Cross-modal Binding and Activated Attentional Networks during Audio-visual Speech Integration: a Functional MRI Study

We evaluated the neural substrates of cross-modal binding and divided attention during audio-visual speech integration using functional magnetic resonance imaging. The subjects (n = 17) were exposed to phonemically concordant or discordant auditory and visual speech stimuli. Three different matching tasks were performed: auditory-auditory (AA), visual-visual (VV) and auditory-visual (AV). Subjects were asked whether the prompted pair were congruent or not. We defined the neural substrates for the within-modal matching tasks by VV--AA and AA--VV. We defined the cross-modal area as the intersection of the loci defined by AV--AA and AV--VV. The auditory task activated the bilateral anterior superior temporal gyrus and superior temporal sulcus, the left planum temporale and left lingual gyrus. The visual task activated the bilateral middle and inferior frontal gyrus, right occipito-temporal junction, intraparietal sulcus and left cerebellum. The bilateral dorsal premotor cortex, posterior parietal cortex (including the bilateral superior parietal lobule and the left intraparietal sulcus) and right cerebellum showed more prominent activation during AV compared with AA and VV. Within these areas, the posterior parietal cortex showed more activation during concordant than discordant stimuli, and hence was related to cross-modal binding. Our results indicate a close relationship between cross-modal attentional control and cross-modal binding during speech reading.

Keywords: cross-modal matching, human voice, integration, intraparietal sulcus, visual

Introduction

The integration of information conveyed through anatomically distinct sensory pathways is necessary for many human behaviors. In particular, auditory and visual multimodal information processing is important for face-to-face communication. Multisensory convergence contributes to the interpretation of the functional significance of stimuli linked by a common causality (Meredith et al. 1987). Speech reading, or lipreading, for example, enhances the perception of spoken language. Combining audible speech with the corresponding visible articulation movements can improve comprehension to the same degree as altering the acoustic signal-to-noise ratio by 15–20 dB (Sumby and Polack, 1954). Several functional neuroimaging approaches have been used to study cross-modal integration in humans. The first indication of audio-visual integration data emerged when activity in the auditory areas was seen during the viewing of visual stimuli. Calvert et al. (1997) found that silent lipreading activated the superior temporal gyrus. Calvert et al. (2000) also found that the left superior temporal sulcus (STS) exhibited a significant response enhancement to congruent audio-visual inputs; this enhancement was supra-additive to that seen in each isolated modality. Hence, they suggested that STS is one site of audio-visual integration.

However, several studies report conflicting results. Using temporally synchronized and desynchronized voice and lip movement stimuli, Olson et al. (2002) found no activation in the STS; rather, they found activation in the claustrum. Jones and Callan (2003) used functional magnetic resonance imaging (fMRI) to assess the relationship between brain activity and the degree of audio-visual integration of speech information during a phoneme categorization task. They found that visual information had a strong influence on speech perception (the McGurk effect), and this was positively correlated with the activity in the left occipito-temporal junction, an area often associated with processing visual motion. Based on these results, they proposed that auditory information modulates visual processing to affect perception. Bushara et al. (2003) measured transient brain responses to audio-visual binding, seen during a sound-induced change in visual motion perception. They found that cross-modal binding was associated with higher activity in multimodal areas, including the insula/frontal operculum, dorsolateral and medial prefrontal cortex, posterior parietal cortex, posterior thalamus, superior colliculus and posterior cerebellar vermis. Using audio-visual motion discrimination tasks, Lewis et al. (2000) found activation in the intraparietal sulcus (IPS) and fronto-parietal network, suggesting a close relationship between attentional selection and cross-modal integration. They did not find cross-modal activation in the STS. In summary, the proposed neural substrates of audio-visual cross-modal integration, particularly between voice sounds and lip movements, remain controversial.

The comparison between cross-modal and unimodal processing should take into account attentional modulation, which may in part explain the discrepancy across studies (Lewis et al., 2000). Subjects must attend to both modalities during cross-modal conditions, but only to a single modality during unimodal conditions. We hypothesized that an auditory and visual cross-modal speech matching task should involve both divided attention and cross-modal integration processes that are not required for the unimodal matching tasks. Hence, comparisons between the cross-modal and unimodal conditions would highlight the neural substrates underlying divided attention and cross-modal binding. As behavioral results have revealed cross-modal facilitation during lipreading (Sumby and Polack, 1954), the neural substrates of the mechanisms underlying cross-modal binding can be evaluated based on the modulation of responses due to the congruency/incongruency of the stimuli.

