Polysensory Interactions along Lateral Temporal Regions Evoked by Audiovisual Speech

Tarra M. Wright[1], Kevin A. Pelphrey[1,2], Truett Allison[3], Martin J. McKeown[1] and Gregory McCarthy[1,4]

[1]Brain Imaging and Analysis Center, Duke University Medical Center, [2]Neurodevelopmental Disorders Research Center, Department of Psychiatry, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, NC, [3]Department of Neurology, Yale University School of Medicine, New Haven, CT and [4]Department of Veterans Affairs Medical Center, Durham, NC, USA

Many socially significant biological stimuli are polymodal, and information processing is enhanced for polymodal over unimodal stimuli. The human superior temporal sulcus (STS) region has been implicated in processing socially relevant stimuli — particularly those derived from biological motion such as mouth movements. Single unit studies in monkeys have demonstrated that regions of STS are polysensory — responding to visual, auditory and somatosensory stimuli, and human neuroimaging studies have shown that lip-reading activates auditory regions of the lateral temporal lobe. We evaluated whether concurrent speech sounds and mouth movements were more potent activators of STS than either speech sounds or mouth movements alone. In an event-related fMRI study, subjects observed an animated character that produced audiovisual speech and the audio and visual components of speech alone. Strong activation of the STS region was evoked in all three conditions, with greatest levels of activity elicited by audiovisual speech. Subsets of activated voxels within the STS region demonstrated overadditivity (audiovisual > audio + visual) and underadditivity (audiovisual < audio + visual). These results confirm the polysensory nature of STS region and demonstrate for the first time that polymodal interactions may both potentiate and inhibit activation.

Introduction

Many socially relevant stimuli are polymodal, and research points to information processing enhancements for polymodal over unimodal stimuli (Cotton, 1935; Sumby and Pollack, 1954; Welch and Warren, 1986; Summerfield, 1987; Stein and Meredith, 1993; Calvert *et al.*, 1998). Such enhancements are not necessarily additive; rather, inputs in different modalities can have multiplicative effects that are not predicted simply from knowledge of the components (Stein and Meredith, 1993). Functional neuroimaging studies in humans indicate that the superior temporal sulcus (STS) region plays a role in processing biological motion including the facial feature movements that are involved in visual speech [(Grafton *et al.*, 1996; Howard *et al.*, 1996; Calvert *et al.*, 1997; Puce *et al.*, 1998; Grossman *et al.*, 2000; Grèzes *et al.*, 2001; Bernstein *et al.*, 2002); reviewed by Allison *et al.* (Allison *et al.*, 2000)]. This region is also involved in signaling the social significance of biological motions (Campbell *et al.*, 2001; Winston *et al.*, 2002; Pelphrey *et al.*, 2003).

Functional neuroimaging studies in humans have suggested several candidate regions for polymodal integration. Calvert *et al.* (Calvert *et al.*, 1997) identified portions of the STS and superior temporal gyrus (STG) that were activated by heard speech and silent lip-reading, and surmised that portions of this activation overlapped with primary auditory cortex (Bernstein *et al.*, 2002). Calvert *et al.* (Calvert *et al.*, 1999) found that congruent nonspeech audiovisual stimuli enhanced activity in auditory [Brodmann Area (BA) 41/42] and motion-sensitive visual (MT/V5) cortices as compared with unimodal presentations (audio or visual alone). Calvert *et al.* (Calvert *et al.*, 2000) utilized audiovisual speech to identify a cluster of voxels in the left STS as a site of polysensory integration. Similarly, Calvert *et al.* (Calvert *et al.*, 2001) identified certain nonsensory as well as sensory cortices that showed overadditive and underadditive responses to congruent and incongruent nonspeech audio and visual stimuli, respectively. In contrast, Olson *et al.* (Olson *et al.*, 2002) demonstrated that the STS/STG region, while responding more strongly to audiovisual speech than to visual speech alone, did not discriminate between synchronized and unsynchronized audiovisual speech, suggesting that the STS is not involved in integrating audio and visual components of speech. Instead, they identified the claustrum as a site of possible polysensory integration.

The studies reviewed above suggest that sensory-specific cortices can be involved in polymodal processing, showing an enhanced response for congruent stimuli and a depressed response for incongruent stimuli. However, 'cross-modal inhibition' is possibly another mechanism for enhancing the response to a sensory stimulus. For instance, using functional magnetic resonance imaging (fMRI), Laurienti *et al.* (Laurienti *et al.*, 2002) demonstrated that a visual (or auditory) stimulus could activate its matching sensory cortex and simultaneously inhibit activity in a nonmatching sensory cortex. It is not yet known whether cross-modal inhibition plays a role in audiovisual speech perception.

Here we report the results of an event-related fMRI experiment of audiovisual speech. Subjects were presented with three randomly intermixed conditions in which a three-dimensional animated figure: (i) moved her mouth concurrently with audible speech; (ii) moved her mouth similarly but without speech; and (iii) did not move her mouth but audible speech was heard. We sought to confirm previous studies of activation in response to mouth movements [e.g. (Puce *et al.*, 1998)]; to investigate the time courses of audio and visual speech-evoked activations; to confirm whether speech sounds with congruent mouth movements were more potent activators of STS and STG than either speech sounds or mouth movements in isolation; to explore the distribution of sensory-specific and polysensory regions along lateral temporal cortex; and to identify potential cross-modal inhibitory interactions between sensory-specific cortices involved in audiovisual speech perception.

Materials and Methods

Subjects

Twelve right-handed healthy subjects (seven females, five males), age range 19–29 years (mean age 23 years), provided written informed consent to participate in a study approved by the Duke University Medical Center Institutional Review Board. All subjects had normal or corrected to normal visual acuity and were paid for participating.

Experimental Stimuli

We created three stimulus conditions using the Poser 4.0® software program (Curious Labs Inc., Santa Cruz, CA). In each, an animated female character was presented from the shoulders up with eyes forward and mouth closed (see Fig. 1). In the first condition, the character moved her mouth concurrently with audible speech (Audiovisual). In the second condition, the figure did not move her mouth yet audible speech occurred (Audio) (i.e. words were heard by the subjects). In the final condition, the character moved her mouth but without audible speech (Visual). The character's vocabulary consisted of 45 monosyllabic words (e.g. 'cat', 'dog', 'flag'). Over 135 trials, subjects heard each vocabulary word twice, once with accompanying mouth movements and once without. Similarly, subjects saw each articulation twice, once with accompanying audible speech and once without. The character's mouth motions and speech sounds were coordinated using the Mimic® software program (LIPSinc Inc., Morrisville, NC) to simulate natural articulation. Each stimulus event occurred over a 1 s duration, and trials were separated by a 21 s intertrial interval (ITI) during which the character was presented with eyes forward and mouth closed.

We used CIGAL to control stimulus presentation. Using an LCD projector (XGA resolution, 900 lumens), stimuli were back-projected upon a translucent screen (~56 cm × 66 cm) placed at the subject's feet. Subjects viewed the stimuli through custom glasses with angled mirrors. Audio stimuli were presented using MR-compatible earphones (Resonance Technology, Los Angeles, CA). We instructed subjects to attend to the screen at all times and to listen carefully. Trials were randomized within runs lasting ~5 min (15 trials per run). We encouraged subjects to complete nine runs. Eleven subjects completed nine and one completed eight runs, for an average of 8.92 runs (134 trials) per subject.

