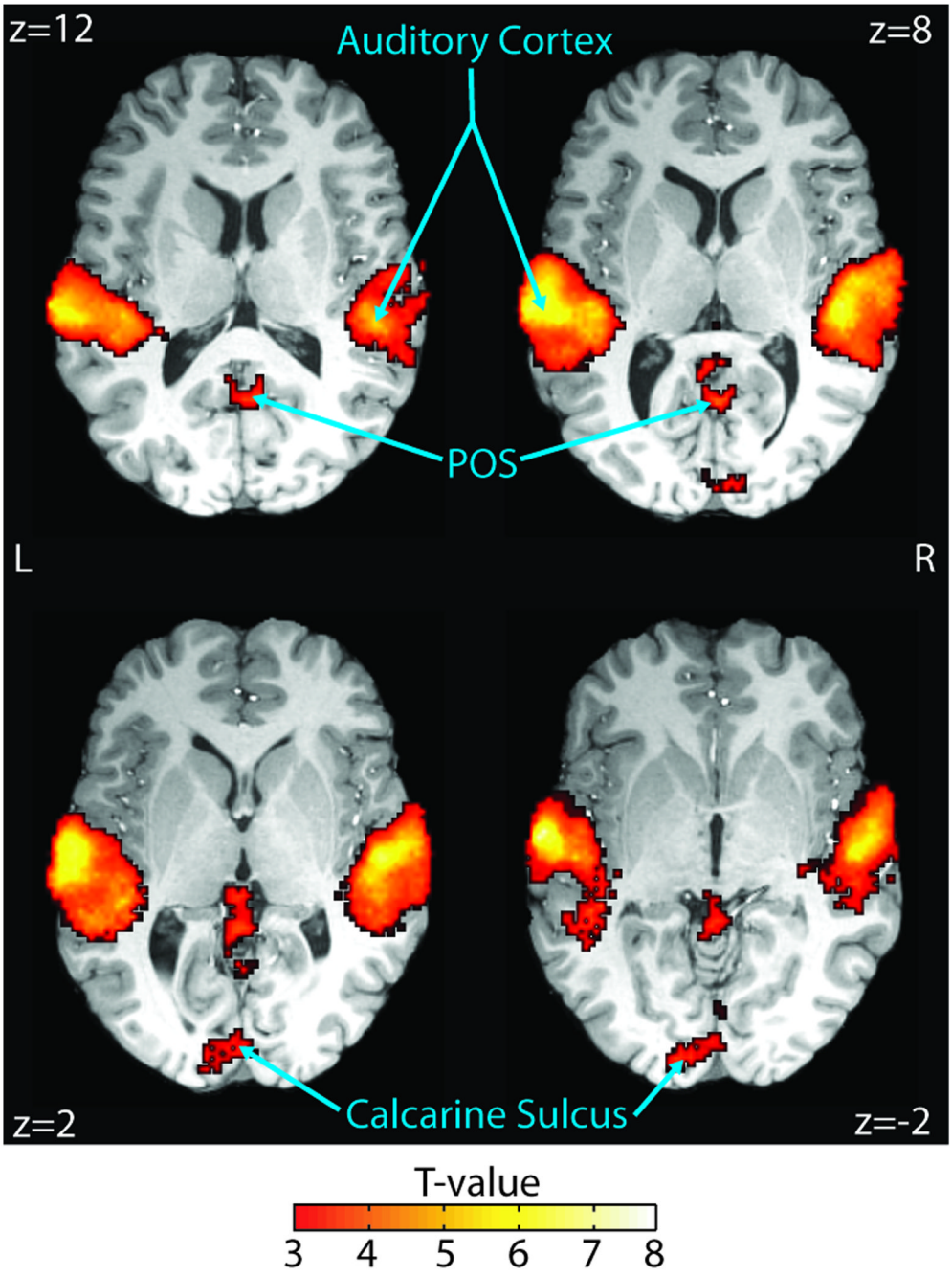

Figure 6.
Quantifying SNR sensitivity. Regions significant in a group-level *t* test of SNR regressor beta estimates (see Methods) are overlaid on a representative brain. The bilateral auditory cortex, including Heschl's gyri, superior temporal gyri/sulci, and left middle temporal gyrus, were significantly sensitive (FDR $p < .05$, $n = 25$) to SNR (FDR $p < .05$, $n = 25$). Importantly, BOLD activity in our *understanding* network is not significantly explained by the small SNR variance. MNI coordinates are reported in white text in each panel. L = left; R = right.