In the present study, we explicitly controlled the attentional modulation in the voice and lip movement matching tasks by preparing audio-visual stimuli in which subjects attend only to...
the appropriate modality. Stimuli are identical across the conditions, but the subject is instructed either to direct their attention to one or the other modality or divide their attention between the modalities. A direct comparison between cross-modal and unimodal tasks was planned in order to clarify the net effect of both attentional modulation and cross-modal binding. Within the areas showing the sum of these effects, we compared the congruent and incongruent cross-modal conditions, subtracting out the effects of attention.

Materials and Methods

Subjects

Seventeen healthy volunteers (10 men and 7 women, mean age 27.8 ± 6.7 years) participated in this study. Sixteen subjects were right-handed and one was left-handed according to the Edinburgh handedness inventory (Oldfield, 1971). None of the subjects had a history of neurological or psychiatric illness. The protocol was approved by the ethical committee of the National Institute for Physiological Sciences, and all subjects gave their written informed consent for the study.

MRI

A time-course series of 124 volumes were acquired using T2*-weighted, gradient echo, echo planar imaging (EPI) sequences using a 3.0 T MR imager (Allegra, SIEMENS, Erlangen, Germany). Each volume consisted of 36 slices, each 3.0 mm thick, with a 0.6 mm gap, to cover the entire cerebral and cerebellar cortex. Oblique scanning was used to exclude the eyelids from the images. The time interval between two successsive acquisitions of the same slice was 4000 ms with a flip angle of 85 degrees and 30 ms echo time. The cluster volume acquisition time was 2400 ms, leaving a 1600 ms silent period (Edmister et al., 1999). The field of view (FOV) was 192 mm and the in-plane matrix size was 64 × 64 pixels. For anatomical reference, T1-weighted MPRAGE \( T_1 = 1400 \text{ ms}, T_2^* = 43.8 \text{ ms}, \text{ field angle (FA)} = 8^\circ, \text{ FOV} = 192 \text{ mm}, \text{ matrix size} = 256 \times 256 \text{ mm} \) were collected at the same positions as the echo planar images, and 3-D MPRAGE \( T_1 = 2500 \text{ ms}, T_2 = 4.38 \text{ ms}, \text{ FA} = 8^\circ, \text{ FOV} = 230 \text{ mm}, \text{ matrix size} = 256 \times 256 \text{ mm}, \text{ slice thickness} = 1 \text{ mm}, \) a total of 192 transaxial images) were obtained for each subject.

Face-Voice Matching Tasks

For the face-voice matching tasks, the stimuli were of a human face pronouncing vowels. The auditory and visual stimuli consisted of a digitally recorded female voice and face (16 bit, 11.025 kHz sampling rate, using Adobe Premiere, Adobe, San Jose, CA). The maximum sound pressure (72 dB at the ear), frequency range and duration of each stimulus were adjusted and presented via earphones using Presentation software (Neurobehavorial Systems, Albany, CA) on a microcomputer (Dimension 8200, Dell Computer Co., Round Rock, TX). To minimize the effects of the noise of the MRI scanner, the auditory stimuli were presented during an interval of scanner silence (Seki et al., 2004) and were started 50 ms after the end of image acquisitions (Fig. 1). The visual stimuli were presented at a visual angle of 5.5 × 10.6°. We used an event-related design to minimize habituation and learning effects. The design consisted of four types of event conditions: auditory-auditory (AA), visual-visual (VV) and auditory-visual (AV) matching tasks, and a still condition (STILL). For the AA, AV and VV matching tasks, both auditory and visual stimuli were presented in order to control for the sensory input. Through the session, the subjects were asked to fixate on the appropriate modality. Stimuli are identical across the conditions, but the subject is instructed either to direct their attention to one or the other modality or divide their attention between the modalities. A direct comparison between cross-modal and unimodal tasks was planned in order to clarify the net effect of both attentional modulation and cross-modal binding. Within the areas showing the sum of these effects, we compared the congruent and incongruent cross-modal conditions, subtracting out the effects of attention.