Data Acquisition

MRI scanning was performed on a General Electric 4T LX NVi scanner system equipped with 41 mT/m gradients, and using a birdcage radio frequency (RF) head coil for transmit and receive (General Electric, Milwaukee, Wisconsin). Sagittal T_1-weighted localizer images were first acquired and used to define a target volume for a semi-automated high-order shimming program. After shimming, the anterior commissure (AC) and posterior commissure (PC) were identified in the mid-sagittal slice and used as landmarks for the prescription of blood oxygen-level dependent (BOLD) contrast images. A series of 60 high-resolution coronal T_1-weighted images [repetition time (T_R) = 450 ms; echo time (T_E) = 20 ms; field of view (FOV) = 24 cm; image matrix = 256^2; slice thickness = 5 mm; in-plane resolution = 0.9375 mm^2] was acquired along the AC–PC line. The T_1-weighted images were used to select 20 contiguous 5 mm coronal slices for functional imaging. Slices were acquired from posterior to anterior along the intercommissural line such that the 20th slice was anchored at the AC (see Fig. 2*A*). Functional images were collected using a spiral imaging sequence sensitive to BOLD contrast (T_R = 1.5 s; T_E = 30 ms; FOV = 24 cm; image matrix = 64^2; flip angle = 62°; slice thickness = 5 mm; in-plane resolution = 3.75 mm). Each imaging run began with five discarded RF excitations to allow for steady-state equilibrium.

Data Analyses

Our data analytic strategy followed closely that taken in prior studies from our laboratory [e.g. (Jha and McCarthy, 2000; Yamasaki *et al.*, 2002; Pelphrey *et al.*, 2003)]. This strategy involved a hypothesis-driven anatomical regions of interest (ROI) approach supplemented with exploratory voxel-based analyses.

The centroid of activation for each functional image volume within each time series was computed and plotted for each subject and imaging run. No subject had greater than a 3 mm deviation in the center of activation in the *x*, *y* or *z* dimensions. The MR signal for each voxel was temporally aligned to correct for the interleaving of slice acquisition within each 1.5 s T_R. Temporal alignment was accomplished by fitting the time series of each voxel with a cubic spline and then resampling this function for all voxels at the onset of each T_R. Epochs time-locked to the stimulus onset were extracted from the continuous time series and averaged according to trial type, with the temporal order relative to stimulus onset maintained. The averaged epochs consisted of the four image volumes preceding (–6 to –1.5 s) and the nine image volumes following (1.5 to 13.5 s) the onset (0 s) of each stimulus event for 14 image volumes or 21 s of functional data. The averaged MR signal time-epochs were used in the analytic procedures described below.

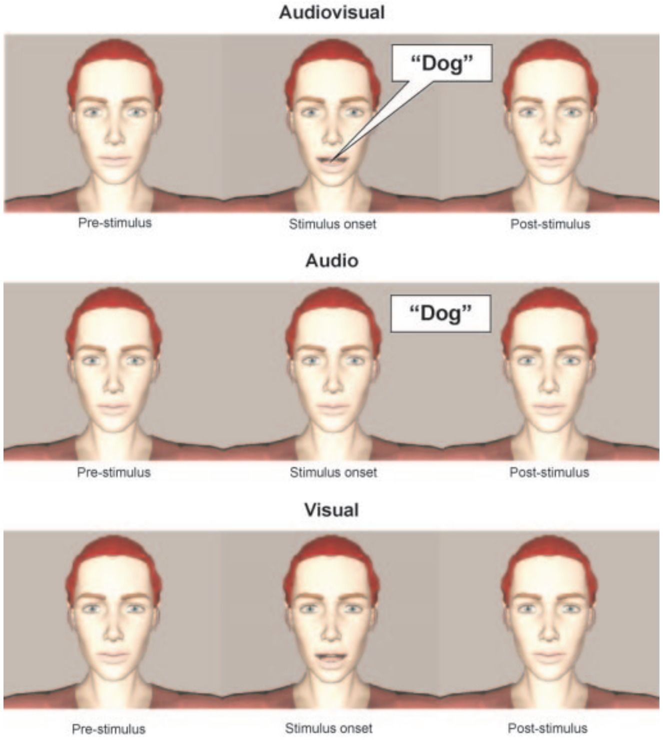

Figure 1. There were three stimulus conditions. In Audiovisual, the animated character moved her mouth with concurrent speech. In Audio, the figure did not move her mouth but audible speech was heard. In Visual, the character's mouth moved similarly to Audiovisual, but audible speech was not heard.

Anatomical ROI

A priori ROI selection was based on pertinent literature [e.g. (Calvert *et al.*, 1997, 2000, 2001; Schlosser *et al.*, 1998)]. Two research assistants who were blind to the subsequent statistical analyses of the data drew ROI on each subject's high-resolution anatomical images. ROI were traced on the left and right intraparietal sulci (IPS), superior temporal sulci (STS), superior temporal gyri (STG), and middle temporal gyri (MTG). Identification of anatomical landmarks and ROI was guided by human brain atlases (Roberts *et al.*, 1987; Mai *et al.*, 1997; Duvernoy, 1999). For each of the eight ROI (four anatomical areas by two hemispheres), labels indicated the distance (in mm) posterior from the AC, facilitating registration of activity from similar ROI across subjects (see Fig. 2*A*). The STS was traced on 12 slices ranging from 0 to 60 mm posterior from the AC. The STG was drawn in parallel fashion to the STS, with 12 slices ranging from 0 to 60 mm posterior from the AC. The IPS was outlined on 11 slices ranging from 40 to 95 mm posterior from the AC. The MTG was outlined on 14 slices ranging from 0 to 70 mm posterior from the AC. In sum, each subject contributed 98 ROI tracings. The average sizes (in functional voxels) of the ROI were: STG = 162; STS = 299; IPS = 237; and MTG = 205.

Time-activation Waveforms from Anatomical ROI

The average signal for all voxels within each ROI was computed for each of the 14 time points and plotted to visualize the time course of the mean hemodynamic response (HDR) for each ROI during each stimulus condition. The HDR time course was examined separately for each slice and hemisphere within each ROI, so that regional and stimulus-condition related effects in the form of the HDR could be evaluated. Repeated measures ANOVAs were performed to evaluate differences in HDR amplitude as a function of stimulus condition, hemisphere and distance from the AC for the average of the 4.5 s and 6.0 s time points for selected ROI. These time points correspond to the peak of the reference waveform

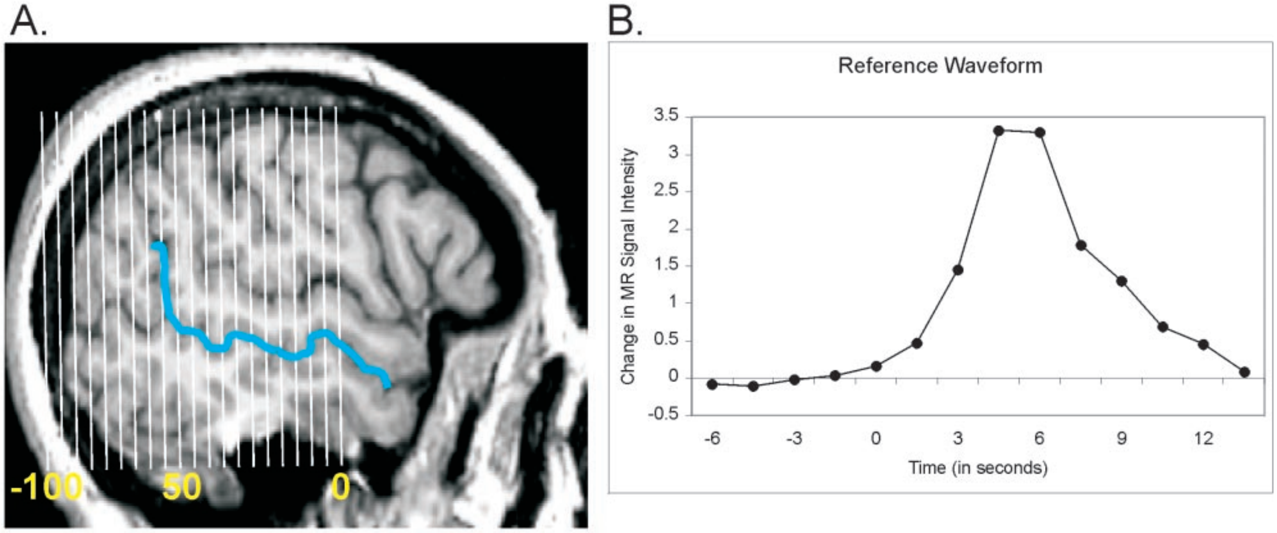

Figure 2. (*A*) For functional imaging, 20 × 5 mm coronal slices were prescribed along the AC–PC line, with the 20th slice anchored at the AC. (*B*) The reference waveform was the mean waveform representing the HDR time course within the STS (blue line) across conditions. The average MR signal value of the prestimulus baseline period was 1422 (SD = 48). Across subjects and voxels, the standard error for this period was 1.56. In this and other figures displaying HDRs, the horizontal axis displays time in seconds with stimulus onset at 0 s.