Table 1

MNI Coordinates of Significant Regions in Linear Contrasts

	x	y	z	t	p*
Audiovisual Synchronous >Audiovisual Offset**					
Left LTO	−42	−84	−2	7.87	0.001
Left Occipito-Temporal Sulcus	−44	−52	−16	5.79	0.004
Left Precentral Sulcus	−34	6	20	6.30	0.003
Left MTL	−28	−20	−14	5.71	0.004
Left Anterior IPS	−52	−24	42	4.42	0.013
Left Middle IPS	−24	−50	52	4.85	0.009
Left Middle IPS	−36	−38	36	4.25	0.015
Left Posterior IPS	−30	−68	28	4.10	0.018
Left Inferior Frontal Gyrus	−38	48	−16	4.05	0.019
Left Anterior Putamen	−20	4	−6	4.36	0.014
Left Tail of Caudate	−34	−24	−4	5.11	0.007
Left Precentral Gyrus	−60	2	46	4.84	0.009
Left Postcentral Sulcus	−52	−20	24	4.40	0.013
Right LTO/Posterior IPS	44	−86	−2	6.27	0.003
Right Occipito-Temporal Sulcus	44	−58	−18	5.10	0.007
Right Middle IPS	28	−50	54	4.45	0.012
Right Anterior Putamen	28	4	−6	4.14	0.017
Right Posterior Putamen	30	−22	0	4.94	0.008
Right Ventral Putamen	30	4	−14	3.91	0.023
Right Tail of Caudate/Posterior Putamen	32	−30	−2	3.82	0.025
Right Tail of Caudate	26	−34	4	4.33	0.014
Right Postcentral Gyrus	60	−10	56	4.69	0.010
Understood > Heard (All Conditions)**					
Left LTO	−44	−72	8	4.54	0.039
Left MTL	−28	−10	−16	3.78	0.048
Left Posterior IPS	−26	−72	38	5.11	0.036
Left Ventral Putamen	−20	6	−8	4.49	0.040

	x	y	z	t	p*
Left Posterior Putamen	−28	−18	2	6.01	0.036
Left Body of Caudate	−16	−6	24	4.65	0.038
Left Postcentral Gyrus/Central Sulcus	−54	−16	54	5.24	0.036
Left Superior Frontal Sulcus	−14	26	50	4.49	0.040
Left SPL	−20	−42	62	6.54	0.036
Right LTO	38	−68	8	4.95	0.036
Right Hippocampus	30	−20	−16	4.96	0.036
Right Postcentral Gyrus/Central Sulcus	52	−8	52	5.09	0.036
Right Posterior Putamen	26	−20	8	4.31	0.041
Right Dorsal Putamen	28	−12	14	5.35	0.036
Right Ventral Putamen	24	4	−6	5.76	0.036
Right Medial Frontal Cortex	16	56	12	4.83	0.037
Right Cingulate Sulcus	12	42	10	4.65	0.038
SNR**					
Left Heschl's/STS/MTG/STG	−64	−8	−2	7.98	0.001
Left Lateral Anterior STS/MTG/STG	−60	4	−16	7.41	0.001
Left Lingual Gyrus	−2	−82	−16	4.93	0.004
Left Calcarine Sulcus	−2	−90	0	4.33	0.009
Right Heschl's Gyrus/STG/STS	56	−18	4	7.08	0.001
Right Parieto-occipital Sulcus	4	−62	10	5.11	0.003
Right Lingual Gyrus	4	−76	−6	3.76	0.023

*FDR-corrected p values.

**Only clusters greater than 20 voxels in size are reported.

All coordinates are reported in MNI space (mm).