Data Analysis

The first 4 vols of each fMRI session were discarded to allow for stabilization of the magnetization, and the remaining 120 vols per session (a total of 360 vols per subject) were used for analysis. The data were analyzed using statistical parametric mapping (SPM99, Wellcome Department of Cognitive Neurology, London, UK) implemented in Matlab (Mathworks, Sherborn, MA) (Friston et al., 1994, 1995a,b). Head motion was corrected with the realignment program of SPM99 (Friston et al., 1995a). There was no trend of head motion correlated with the task. Following realignment, all images were co-registered to the high-resolution, 3-D, T1-weighted MRI, in reference to the anatomical T1-weighted MRI from locations identical to those in the fMRI images. The parameters for affine and nonlinear transformation into a template of T1-weighted images already fit to standard stereotaxic space (MNI template) (Evans et al., 1993) were estimated with the high-resolution, 3-D, T1-weighted MR images by least square means (Friston et al., 1995a,b). The parameters were applied to the coregistered fMRI data. The anatomically normalized fMRI data were filtered using a Gaussian kernel of 8 mm (full width at half maximum) in the x, y, and z axes.

Statistical Analysis

Statistical analysis in the present study was conducted at two levels. First, individual task-related activation was evaluated. Second, the summary data for each individual were incorporated into the second-level
Individual Analysis

The signal was scaled proportionally by setting the whole-brain mean value to 100 arbitrary units. The signal time course for each subject was modeled using a box-car function convolved with a hemodynamic response function, session effect, and high-pass filtering (112 s). The explanatory variables were centered at 0. To test hypotheses about regionally-specific condition effects, the estimates for each of the model parameters were compared with the linear contrasts. First, we delineated the areas activated during the AA, VV and AV tasks compared with those activated during the STILL periods of the same session. AA-VV and VV-AA comparisons were conducted to depict the neural substrates of within-modal matching. Cross-modal areas were defined as those more prominently activated when there was matched information from different sensory modalities.

Figure 1. Sequence of the face-voice matching tasks and MR acquisition. (a) The time interval between two successive acquisitions of the same slice was 4000 ms. The cluster volume acquisition time was 2400 ms, leaving a 1600 ms silent period (red arrow) during which the face-voice stimuli were presented. (b) A single frame of two faces pronouncing a vowel was presented. At the same time two consecutive voices pronouncing a vowel such as /a/ followed by /e/ was provided. The first voice was synchronized in time with the faces. (c) Prior to the stimulus presentation, the instruction cue was provided for 800 ms. The instruction cue indicates which comparison should be performed. The subjects were requested to make the button press as soon as possible once the instruction marker (red hair cross) changed in shape to horizontal bars or arrows.

Analysis using a random effect model (Friston et al., 1999a) to make inferences at a population level.
two different sensory modalities than when the matching was within either modality. Regions activated by both AV-AA and AV-VV were depicted by the intersection of the areas defined by each contrast. The statistical threshold of each contrast was set at $P < 0.05$, corrected for multiple comparisons at the cluster level for the entire brain. Within these cross-modal areas, we made a comparison between the AV sessions with concordant stimuli and those with discordant stimuli. The statistical threshold was set at $P < 0.05$, corrected for multiple comparisons within the limited search volume.

**Group Analysis with the Random Effect Model**
The weighted sum of the parameter estimates in the individual analysis constituted 'contrast' images, which were used for the group analysis (Friston et al., 1999a). Contrast images obtained via individual analysis represented the normalized task-related increment of the MR signal of each subject. For each contrast, a one sample $t$-test was performed for each and every voxel within the brain to obtain population inferences. The resulting set of voxel values for each contrast constituted a statistical parametric map of the $t$ statistic ($SPM(t)$). The $SPM(t)$ was transformed to normal distribution units ($SPM(Z)$). The threshold for $SPM(Z)$ was set at $Z > 3.09$ and $P < 0.05$, with a correction for multiple comparisons at the cluster level for the entire brain (Friston et al., 1996). Similar to the individual analysis, cross-modal areas were depicted. Within these areas, we compared the AV sessions with concordant stimuli (AV congruent) and those with discordant stimuli (AV incongruent).

**Results**

**Task Performance**
The mean ($\pm$ SD) percentage of correct responses was 84.4 $\pm$ 8.2% for the AV task, 90.0 $\pm$ 6.0% for the AA task, 86.3 $\pm$ 8.5% for the VV task and 98.5 $\pm$ 2.5% for the STILL condition. The performance on the STILL task was significantly better than on the other tasks, and the AA task performance was better than the AV [$P < 0.0001$, one-way analysis of variance (ANOVA) followed by Fisher’s PLSD]. The mean ($\pm$ SD) reaction time was $414 \pm 109$ ms for the AV task, $394 \pm 110$ ms for the AA task, $367 \pm 111$ ms for the VV task and $537 \pm 108$ ms for the STILL condition. The reaction times for the STILL task were significantly slower than those for the other tasks ($P < 0.0001$, one-way ANOVA followed by Fisher’s PLSD).