(Fig. 2*B*). Additional two-factor ANOVAs were performed to examine deviations from additivity in HDR Audio and Visual amplitudes along the STS and STG. To allow for further tests of overadditivity and underadditivity, we also computed a composite waveform by summing the average HDRs to Audio and Visual presented alone (Audio + Visual), and we compared this composite waveform to those waveforms from the original three experimental conditions.

We calculated standard deviation and standard error estimates for the measure Audio + Visual using error propagation techniques. The SD of Audio + Visual, for instance, was calculated using the formula

$$\text{SDAudio + Visual} = \sqrt{}(\text{SDAudio}^2 + \text{SDVisual}^2)$$

ROI were also used to group and count activated voxels that were identified in a correlation analysis (described below).

Time-course correlation with a reference waveform

We conducted a correlation analysis with an empirically defined reference waveform to identify subsets of Audio, Visual and Audiovisual activated voxels within the individual image volumes. The reference waveform was the mean waveform representing the average HDR time course within the STS across conditions and subjects (see Fig. 2*B*). We generated a *t* statistic for each voxel across runs by correlating the averaged (across runs) 21 s MR signal time-epochs (generated as described above) from each voxel with the reference waveform shown in Figure 2*B*. After correlating the time-activation waveform from each voxel with the reference waveform, *t* statistics were calculated from the correlation coefficients, and activated voxels were defined as those with suprathreshold *t* values, with the threshold for activation set at $t > 1.96$. Counts of activated voxels within each anatomically defined ROI were converted to percentages relative to the number of voxels in that ROI. 'Non-activated voxels', defined here as those with sub-threshold correlations, may nevertheless evince an evoked HDR that would have exceeded the threshold for activation if the signal-to-noise ratio had been improved by averaging additional trials [see (Huettel and McCarthy, 2001)], or if a different reference waveform was used. We therefore routinely examine the averaged epochs of all conditions for voxels that may only show suprathreshold activation for a single condition.

Common Space Voxel-based Analyses

To explore the extent to which populations of voxels demonstrated different patterns of activity as a function of stimulus condition and to identify possible regions of activity outside of the anatomical ROI, we performed voxel-based analyses on the group-averaged data. Across-subjects averaged functional time course volumes and *t* statistic activation maps were computed for each of the three original stimulus conditions and the calculated Audio + Visual condition, combining data from all 12 subjects. These averages were created by first spatially normalizing the data (e.g. the *t* statistic maps) from the individual subjects and then averaging these normalized data across subjects. We created the average time courses by taking the simple arithmetic mean of the 12 subjects. The group-average *t* statistic maps were created by the average *T* method as described by Lazar *et al*. (Lazar *et al*., 2001), where

$$T_A = \sum_{i=1}^{k} \frac{T_i}{\sqrt{k}}$$

Before averaging, the images were spatially normalized to a template image set from a representative subject. Alignment factors for the functional images were calculated on a slice-by-slice basis using custom software written by one of us (M.J.M.). This software implemented a non-linear optimization of translation, rotation, and stretch values (six parameters) based on the cost function of maximizing the correlation between the (low and high passed filtered) template slice and the to-be-normalized current slice. The normalization algorithm used the high-resolution anatomical images without regard to the functional data. Before normalization, the brain was extracted from each subject's anatomical images to eliminate the influence of high contrast by extraneous regions such as the skull and neck. The averaged and spatially normalized data were used to identify and interrogate unexpected regions of group-consistent positive or negative activation. The group-averaged data were also used to compare the patterns of activation observed in each stimulus condition. Activated voxels were defined as those with suprathreshold *t* values, with the threshold for activation set at $t(13) > 5.2$ ($P < 0.000171$, two-tailed, uncorrected). This *t* value survives a Bonferroni correction for 11 622 tests (i.e. the number of functional voxels in the template brain) using the Dubey and Armitage-Parmar correction for a correlation among tests of 0.30 (Sankoh *et al*., 1997). Activated voxels were displayed and archived in individual *t* statistic maps and were superimposed on the subjects' anatomical images for inspection.

Results

Anatomical ROI Analyses

Influences on the Magnitude of the HDRs along the STS and STG

Time courses for the STG and STS (averaged across hemispheres) are illustrated by stimulus condition in the insets of Figure 3*A* and *B*, respectively. In both ROI, we observed positive HDRs 3–6 s following stimulus onset (time point = 0), and the magnitude of Audiovisual appeared greater than Visual or Audio. We observed activity to Visual primarily in the STS. Positive HDRs were not observed in the MTG or IPS for any stimulus condition; thus, we do not discuss these regions further.

We examined the STG and the STS on a slice-by-slice basis. Thirty-six (three conditions by 12 slices) individual HDR waveforms are presented in each of the two main panels of Figure 3*A* and *B*. The *x*-axes represent distance in 5 mm increments from the AC; within each bin, increasing time (in seconds) is displayed from right to left (–6 s to 13.5 s). In the STG (Fig. 3*A*), HDRs were above baseline for Audio and Audiovisual in all slices, but the HDRs were greatest in the middle slices (15–35 mm). Along the extent of the STG, the average ROI response to Visual did not rise above baseline. For the STS (Fig. 3*B*), Visual HDRs (green lines) were confined to posterior portions (30–55 mm). In contrast, we observed Audiovisual HDRs (red lines) along the extent of the STS, although the largest response amplitudes were in posterior slices (35–50 mm). Audio (blue lines) evoked HDRs in both anterior (0–15 mm) and posterior (25–55 mm) slices. Only Audio and Audiovisual evoked significant HDRs in anterior slices of the STS (0–20 mm). As shown in the lower panels of Figure 3*A* and *B*, the percentages of activated voxels followed patterns of distribution similar to those observed for the magnitudes of response in the waveform analyses.

To evaluate the effects displayed in Figure 3*A* and *B*, we calculated peak amplitude scores for each subject by averaging the HDR values at the 4.5 s and 6 s time points. Using these scores, we then conducted two 3 (Condition: Audio, Audiovisual, Visual) × 2 (Hemisphere: Right versus Left) ANOVAs separately for the STG and STS. Distance from the anterior commissure (Slice) was included as a covariate in these analyses.

In the STG, Condition was significant, $F(2,570) = 41.05$, $P < 0.0005$. Pre-planned contrasts revealed that Audiovisual [$M = 6.97$ (SE = 0.39)] was greater than Audio [$M = 5.16$ (SE = 0.30)], $F(1,285) = 63.96$, $P < 0.0005$, and Visual [$M = 1.02$ (SE = 0.26)], $F(1,285) = 250.56$, $P < 0.0005$. Activity in the left hemisphere [$M = 4.90$ (SE = 0.37)] was greater than in the right hemisphere [$M = 3.88$ (SE = 0.37)], $F(1,285) = 3.94$, $P < 0.05$. The Condition × Hemisphere interaction and Slice were not significant. However, the Condition × Slice interaction was significant, $F(2,570) = 11.60$, $P < 0.0005$, suggesting that the A–P distributions of activity varied as a function of condition. To further explore this interaction, we compared the Slice and Audiovisual correlation ($r = 0.13$, $P = 0.03$) to the Slice and Audio ($r = 0.06$, $p = 0.33$) and Slice and Visual ($r = -0.15$, $P < 0.01$) correlations using Cohen and Cohen's (Cohen and Cohen 1983) method for determining the significance of the difference between two dependent correlations. The Slice and Audio correlation was significantly smaller than the Slice and Audiovisual correlation, $t(285) = -2.07$, $P < 0.05$, as was the Slice and Visual correlation, $t(285) = -4.45$, $P < 0.0005$.

