ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 21 (2004) 725–732

Spatial and temporal factors during processing of audiovisual speech: a PET study

E. Macaluso,[a,*] N. George,[b] R. Dolan,[c] C. Spence,[d] and J. Driver[a,c]

[a] *Institute of Cognitive Neuroscience, University College London, London, UK*
[b] *Laboratoire de Neurosciences Cognitives et Imagerie Cerebrale, LENA-CNRS UPR 640, Paris, France*
[c] *Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK*
[d] *Department of Experimental Psychology, University of Oxford, Oxford, UK*

Received 14 July 2003; revised 16 September 2003; accepted 17 September 2003

Speech perception can use not only auditory signals, but also visual information from seeing the speaker's mouth. The relative timing and relative location of auditory and visual inputs are both known to influence crossmodal integration psychologically, but previous imaging studies of audiovisual speech focused primarily on just temporal aspects. Here we used Positron Emission Tomography (PET) during audiovisual speech processing to study how temporal and spatial factors might jointly affect brain activations. In agreement with previous work, synchronous versus asynchronous audiovisual speech yielded increased activity in multisensory association areas (e.g., superior temporal sulcus [STS]), plus in some unimodal visual areas. Our orthogonal manipulation of relative stimulus position (auditory and visual stimuli presented at same location vs. opposite sides) and stimulus synchrony showed that (i) ventral occipital areas and superior temporal sulcus were unaffected by relative location; (ii) lateral and dorsal occipital areas were selectively activated for synchronous bimodal stimulation at the same external location; (iii) right inferior parietal lobule was activated for synchronous auditory and visual stimuli at different locations, that is, in the condition classically associated with the 'ventriloquism effect' (shift of perceived auditory position toward the visual location). Thus, different brain regions are involved in different aspects of audiovisual integration. While ventral areas appear more affected by audiovisual synchrony (which can influence speech identification), more dorsal areas appear to be associated with spatial multisensory interactions.

Keywords: PET; Spatial; Audiovisual

Introduction

Many events in daily life produce multiple signals that the brain can register via more than one sensory modality. A typical example is listening to someone while seeing the movements of his or her mouth and body. In such cases, the content and spatial source of the spoken message are not only available in audition. Visual cues (lip movements and other visual information from mouth and face) can aid in hearing what has been said and in perceiving where the speech signal came from, especially in noisy environments (e.g., Bertelson and de Gelder, in press; Driver and Spence, 1994; Sumby and Pollack, 1954). Combining related signals from different modalities about a common event is often called multisensory integration (Stein and Meredith, 1993).

Behavioral studies have used various paradigms to examine factors influencing multisensory integration (e.g., see Bertelson, 1998; Driver and Spence, 2000; Spence and Driver, in press; Stein and Meredith, 1993, for reviews). Two extensively researched factors are *temporal* synchrony (or asynchrony), plus common (or different) external *locations*, for signals in different sensory modalities. Many physiological studies of multisensory neurons have indicated that combining inputs from different modalities can produce the greatest increase in firing rates (as compared with unimodal baselines) when the multisensory inputs are approximately synchronous and come from approximately the same external position (e.g., Stein and Meredith, 1993).

While the neural basis of multisensory integration has been extensively studied in animals (e.g., Graziano and Gross, 1995; Stein and Meredith, 1993), there have been fewer studies of its neural basis in humans to date (though see Driver and Spence, 2000; Eimer, in press; King and Calvert, 2001; McDonald and Ward, 2000). Most neuroimaging studies on this topic have focused primarily on either just the role of stimulus location in determining multisensory interactions (e.g., Macaluso et al., 2000; Misaki et al., 2002), or on just the role of synchronously matching inputs to different senses (e.g., Bushara et al., 2001, 2003; Calvert et al., 1999, 2000).

In the audiovisual domain, existing human imaging studies mainly focused on the role of temporal synchrony (and/or semantic congruency) during multisensory stimulation (e.g., Bushara et al., 2001; Calvert et al., 2000; see also Bushara et al., 2003). Calvert et al. (2000) compared brain activity while participants were presented with a congruent audiovisual version of a story (i.e., the face and mouth of the person reading the story was seen while their voice was heard) versus activity during an

* Corresponding author. Institute of Cognitive Neuroscience, University College London, Alexandra House, 17 Queen Square, London WC1N 3AR, UK. Fax: +44-207-813-28-35.
E-mail address: e.macaluso@fil.ion.ucl.ac.uk (E. Macaluso).
Available online on ScienceDirect (www.sciencedirect.com.)