**Group Analysis with Random Effect Model (Figs 2–5, Tables 1–3)**
As the left hander showed a activation pattern similar to those of the right-handed subjects, the group analysis was conducted with all subjects. Compared with rest, the auditory matching task activated the bilateral transverse (GTT) and superior temporal gyri (G Ts), left inferior frontal gyrus (GFi), dorsal premotor cortex (PMd), inferior (LPi) and superior parietal lobule (LPs), fusiform gyrus (GF), the right cerebellum and lingual gyrus (GL), and the supplementary motor area (SMA). The visual matching task activated the bilateral occipito-temporal junction extending to the GL, GF, G Ts, G Fi, middle frontal gyrus (GFM), L Pi, L Ps, left PMd, right cerebellum and anterior cingulate gyrus (ACG). The audio-visual matching task activated the bilateral G Ts, PMd, L Pi, L Ps, occipito-temporal junction, cerebellum, left GF, G FM, G Tm, intraparietal sulcus (IPS), thalamus, right GL, GF, insula and ACG (Fig. 2).

Compared with the visual matching condition, the auditory matching task induced more prominent activation in the bilateral G Ts and S Ts anterior to the Vpc line [an imaginary vertical line in the mid-sagittal plane passing through the anterior margin of the posterior commissure (Talairach and Tournoux, 1988)], including the part of the frontal operculum and the left G Ts extending to the STS posterior to the Vpc, and the left lingual gyrus (AA–VV). In these areas, the task-related increment during VV was similar to that during AV (Fig. 3). Compared with the auditory matching condition, the visual matching task activated the bilateral G Fi and G Fm, the right occipito-temporal junction extending to the GL and GF, the IPS, left cerebellum and ACG (VV–AA). In these areas, the task-related changes during the VV condition were similar to those during AV (Fig. 3).

More prominent activation during the AV task compared with both the AA and VV tasks was observed in the bilateral PMd and L Ps, left IPS and right cerebellum. Within these areas, concordant AV stimuli activated the bilateral LPs and left IPS more prominently than discordant stimuli ($P < 0.05$, paired $t$-test with random effect model; Fig. 4 and Table 3). This congruency effect was not observed during AA or VV conditions. A typical individual dataset is presented to reveal the more prominent activation during concordant than discordant AV (Fig. 5).

The reverse contrasts (VV–AV masked with VV–STILL, and AA–AV masked with AA–STILL) were used to identify brain areas with significantly lower activity during cross-modal matching than during unimodal matching conditions (Table 4 and Fig. 6). During the AV condition compared with the AA condition, there was a decrease in signal in the bilateral primary and association auditory cortex, and the cerebellum. Compared with the VV condition, the activity in the right MT/V5 area was reduced during the AV condition (Fig. 6).

**Discussion**

**Task Design**
In the present study, we hypothesized that the cross-modal matching of voice and lip movements requires both divided attention and cross-modal binding.
Attention
To isolate the divided attention component, we explicitly targeted attentional modulation by preparing audio-visual stimuli in which subjects attend only to the appropriate modality. Stimuli were identical across the conditions, while the subject was required to shift their attention to each modality or across modalities according to the cued instructions. The motor responses based on the unimodal and cross-modal comparisons were identical.

When recording brain activity during the tasks, different types of signals correspond to the activation of the attentional mechanism ('source' signals) and its interaction with the