In the STS, Condition was significant, $F(2,570) = 36.51$, $P < 0.0005$. Audiovisual [$M = 4.41$ (SE = 0.26)] was greater than Audio [$M = 3.15$ (SE = 0.22)], $F(1,285) = 24.04$, $P < 0.0005$ and Visual [$M = 2.0$ (SE = 0.21)], $F(1,285) = 73.58$, $P < 0.0005$. Activity was greater in the right hemisphere [$M = 3.58$ (SE = 0.28)] than in the left hemisphere [$M = 2.79$ (SE = 0.28)], $F(1,285) = 3.93$, $P < 0.05$. The Condition × Hemisphere interaction was not significant. The main effect of Slice was significant, $F(1,285) = 12.39$, $P < 0.001$. Across conditions, the positive correlation between Slice and amplitude was $r(288) = 0.203$, $P < 0.001$. The Condition × Slice interaction was marginally significant, $F(2,570) = 2.70$, $P = 0.068$. Preplanned contrasts revealed that the Condition × Slice interaction was significant for Visual versus Audiovisual, $F(1,285) = 5.27$, $P < 0.05$, but not for Audio versus Audiovisual. We compared the Slice and Visual ($r = 0.253$, $P < 0.0005$) and Slice and Audiovisual ($r = 0.12$, $P < 0.05$) correlations, and found them to differ significantly, $t(285) = 2.98$, $P < 0.005$.

To determine whether the peak amplitude values within the STG and STS exhibited over- or underadditivity, we conducted two (Audio, yes/no; Visual, yes/no) × 2 (Hemisphere: Right versus Left) ANOVAs separately for the STG and STS. Distance from the anterior commissure (Slice) was included as a covariate. We created a 'Rest' (Audio = no; Visual = no) condition for each subject by averaging the amplitude (across conditions) of the two time points (–3.0 s and –1.5 s) before stimulus onset. The important effect was the Audio × Visual interaction. In the STG, this interaction was significant, $F(1,1143) = 433.76$, $P < 0.05$, and driven by overadditivity (inset of Fig. 3*A*). In the STS, the Audio × Visual interaction was not significant. Thus, the STS overall did not deviate from additivity (inset of Fig. 3*B*).

Voxel-based Analyses

Patterns of Activation

Figure 4 presents 14 coronal images from the normalization template beginning 65 mm posterior from the AC (top left) and moving anteriorly (in 5 mm increments) to the AC (bottom right). Overlaid upon the anatomical template images are three averaged across-subjects *t* statistic or positive activation maps (see Methods), one for each condition – Audiovisual (red), Audio (blue) and Visual (green). The maps represent the average of the spatially normalized statistical significance values for the correlations (on a voxel-by-voxel basis) between the reference waveform (Fig. 2*B*) and the HDR waveform of voxels at a threshold of $t > 5.2$. Only positive activations are displayed. We observed discrete areas of positive activation in the STS and STG bilaterally (40–55 mm; framed in white squares) where Audio and Visual elicited overlapping activations (i.e. polysensory areas). Audiovisual elicited activity in these same areas. Throughout, where there was overlap of activation to Audiovisual and Audio and/or Visual, the spatial extent of activation to Audiovisual was greater than the extent to Audio and/or Visual. Audio elicited activation in the STS and the STG 5–55 mm; Audio activation was right lateralized 5–15 mm, bilateral 20–50 mm and left lateralized 55 mm. Within the STS and STG, activation to Audio generally overlapped with activation to Audiovisual. Visual activation was mostly right lateralized, with the exception of bilateral activation 45–55 mm. A subset of voxels (60–65 mm) activated to Visual (framed in yellow circles). Activations elicited by Visual (green color map) and Audio (blue color map) are displayed in sagittal orientation in the inset of Figure 4 (bottom

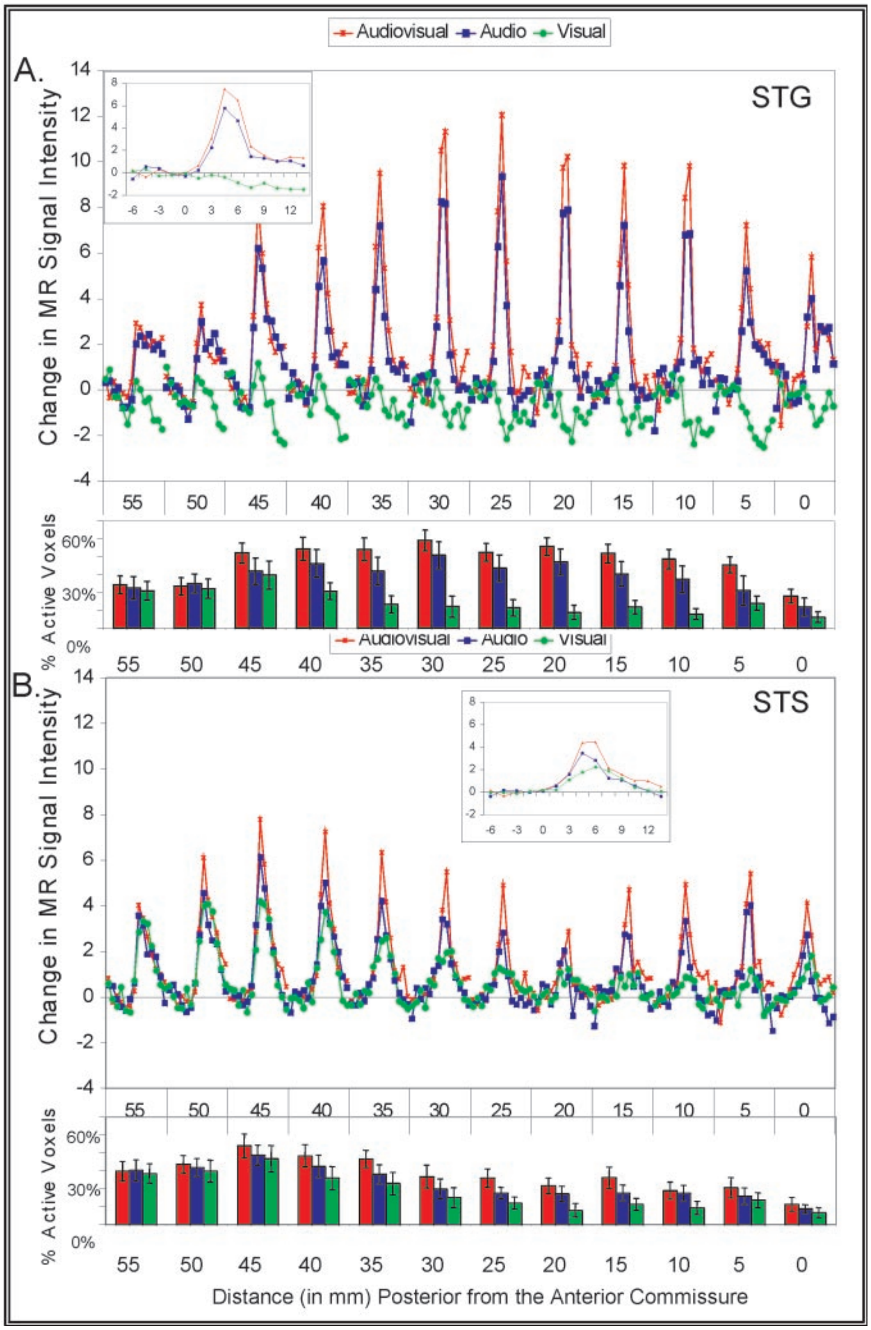

Figure 3. Thirty-six (three conditions by 12 slices) individual HDR waveforms are presented in each of the two main panels. The *x*-axes represent distance in 5 mm increments from the AC; within each bin, increasing time (in s) is displayed from right to left (–6 to 13.5 s). (*A*) HDR waveforms from all voxels of the anatomically defined STG, displayed as a function of distance posterior from the AC and stimulus condition. Inset are the mean HDR waveforms across all slices of the STG, by stimulus condition. The bottom panel shows the percentage of activated voxels in the STG on a slice-by-slice basis and by condition. (*B*) HDR waveforms from all voxels of the anatomically defined STS, displayed as a function of distance posterior from the AC and stimulus condition. Inset are the mean HDR waveforms across all slices of the STS, by stimulus condition. The bottom panel shows the percentage of activated voxels in the STS on a slice-by-slice basis and by condition.

right corner). Audio was localized to the STG and upper bank of the STS. Visual was localized to the STS and to a second area that was inferior and posterior to this region, corresponding to area MT/V5 (framed by a yellow circle). This cluster of voxels is the same area identified earlier in the coronal images (60–65 mm) and circled in yellow. Audiovisual activations were observed in the STS and STG 5–60 mm. Audiovisual was bilateral 5–45 mm and left lateralized 40–50 mm. In summary, the patterns of activation were consistent with the prior ROI analyses and indicated that Visual tended to elicit more posterior activation, localized primarily to the STS. Audio elicited activation in more anterior slices, localized primarily to the STG. Audiovisual elicited responses throughout the STG and STS.

Area MT/V5

We identified an area of activation in MT/V5 that we did not observe in the anatomical ROI analyses. Using the across-subjects functional time-course data, we interrogated that subset

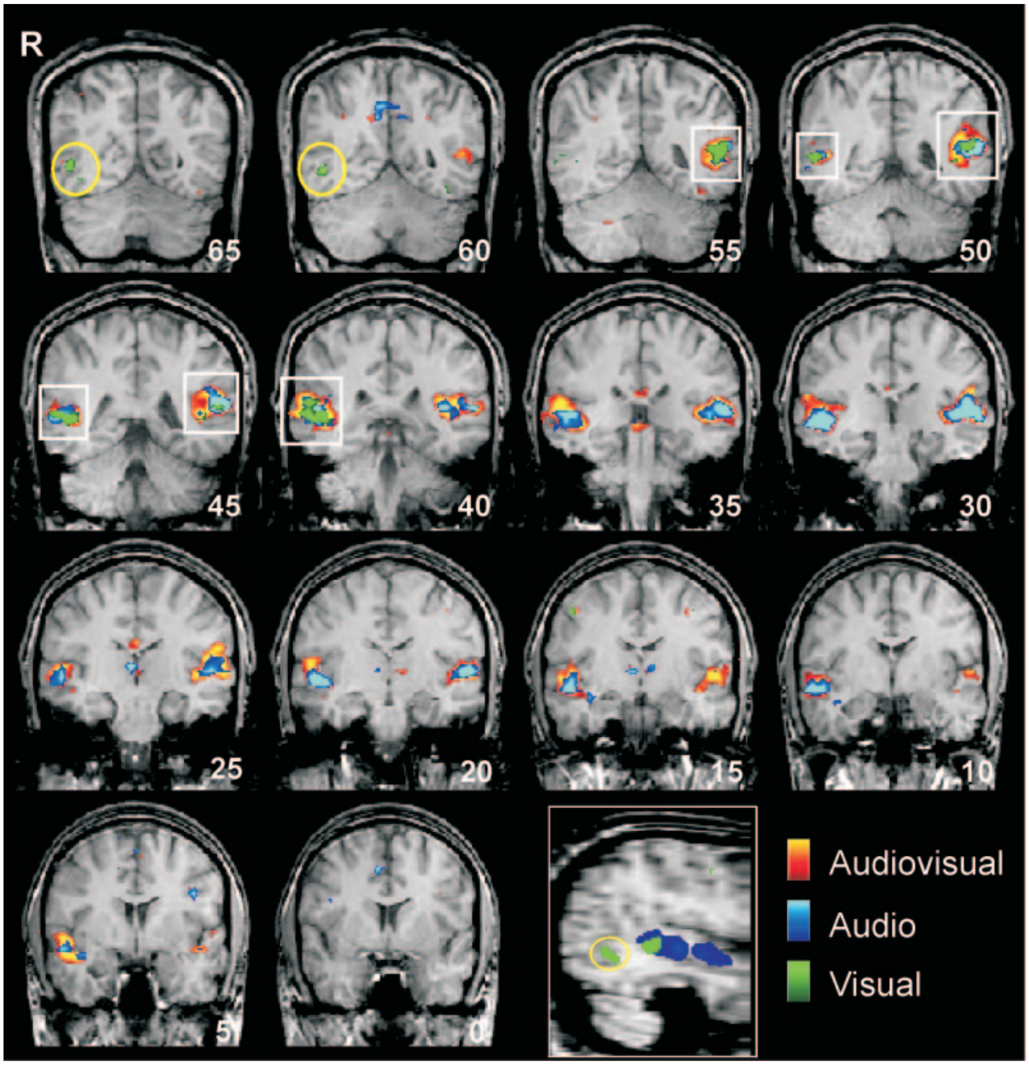

Figure 4. Across-subjects activation *t* statistic maps for the three conditions overlaid on the template subject's anatomical images. Numbers in the lower right hand corner of each image represent distance (in mm) posterior from the AC. Red, green and blue activation maps represent the magnitude of the correlation on a voxel-by-voxel basis between the reference waveform and the HDR time course for each condition. The averaged correlations were converted to *t* statistics and threshold set at $t \geq 5.2$. Voxels where Audio and Visual elicited overlapping activations are framed by white squares. Areas framed in yellow circles include a subset of voxels in area MT/V5 (60–65 mm) that activated to Visual (see also figure inset for sagittal view).

of voxels in area MT/V5 (60–65 mm, framed by yellow circles in Fig. 4). As shown in Figure 5, this area responded positively to Visual and to Audiovisual, but the HDR dropped below baseline in response to Audio.

Polysensory Cortex

In identifying areas of polysensory cortex, we were interested in areas of overlap between the Audio and Visual activations. We interrogated this intersection, which consisted of voxels in the STS and STG (Fig. 6*A*). The HDR waveforms from this analysis are presented by stimulus condition in Figure 6*B*. As can be seen, each of the three conditions produced positive HDRs. Audiovisual was greater than the response to Audio or Visual; and Audio was greater than Visual. To evaluate these potential differences, peak amplitude scores were calculated for each subject's average HDR across the voxels comprising the intersection area by averaging the HDR values at the 4.5 and 6 s time points. As illustrated in Figure 6*C*, and consistent with the ROI analyses presented previously, Audiovisual [(M = 7.91) (SD = 4.58)] was greater than Audio [(M = 5.67) (SD = 3.84)], (diff = 2.24; $t(12)$ = 3.53, P = 0.004) and Visual [(M = 3.71) (SD = 6.65)], (diff = 4.20; $t(12)$ = 2.98, P = 0.012).

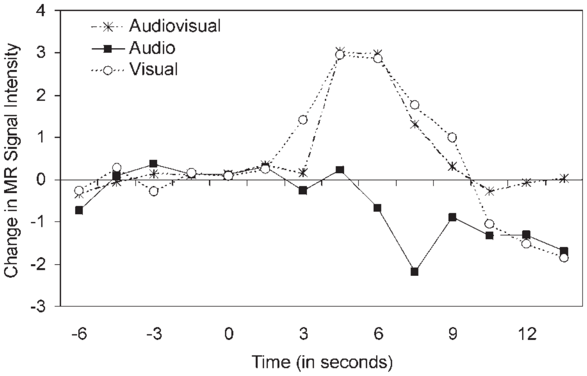

Figure 5. HDRs from area MT/V5. This area responded to Visual and Audiovisual but not to Audio.