**Figure 3.** (a) Neural substrates of within-modal matching by comparing AA-VV. The more pronounced activation seen during AA than VV ($P < 0.05$ corrected) is superimposed on surface-rendered high-resolution MRIs viewed from the right (left of the second row) and left (right of the second row). Top and bottom rows show the task-related activation (% signal change) in each area during each condition. (b) Neural substrates of within-modal matching by VV-AA. The more pronounced activation seen during VV than AA ($P < 0.05$ corrected) is superimposed on surface-rendered high-resolution MRIs viewed from the right (left of the fourth row), left (middle of the fourth row) and top (right of the fourth row). The third and bottom rows show the task-related activation (% signal change) in each area during each condition. *$P < 0.05$, **$P < 0.01$, one-way ANOVA followed by Fisher's LSD.
sensory systems (‘site’ signals) (Corbetta, 1998). In the present study, attention shifting is initiated by the instructions to the subject. A source signal would be associated with a modality-related shift of attention and would be recorded in areas that implement the attentional mechanism and/or in sensory areas responsible for stimulus analysis. During the VV task, for example, a source signal may prime visual processes for a more efficient response. Once a stimulus is presented, stimulus analysis may be enhanced by attention. This would produce modulation of visual processing (‘site’ signal), and this signal would mark the site of the interaction between the source attentional signals and visual processes (Corbetta, 1998). The same situation would occur in auditory processing during the AA task. Whereas source signals provide information regarding the organization of attention systems, site signals provide information on how sensory systems are affected by attention. Hence, VV-AA will show the differences in site signals representing the interaction of source attentional signals with visual processes. Similarly, AA-VV will reveal similar data for auditory processes. Here we assume that the same ‘source’ signals represent the attentional mechanisms in each condition, because they are directed to a single modality. On the other hand, during the AV task, compared with the unimodal tasks, the source signal would be recorded in areas that implement attention mechanisms, given that divided attention is more demanding than unimodal attention, which may reflect the activation illustrated in Figure 4.

**Cross-modal Binding**

Additional neural processing during the AV condition compared with the unimodal VV or AA conditions includes cross-modal binding, in addition to cross-modal divided attention, arousal effects or other task-related effects, such as difficulty. In particular, the AV condition includes the switching of the direction of the spatial attention indicated by the visual cue, whereas AA and VV do not. Hence, the parieto-premotor cortex depicted by the AV-AA and AV-VV comparisons (Fig. 4) partly represents the neural substrates of switching spatial attention. This is concordant with the previous report indicating that switching the direction of the spatial attention is controlled by parieto-premotor cortical networks (Hopfinger et al., 2000).

Due to the methodological limitation, congruency cues in spatial location of the face and corresponding voice were missing. This might have biased one’s perception and cross-modal processing mechanisms. To extract the component of cross-modal binding from these confounds, we utilized the cross-modal response enhancement by comparing a concordant stimulus with a discordant one. Based on an analogy with electrophysiological studies in nonhuman primates and other...
mammals in the superior colliculus or cortex (Wallace et al., 1992, 1996; King and Palmer, 1985), Calvert et al. (2000) postulated that response enhancement and depression are hallmarks of these intersensory interactions in humans.

### Unimodal Matching

The VV-AA comparison revealed activation in the right occipito-temporal junction (hMT, the homologue of the simian MT) and LPi, which are part of the dorsal motion pathway, and bilateral prefrontal areas.

In monkeys, the cortical processing of visual motion is thought to involve anatomically inter-connected visual areas known as the dorsal motion pathway and bilateral prefrontal areas.

In humans, areas V1 and V2 are responsive to visual motion, but more selective responses can be obtained from extrastriate visual areas. For instance, hMT is strongly activated by visual motion stimuli and by visual motion discrimination tasks (Corbetta et al., 1991; Zeki et al., 1991; Dupont et al., 1994; Orban et al., 1995; Tootell et al., 1995a,b; Beauchamp et al., 1997). Additionally, the same stimuli and tasks concurrently activate areas in dorsal occipital and posterior parietal cortex. Bilateral lesions of the lateral occipital cortex (including hMT) can selectively compromise visual motion perception, while leaving auditory and somatosensory motion perception intact (Zihl et al., 1983, 1991; Rizzo et al., 1995). Hence, along the dorsal visual stream, areas up to hMT may be specific to visual motion processing.

The activation in the occipito-temporal junction may include the posterior STS region that is adjacent to MT/V5. The posterior STS is known to be activated during the perception of human body movement (Bonda et al., 1996; Howard et al., 1996; Puce et al., 1998).