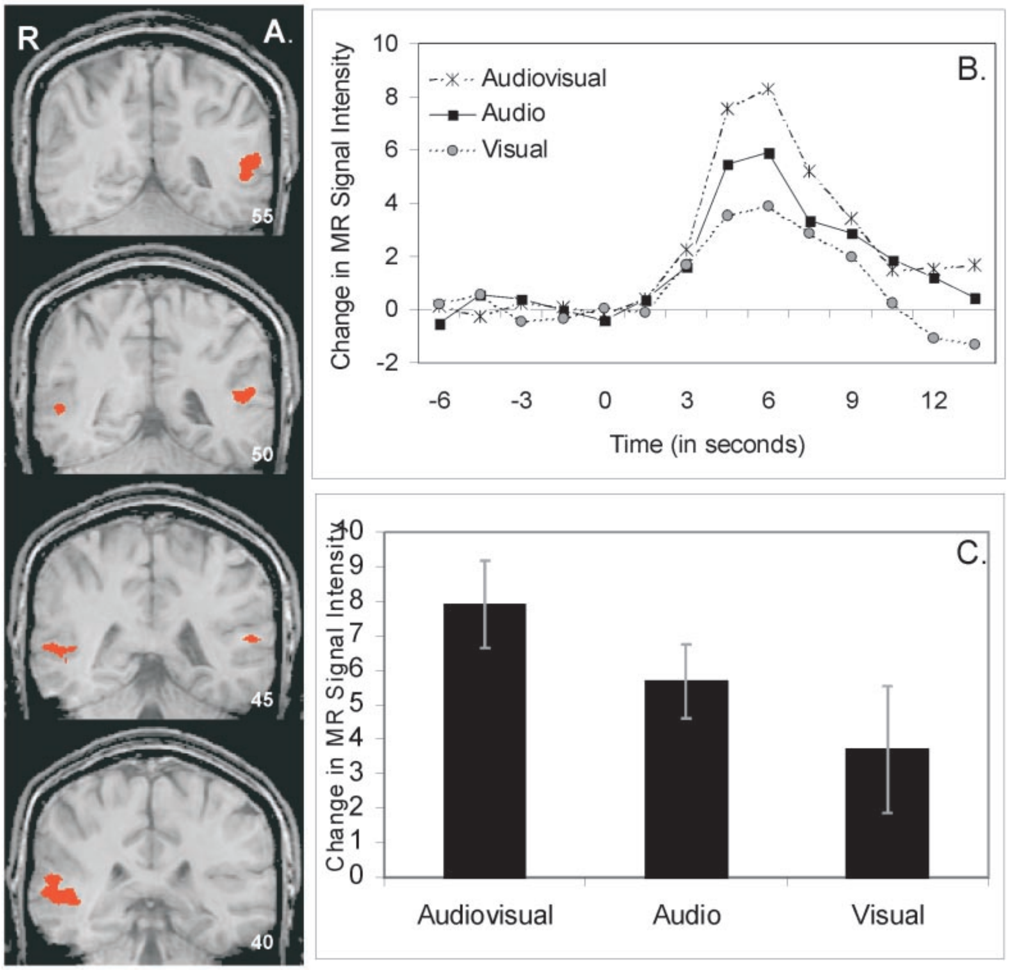

Figure 6. (*A*) Activated voxels comprising the intersection of Audio and Visual. Numbers in the lower right hand corner of each image represent distance (in mm) posterior from the AC. Shown on the template subject's anatomical images are voxels in the STS and STG on slices 40–55 mm posterior from the AC that activate to all three conditions ($t \geq 5.2$). (*B*) HDR waveforms from these voxels. (*C*) Peak amplitude scores of the HDRs, calculated for each subject by averaging the HDR values at 4.5 and 6 s.

Polymodal Interactions

An across-subjects t statistic map comparing Audiovisual to Audio + Visual across the two post-stimulus time points from 4.5 to 6.0 s is presented in Figure 7*A*. This map was generated by calculating (on a voxel-by-voxel basis) the difference between Audiovisual and Audio + Visual averaged across the two time points. Voxels showing suprathreshold Audiovisual > Audio + Visual differences ($t \geq 1.96$, $P \leq 0.03$, one-tailed, uncorrected) and the magnitude of the size effects are indicated by the red to yellow color map for slices 5–35 mm posterior from the AC. These overadditive activations were bilaterally distributed and located primarily in the STG. Voxels showing suprathreshold Audiovisual < Audio + Visual differences ($t \leq -1.96$, $P \leq 0.03$, one-tailed, uncorrected) are indicated by the blue color map for slices 40–50 mm posterior from the AC. These activations were right lateralized, localized primarily to the STS, and were posterior to the overadditive activations.

We interrogated the identified overadditive and underadditive areas using the across-subjects average functional time-course data. The waveforms resulting from this analysis are shown for the overadditive and underadditive regions in Figure 7*B* and *C*, respectively. Examinations of the waveforms from the three stimulus conditions for overadditive and underadditive areas revealed an interesting pattern of responses. As shown in the inset of Figure 7*B*, in those voxels showing overadditive responses, the response to Visual dropped below baseline. In contrast to the pattern observed for overadditive voxels, those voxels demonstrating underadditive responses were equally responsive to the three stimulus conditions (see inset of Fig. 7*C*).

Discussion

The present findings confirm the results of previous studies that reported activation in the STS region to isolated mouth movements [e.g. (Puce *et al*., 1998; Bernstein *et al*., 2002)] and to audiovisual speech [e.g. (Calvert *et al*., 1997, 1999, 2000)]. This study also confirms that the auditory and visual components of speech, when presented in isolation, activate overlapping regions of temporal cortex [e.g. (Calvert *et al*., 1997)]. Further, these results confirm that lateral temporal activity is enhanced by polymodal stimulation because audiovisual speech increased the amplitude of activation in the STS and STG beyond that evoked by auditory and visual speech alone. The STS and STG demonstrated a maximal response to audiovisual speech, confirming the sensitivity of the STS region to the context of a biologically relevant motion; in this case, whether or not visual articulatory information was paired with appropriate auditory information. Interrogation of the waveforms from areas of overadditivity and underadditivity revealed that these

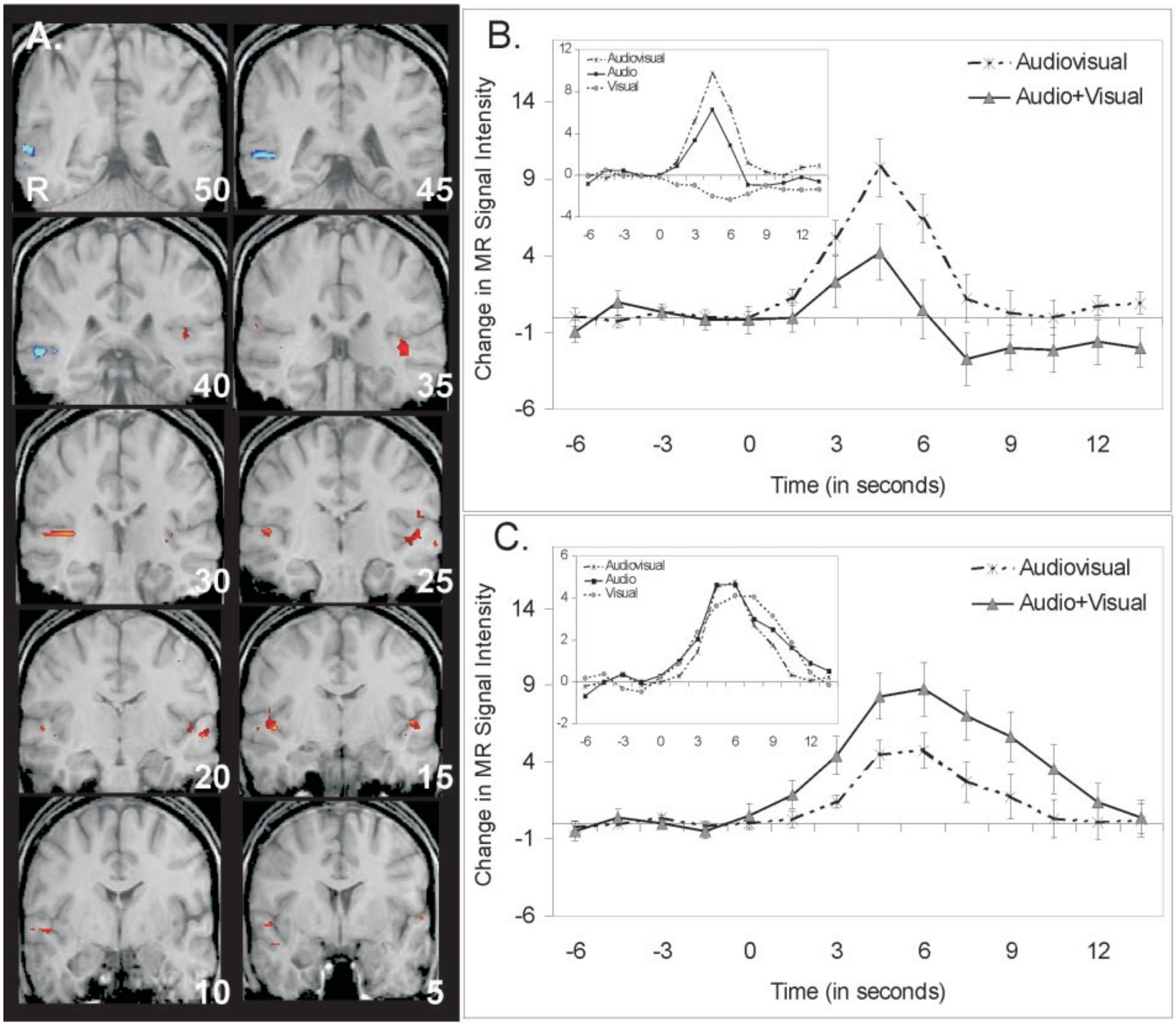

Figure 7. (*A*) Across-subjects *t* statistic map comparing Audiovisual to Audio + Visual, averaging the 4.5 and 6 s time points. Numbers in the lower right corner of each image represent distance (in mm) posterior from the AC. Overadditive voxels are shown in red (Audiovisual > Audio + Visual) and underadditive voxels are shown in blue (Audiovisual < Audio + Visual), ($t \geq 1.96$, $P \leq 0.03$, one-tailed, uncorrected). HDRs to Audiovisual and Audio + Visual from the overadditive (*B*) and underadditive voxels (*C*) are shown. Inset are the HDRs to the three stimulus conditions within the overadditive and underadditive voxels.

interactions could be driven by suppression of the response to Visual in overadditive areas and by maximal responses to every stimulus condition in underadditive areas. Somewhat unexpectedly, the polysensory voxels and voxels showing polymodal interactions identified here were more numerous in the STG than in the STS.

Activation to Visual Speech

This study demonstrated activation to visual speech primarily in the right posterior STS region (40–55 mm posterior from the AC). Puce *et al.* (Puce *et al.*, 1998) reported a focus of nonlinguistic mouth movement-elicited activation in the right posterior STS (Talairach coordinates $x = 50$, $y = -49$, $z = 3$), a pattern of activation to mouth movements consistent with that observed in the present study. Similarly, Bernstein *et al.* (Bernstein *et al.*, 2002) reported a focus of activation in response to lip-reading in the right posterior STS. Like Puce *et al.* (Puce *et al.*, 1998), we also observed an additional focus of mouth movement-elicited activity localized to area MT/V5 (see their Fig. 7*A*).

A recent fMRI study demonstrates that, while visual speech elicits activation in the STS region, this activation is not necessarily within primary auditory cortex (Bernstein *et al.*, 2002). Their observation does not support the proposal of Calvert *et al.* (Calvert *et al.*, 1997) that visual speech is processed by a network involving primary auditory cortex. However, Calvert and colleagues have more recently proposed that visual speech does not directly stimulate auditory cortex. Rather, polymodal enhancements more likely result from feedback following integration of audiovisual speech in polysensory cortex (Calvert *et al.*, 1998, 2000; Calvert, 2001). Calvert *et al.* (Calvert *et al.*, 1997) reported activation to lip-reading in areas including bilateral STG, more anterior to the visual-evoked activation observed in the current study. The STS area that activated to Visual in the current study is a candidate for a visual speech processing area.

The Role of the STS and STG in Social Perception

Social perception refers to the initial stages of evaluating the intentions of others by analysis of biological motions including walking, reaching, eye movements, and mouth movements; hence, social perception is a component of the larger domain of social cognition (Allison *et al.*, 2000). In the present study, activity was greater in response to the pairing of visual articulatory information with appropriate auditory information as compared with visual or auditory information alone. This finding is consistent with previous demonstrations of contextual modulation of responses in the STS region [e.g. (Campbell *et al.*, 2001; Pelphrey *et al.*, 2003)]. The differential response of the STS region to congruent and incongruent speech (Calvert *et al.*, 2000) also supports the sensitivity of the STS region to the social 'appropriateness' of a stimulus (Pelphrey *et al.*, 2003). However, in contrast to the results from Calvert's group, Olson *et al.* (Olson *et al.*, 2002) demonstrated that the STS/STG region did not discriminate between synchronized and unsynchronized audiovisual speech. Instead, they identified the claustrum as a site of possible integration. We did not identify activity in the claustrum for any condition. A weakness of the present study was the failure to include a polymodal incongruent speech condition or a condition incorporating two unimodal cues as a comparison condition for the audiovisual speech condition. Inclusion of this control condition would have allowed us to better dissect the true polysensory effects and greater or less attentional processes when two stimuli are compared with the perception of one stimulus [for a discussion, see Teder-Salejarvi *et al.* (Teder-Salejarvi *et al.*, 2002)].

Polysensory Regions

We observed a population of voxels primarily in the STG 40–55 mm posterior from the AC where Audio intersected Visual activation (Fig. 6*A*). This polysensory area demonstrated supra-threshold positive responses to all three conditions, and the peak HDRs followed the Audiovisual > Audio > Visual pattern. This region may represent the human analogue of the superior temporal polysensory area in macaques. That is, single unit recording studies in monkeys have demonstrated that this region contains neurons that respond to audio, visual, and somatosensory stimuli (Cusick, 1997; Bruce *et al.*, 1981). A similar conclusion regarding the polysensory nature of the STG was drawn by Howard *et al.* (Howard *et al.*, 1996), in discussing their finding that overlapping regions of the STG were activated in response to human walking and speech perception.

Polymodal Interactions

Comparison of this polysensory region in STG to the underadditive (Audiovisual < Audio + Visual) voxels of Figure 7*A* revealed overlap between the two groups on slices 40–50 mm posterior from the AC in the right STG and STS. Thus, polysensory regions can be underadditive in response to *congruent* audiovisual stimuli. However, Calvert *et al.* (Calvert *et al.*, 2000) demonstrated underadditivity in response to *incongruent* speech in the left STS and right STG. In these regions, Calvert *et al.* identified a cluster of voxels 50 mm posterior from the AC that also demonstrated overadditivity to congruent audiovisual stimuli. However, we did not observe overadditivity in the STG polysensory areas observed by Calvert *et al.* (Calvert *et al.*, 2000).

Underadditivity might have resulted from response saturation; that is, the individual responses of these voxels to Audio, Visual and Audiovisual approached maximum response capability. Thus, comparison of Audiovisual to Audio + Visual might not accurately reflect the capability of these polysensory voxels to demonstrate overadditivity. This possibility could be evaluated by incorporating the principle of inverse effectiveness in the experimental design, where polymodal enhancements are maximal when the individual stimuli are minimally effective (Stein and Meredith, 1993; Callan *et al.*, 2002).