A previous functional MRI study showed that mere observation of mouth actions activates bilateral ventral Brodmann's area (BA) 6 and 44, plus the right BA 45 in the inferior frontal gyrus (Buccino et al., 2001). Other studies revealed activity in Broca's area during prepositional speech tasks, including reading words (Price et al., 1994), word generation (McCarthy et al., 1993), decoding syntactically complex sentences (Just et al., 1996) and phoneme discrimination (Zatorre et al., 1992, 1996). Auditory phonemic-discrimination activated secondary auditory cortices bilaterally, as well as Broca's area (Zatorre et al., 1992, 1996). Zatorre et al. (1996) proposed that Broca's area is recruited

---

**Figure 5.** Statistical parametric maps from single subject. The neural substrates of cross-modal matching that were revealed by the intersection of AV-VW and AV-AA (P < 0.05 corrected) are superimposed on the surface-rendered high-resolution MRI of this subject (center). Event-related activation by AV congruent stimuli (red line) and AV incongruent stimuli (blue line) in the bilateral PMd and LPS and left LPi are shown.
Table 1

<table>
<thead>
<tr>
<th>Task</th>
<th>Cluster size</th>
<th>x (mm)</th>
<th>y (mm)</th>
<th>z (mm)</th>
<th>Z-value</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA-STILL</td>
<td>&lt;0.001</td>
<td>3174</td>
<td>-60</td>
<td>-16</td>
<td>6.64</td>
<td>Lt Gm(43)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>2173</td>
<td>51</td>
<td>-10</td>
<td>2.02</td>
<td>Rt Gm(23)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>2206</td>
<td>-36</td>
<td>4</td>
<td>5.43</td>
<td>Lt PmD(6)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>469</td>
<td>21</td>
<td>-69</td>
<td>4.30</td>
<td>Rt cerebellum</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>397</td>
<td>-38</td>
<td>-49</td>
<td>4.85</td>
<td>Lt LPL(40)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.002</td>
<td>333</td>
<td>-12</td>
<td>-62</td>
<td>5.19</td>
<td>Lt LPLs(7)</td>
</tr>
<tr>
<td>VV-STILL</td>
<td>&lt;0.001</td>
<td>3045</td>
<td>51</td>
<td>-65</td>
<td>5.13</td>
<td>Rt Gm(23/37)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.013</td>
<td>51</td>
<td>-10</td>
<td>18</td>
<td>4.10</td>
<td>Lt GFm(37)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>2285</td>
<td>-48</td>
<td>16</td>
<td>5.41</td>
<td>Lt Gmi(44)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>2236</td>
<td>-62</td>
<td>18</td>
<td>5.48</td>
<td>Rt Gm(22)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>1391</td>
<td>51</td>
<td>-10</td>
<td>6.25</td>
<td>Rt Gm(22)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.012</td>
<td>1216</td>
<td>6</td>
<td>24</td>
<td>4.47</td>
<td>Rt GFmi(43)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>700</td>
<td>-38</td>
<td>-49</td>
<td>4.51</td>
<td>Lt LPL(40)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>534</td>
<td>45</td>
<td>-30</td>
<td>4.31</td>
<td>Lt LPLs(7)</td>
</tr>
<tr>
<td></td>
<td>.008</td>
<td>289</td>
<td>-2</td>
<td>23</td>
<td>4.28</td>
<td>Lt ACG(32)</td>
</tr>
<tr>
<td>AV-STILL</td>
<td>&lt;0.001</td>
<td>4749</td>
<td>-60</td>
<td>-16</td>
<td>6.58</td>
<td>Rt Gm(22)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>358</td>
<td>16</td>
<td>25</td>
<td>5.13</td>
<td>Lt GFmi(44)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>2913</td>
<td>-10</td>
<td>24</td>
<td>5.11</td>
<td>Lt LRmi(40)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>2523</td>
<td>-58</td>
<td>-26</td>
<td>5.50</td>