Cross-modal Inhibition in Audiovisual Speech Perception

The current study reports areas within the STG and STS that demonstrate overadditivity. The HDRs to Visual within some of these overadditive areas were suppressed (i.e. the waveforms drop below baseline). Hence, we conclude that overadditivity in these areas resulted from summation of a negative response to Visual and a positive response to Audio. This pattern of results suggests cross-modal inhibition for audiovisual stimuli within these regions. A recent MEG study investigating the integration of auditory and visual aspects of letters found underadditive interaction effects (Raij *et al.*, 2001). Similarly, in an fMRI study by Laurienti *et al.* (Laurienti *et al.*, 2002), nonspeech auditory stimuli elicited activation in auditory cortex and deactivation in extrastriate visual cortex. Conversely, a visual stimulus (flickering checkerboard) produced activation in striate and extrastriate visual cortex and deactivation in auditory cortex. Areas of overadditivity were identified, but further investigation revealed that the overadditive response was due to the decrease in the HDR during the nonmatching stimulus condition. In the present study, the overadditive areas that demonstrated suppression of the response to Visual, together with the area MT/V5, where the response to Audio dropped below baseline (Fig. 5), suggest that cross-modal inhibition may be involved as a mechanism in audiovisual speech perception. These findings are similar to recent findings in the domain of visual object perception. Allison *et al.* (Allison *et al.*, 2002) presented evidence for category-selective inhibitory interactions in humans in face and word perception. They identified word- and face-specific sites via subdural recordings made from the fusiform gyri and adjacent cortex. At approximately one-half of word-specific N200 sites, faces evoked a surface-positive potential (P200). Conversely, at about one-half of face-specific N200 sites, words evoked P200 responses. Allison *et al.* (Allison *et al.*, 2002) argued that the P200 represents inhibition of category-specific neurons, and provided a model of synaptic connectivity between neurons selectively activated by faces and letter-strings to account for their results. Their results were within the visual modality, but their model of category-sensitive inhibition might be applicable to cross-modality inhibition as well.

Notes

The authors wish to thank Charles Michelich, Allen Song, Jeremy Goldstein, Sarah Hart, Karen Emberger, Peter B. Mack, Ronald Viola and Lilly Kinross-Wright for their assistance in data acquisition and analysis and manuscript preparation. We thank Dr Gary Glover of Stanford University for providing the spiral pulse sequence used for image acquisition. This research was supported by the Department of Veterans Affairs and NIMH grant MH-05 286. T.M.W. was funded by the Howard Hughes Medical Institute Medical Student Fellows Program. K.A.P. was supported by NICHD 1-T32-HD40127.

Address correspondence to Gregory McCarthy, Duke-UNC Brain Imaging and Analysis Center, Box 3918, Duke University Medical Center, Durham, NC 27710, USA. Email: gregory.mccarthy@duke.edu.

References

Allison T, Puce A, McCarthy G (2000) Social perception from visual cues: role of the STS region. Trend Cogn Sci 4:267–278.

Allison T, Puce A, McCarthy G (2002) Category-sensitive excitatory and inhibitory processes in human extrastriate cortex. J Neurophysiol 88:2864–2868.

Bernstein LE, Auer ET Jr, Moore JK, Ponton CW, Don M, Singh M (2002)

Visual speech perception without primary auditory cortex activation. Neuroreport 13:311–315.

Bruce C, Desimone R, Gross CG (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J Neurophysiol 46:369–384.

Calvert GA (2001) Polymodal processing in the human brain: insights from functional neuroimaging studies. Cereb Cortex 11:1110–1123.

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS (1997) Activation of auditory cortex during silent lipreading. Science 276:593–596.

Calvert GA, Brammer MJ, Iversen SD (1998) Polymodal identification. Trends Cogn Neurosci 2:247–253.

Calvert GA, Brammer MJ, Bullmore ET, Campbell R, Iversen SD, David AS (1999) Response amplification in sensory-specific cortices during polymodal binding. Neuroreport 10:2619–2623.

Calvert GA, Campbell R, Brammer MJ (2000) Evidence from functional magnetic resonance imaging of polymodal binding in the human heteromodal cortex. Curr Biol 10:649–657.

Calvert GA, Hansen PC, Iversen SD, Brammer MJ (2001) Detection of audiovisual integration sites in humans by application of electrophysiological criteria to the BOLD effect. Neuroimage 14:427–438.

Campbell R, MacSweeney M, Surguladze S, Calvert G, McGuire P, Suckling J, Brammer MJ, David AS (2001) 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). Cogn Brain Res 12:233–243.

Cohen J, Cohen P (1983) Applied multiple regression/correlation analysis for the behavioral sciences, 2nd edn. Hillside, NJ: Lawrence Erlbaum.

Cotton JC (1935) Normal ‘visual hearing’. Science 82:592–593.

Cusick CG (1997) The superior temporal polysensory region in monkeys. In: Cerebral cortex: extrastriate cortex in primates (Rockland K *et al.*, eds), vol. 12, pp. 435–468. New York: Plenum.

Duvernoy HM (1999) The human brain: surface, three-dimensional sectional anatomy with MRI, and blood supply. New York: Springer.

Grafton ST, Arbib MA, Fadiga L, Rizzolatti G (1996) Localization of grasp representations in humans by positron emission tomography. Exp Brain Res 112:103–111.

Grézes J, Fonlupt P, Bertenthal B, Delon-Martin C, Segebarth C, Decety J (2001) Does perception of biological motion rely on specific brain regions? Neuroimage 13:775–785.

Grossman E, Donnelly M, Price R, Pickens D, Morgan V, Neighbor G, Blake R (2000) Brain areas involved in perception of biological motion. J Cogn Neurosci 12:711–720.

Howard RJ, Brammer M, Wright I, Woodruff PW, Bullmore ET, Zeki S (1996) A direct demonstration of functional specialization within motion-related visual and auditory cortex of the human brain. Curr Biol 6:1015–1019.

Huettel SA, McCarthy G (2001) The effects of single-trial averaging upon the spatial extent of fMRI activation. Neuroreport 12:2411–2416.

Jha A, McCarthy G (2000) The influence of memory load upon delay-interval activity in a working-memory task: an event-related functional MRI study. J Cogn Neurosci 12:90–105.

Laurienti PJ, Burdette JH, Wallace MT, Yen Y, Field AS, Stein BE (2002) Deactivation of sensory-specific cortex by polymodal stimuli. J Cogn Neurosci 14:420–449.

Lazar NA, Luna B, Sweeney JA, Eddy WE (2002) Combining brains: a survey of methods for statistical pooling of information. Neuroimage 16:538–550.

Mai JK, Assheuer J, Paxinos G (1997) Atlas of the human brain. San Diego, CA: Academic Press.

Olson IR, Gatenby JC, Gore JC (2002) A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. Brain Res Cogn Brain Res 14:129–138.

Pelphrey KA, Singerman JD, Allison T, McCarthy G (2003) Brain activation evoked by the perception of gaze shifts: influence of context. Neuropsychologia 41:156–170.

Perrett DI, Hietanen JK, Oram MW, Benson PJ (1992) Organization and functions of cells responsive to faces in the temporal cortex. Philos Trans R Soc Lond B Biol Sci 335:25–54.

Puce A, Allison T, Bentin S, Gore JC, McCarthy G (1998) Temporal cortex activation in humans viewing eye and mouth movements. J Neurosci 18:2188–2199.

Raij T, Uutela K, Hari R (2001) Audiovisual integration of letters in the human brain. Neuron 28:617–625.

Roberts M, Hanaway J, Morest DK (1987) Atlas of the human brain in section. Philadelphia, PA: Lea & Febiger.

Sankoh A, Huque M, Dubey S (1997) Some comments on frequently used multiple adjustements methods in clinical trials. Stat Med 16:2529–2542.

Schlosser MJ, Aoyagi N, Fulbright RK, Gore JC, McCarthy G (1998) Functional MRI studies of auditory comprehension. Hum Brain Mapp 6:1–13.

Stein BE, Meredith MA (1993) Merging of the senses. Cambridge, MA: MIT Press.

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–225.

Summerfield Q (1987) Speech perception in normal and impaired hearing. Br Med Bull 43:909–925.

Teder-Salejarvi WA, McDonald JJ, Di Russo F, Hillyard SA (2002) An analysis of audio-visual polymodal integration by means of event-related potential (ERP) recordings. Brain Res Cogn Brain Res 14:106–114.

Welch RB, Warren DH (1986). In: Handbook of perception and human performance (Boff KR, Kaufman L, Thomas JP, eds), vol. 1, pp. 1–36. New York: Wiley.

Winston JS, Strange BA, O’Doherty J, Dolan RJ (2002) Automatic and intentional brain responses during evaluation of trustworthiness of faces. Nat Neurosci 5:277–283.

Yamasaki H, LaBar KS, McCarthy G (2002) Dissociable prefrontal brain systems for attention and emotion. Proc Natl Acad Sci 99:11447–11451.