<td>Rt cerebellum</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>1385</td>
<td>-62</td>
<td>50</td>
<td>5.99</td>
<td>Lt LPLs(7)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>835</td>
<td>61</td>
<td>-16</td>
<td>5.41</td>
<td>Lt Gm(22)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>508</td>
<td>-4</td>
<td>17</td>
<td>4.56</td>
<td>Lt ACG(32)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>387</td>
<td>-12</td>
<td>9</td>
<td>3.28</td>
<td>Lt thalamus</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>286</td>
<td>31</td>
<td>-62</td>
<td>5.40</td>
<td>Lt PMd(6)</td>
</tr>
<tr>
<td></td>
<td>0.013</td>
<td>232</td>
<td>-30</td>
<td>48</td>
<td>3.97</td>
<td>Lt Gm(10)</td>
</tr>
<tr>
<td></td>
<td>0.022</td>
<td>206</td>
<td>32</td>
<td>-25</td>
<td>4.45</td>
<td>Rt insula(13)</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>Task</th>
<th>Cluster size</th>
<th>x (mm)</th>
<th>y (mm)</th>
<th>z (mm)</th>
<th>Z-value</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA-VV</td>
<td>&lt;0.001</td>
<td>625</td>
<td>53</td>
<td>-7</td>
<td>10</td>
<td>4.84</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>278</td>
<td>-57</td>
<td>-7</td>
<td>11</td>
<td>4.42</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>211</td>
<td>-57</td>
<td>-30</td>
<td>16</td>
<td>4.06</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>184</td>
<td>-18</td>
<td>-58</td>
<td>-2</td>
<td>3.85</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>111</td>
<td>-44</td>
<td>-25</td>
<td>-4</td>
<td>3.89</td>
</tr>
<tr>
<td>VV-AA</td>
<td>&lt;0.001</td>
<td>874</td>
<td>44</td>
<td>-65</td>
<td>-14</td>
<td>4.72</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>812</td>
<td>57</td>
<td>28</td>
<td>15</td>
<td>5.28</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>449</td>
<td>42</td>
<td>-50</td>
<td>-3</td>
<td>4.50</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>291</td>
<td>-49</td>
<td>48</td>
<td>-2</td>
<td>5.12</td>
</tr>
<tr>
<td></td>
<td>0.011</td>
<td>136</td>
<td>-58</td>
<td>26</td>
<td>16</td>
<td>3.87</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>700</td>
<td>-38</td>
<td>-49</td>
<td>41</td>
<td>4.51</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>534</td>
<td>45</td>
<td>-30</td>
<td>4.31</td>
<td>Lt LPLs(7)</td>
</tr>
<tr>
<td></td>
<td>.008</td>
<td>289</td>
<td>-2</td>
<td>23</td>
<td>4.28</td>
<td>Lt ACG(32)</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>Task</th>
<th>Cluster size</th>
<th>x (mm)</th>
<th>y (mm)</th>
<th>z (mm)</th>
<th>Z-value</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV-AA-VV</td>
<td>&lt;0.001</td>
<td>2523</td>
<td>-58</td>
<td>-26</td>
<td>5.50</td>
<td>Rt cerebellum</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>1385</td>
<td>-62</td>
<td>50</td>
<td>5.99</td>
<td>Lt LPLs(7)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>835</td>
<td>61</td>
<td>-16</td>
<td>5.41</td>
<td>Lt Gm(22)</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001</td>
<td>508</td>
<td>-4</td>
<td>17</td>
<td>4.56</td>
<td>Lt ACG(32)</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>387</td>
<td>-12</td>
<td>9</td>
<td>3.28</td>
<td>Lt thalamus</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>286</td>
<td>31</td>
<td>-62</td>
<td>5.40</td>
<td>Lt PMd(6)</td>
</tr>
<tr>
<td></td>
<td>0.013</td>
<td>232</td>
<td>-30</td>
<td>48</td>
<td>3.97</td>
<td>Lt Gm(10)</td>
</tr>
<tr>
<td></td>
<td>0.022</td>
<td>206</td>
<td>32</td>
<td>-25</td>
<td>4.45</td>
<td>Rt insula(13)</td>
</tr>
</tbody>
</table>

AA-VV

The AA-VV comparison activated the anterior portion of the GTS and STS bilaterally, and the posterior GTS corresponding to planum temporale (PT) and STS on the left side. Compared with visual motion processing, the neural substrates of auditory speech processing are not as well understood. Activations along the STS are often reported in neuroimaging studies of human speech processing (Zatorre et al., 1992; Dehaene et al., 1997); however, their exact role is not clear. Belin et al. (2000) reported voice-selective regions located bilaterally along the upper bank of the superior temporal sulcus (STS) of humans (x = 62, y = -14, z = 0; left: x = -58, y = -18, z = -4, in Talairach's coordinates). They postulated that this area was the human homologue of T9a of the macaque (Seltzer and Pandya, 1978); this area forms part of a hierarchically organized system that is specialized for extracting auditory object features (Rauschecker, 1997; Kaas and Hackett, 1999). Belin et al. (2000) concluded that this area may be involved in the high-level analysis of complex acoustic information, such as the extraction of speaker-related cues, and transmission of this information to other areas for multimodal integration (Mesulam, 1998).

for fine-grained phonetic analysis by means of articulatory decoding. The phonetic analysis of speech depends not only on auditory information but also on access to information about the articulatory gestures associated with a given speech sound (Liberman and Whalen, 2000). This access might be accomplished through a 'mirror' system, including Broca's area, which forms a link between the observer and the speaker (Rizzolatti and Arbib, 1998; Iacoboni et al., 1999).
The areas in and around the posterior IPS (IPP) are known to be polymodal areas. In macaque monkeys, the ventral intraparietal area (VIP), located in the fundus of the IPS, is known to contain cells with distinct polysensory receptive fields. The VIP is known to receive direct projections from motion-related visual areas MT, MST and the surrounding polymodal cortex, as well as from auditory-related cortices (Maunsell and Van Essen, 1983; Lewis and Van Essen, 2000); hence this area is part of the neural substrates for visual motion processing. Individual neurons in the lateral intraparietal area, LIP, respond selectively to the locations of both visual and auditory targets (Striclanne et al., 1996). The neurons in VIP and LIP can represent visuospatial information in a frame of reference that is non-retinotopic (e.g. head- or world-centered) (Duhamel et al., 1998; Snyder et al., 1998). Furthermore, using fMRI, Bremmer et al. (2001) revealed that the depth of the human IPS was equivalent to that of the monkey VIP using polysensory (visual, auditory and somatosensory) stimuli conveying motion information. These results strongly indicate that the cortical areas in and around the IPP are polysensory areas (Lewis et al., 2000).

The present findings also suggest that the neural substrates of cross-modal binding are closely related to those for divided attention. This also has been implicated in tasks involving cross-modal spatial attention (Bushara et al., 1999; Eimer, 1999). Using audio-visual cross-modal motion discrimination, Lewis et al. (2000) found activation in the IPS and dorsal premotor cortex. Lewis et al. (2000) suggested that the control of attention and the motion computations themselves may be intimately intertwined by common mechanisms within the fronto-parietal networks. Also, it has been proposed that the LPs are involved in the orienting of attention (Corbetta et al., 1993).

The parietal–premotor cortical network is related to top-down attentional modulation (Hopfinger et al., 2000). The region of the LPI close to the IPS is important for attention shifting whereas the LPs are active both during shifts of attention in the visual periphery and also when attention is focused on target stimuli and no attentional shifts are required (Hopfinger et al., 2000; Corbetta, 1998). They are part of a network that is important for the control of attention (Driver and Spence, 1998; Mesulam, 1998). Numerous studies have reported that the parietal–premotor network is activated when attention is directed to vision (Posner et al., 1987; Corbetta et al., 1993; LaBar et al., 1999), audition (Pugh et al., 1996; Binder et al., 1997), and cross-modal stimuli (O’Leary et al., 1997; Bushara et al., 1999). The parietal–premotor network may act to direct attention to targets within the same or different sensory modalities (Lewis et al., 2000).

The present study revealed that the neural substrates of divided attention and cross-modal integration overlap in the IPP and LPs. This is consistent with the idea that cross-modal areas allow multidimensional integration through two interactive processes (Mesulam, 1998): the formation of a directory pointing to the distributed sources of the related information; and the establishment, by local neuronal groups, of convergent cross-modal associations related to a target event. Divided attention may enhance the process of the former, while cross-modal response enhancement (congruency effect) is related to the latter. Cross-modal IPP and the adjacent LPs thus enable the binding of modality-specific information into multimodal representations that have both distributed and convergent components.
The IPP was also active during tactile–visual cross-modal shape matching (Saito et al., 2003), audio–visual motion speed discrimination tasks (Lewis et al., 2000) and even olfactory–visual integration (Gottfried and Dolan, 2003), although the locations slightly differed from the present study. Considering this lack of specificity of IPP activation, it is unlikely that the IPP contains the amodal representations of speech itself. Instead, the IPP and adjacent LPs may represent a node through which the senses can access each other directly from their sensory-specific systems.

Finally, during AV condition, activities in the unimodal areas were reduced whereas multimodal areas were enhanced. This is consistent with the previous study (Bushara et al., 2003), suggesting that a reciprocal and ‘competitive’ interaction between multimodal and unimodal areas underlies the integration of simultaneous signals from different sensory modalities.

**Notes**

This study was supported by a Grant-in-Aid for Scientific Research (B14380380 (N.S.) from the Japan Society for the Promotion of Science, and Special Coordination Funds for Promoting Science and Technology from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

Address correspondence to Norihiro Sadato, Section of Cerebral Integration, Department of Cerebral Research, National Institute for Physiological Sciences, Myodaiji, Okazaki, Aichi, 444-8585, Japan. Email: sadato@nips.ac.jp.

**References**