1053-8119/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2003.09.049

incongruent condition, with a different story presented to each modality, thus disrupting audiovisual synchrony (plus phonological and semantic coherence). The congruent condition, which should presumably lead to more successful and useful multisensory integration, produced increased activity in superior temporal sulcus (STS) and inferior parietal lobule. It also led to increased activity in occipital visual areas and in auditory cortex, both of which would typically be considered to respond to stimulation from just one modality but not the other. These results together with other recent imaging evidence now indicate that multisensory integration may affect not only higher-level multisensory areas of association cortex, but may also affect the activity in sensory-specific cortices (see also Calvert et al., 1997; Driver and Spence, 2000; Macaluso and Driver, 2001).

As most audiovisual imaging studies, Calvert et al. (2000) presented visual input on a screen in front of the subjects, while auditory stimuli were delivered elsewhere, over headphones. Such procedures cannot directly address the role of stimulus position in multisensory integration (Spence et al., 2003). As noted already, electrophysiological and behavioral data indicate that the relative location of unimodal inputs during multisensory stimulation (as well as any temporal synchrony) can play a critical role in determining multisensory interactions (e.g., Bertelson, 1998; Bertelson and de Gelder, in press; Driver and Spence, 1998; Spence et al., 2000; Stein and Meredith, 1993). Moreover, in some cases, the observed behavioral phenomena (e.g., 'ventriloquist effect', whereby auditory stimuli can be mislocalized toward visual stimuli) can depend on visual stimuli being spatially discrepant with respect to the sound, yet synchronous with it (Bertelson, 1998; Bertelson and de Gelder, in press; Recanzone, 1998). By contrast, other psychological phenomena (e.g., speech identification and the well-known McGurk effect, whereby perceived speech sounds can be influenced by seen lip movements; McGurk and MacDonald, 1976) are affected primarily just by temporal synchrony (Van Wassenhove et al., 2002), rather than by the relative spatial location of auditory and visual inputs (e.g., see Bertelson et al., 1994; Colin et al., 2001). Separating brain activations related to such different psychological processes may thus require a design in which temporal synchrony and relative spatial location are *both* manipulated, in an orthogonal manner, for auditory and visual inputs.

Accordingly, the present study used Positron Emission Tomography (PET) to investigate the neural consequences of both relative stimulus position and temporal synchrony during the combined presentation of auditory and visual speech inputs. Subjects performed a semantic monitoring task (listening out for animal names in a list of words) while always looking at a video monitor that showed a face mouthing the words that were spoken. In different blocks of trials, these audiovisual signals were either presented synchronously or asynchronously (in the latter case, the auditory stimulus led by 240 ms, which is outside the usual temporal window for the audiovisual integration that produces McGurk-like multisensory phenomena; see Van Wassenhove et al., 2002). Orthogonally to this, the visual and auditory sources were either presented at the same location or in opposite hemifields, using the free-field situation that is permitted by PET scanning (but typically not by fMRI scanning).

Given the results of previous imaging studies on audiovisual integration in speech processing (e.g., Calvert et al., 2000), we expected that synchronous versus asynchronous conditions should activate multisensory association areas (e.g., superior temporal cortex), and possibly unimodal visual and/or auditory regions also (cf. Calvert et al., 2000). Critically, our orthogonal manipulation here of both temporal and spatial audiovisual relations should reveal not only how the relative location of visual and auditory signals affects activity, but also any relation of this to synchrony. Any areas showing multisensory responses of the type identified by Stein et al. (e.g., Stein and Meredith, 1993) might be expected to respond maximally when vision and audition are not only synchronous but also spatially coincident. Areas involved in extracting lipread information from visual input for integration with the speech sounds might be activated more strongly in synchronous conditions versus the asynchronous conditions (Van Wassenhove et al., 2002). Such activation might apply regardless of relative stimulus location, given psychological findings that McGurk-like effects of lip movements upon speech perception are typically uninfluenced by the relative location of the visual and auditory information (see Bertelson et al., 1994; Colin et al., 2001). Finally, any activations that are potentially related to ventriloquist-like phenomena should presumably arise only when visual and auditory stimuli are synchronous, but at different locations, since this is the classical situation for producing ventriloquism (e.g., see Bertelson, 1998).

Methods

Subjects

Eight healthy paid male volunteers (age = 36 ± 3 y) provided written informed consent to participate, which was approved by the National Hospital for Neurology and Neurosurgery Ethics Committee. All were right-handed, had normal or corrected vision, and were native English speakers.

Word stimuli

We used a semantic monitoring task for the spoken word stimuli to ensure that subjects concentrated on these word stimuli, but in a task whose performance should not vary with our audiovisual spatial and temporal manipulations, so that our imaging results would not be confounded with task difficulty. Standard nouns of one to three syllables were chosen from the MRC Psycholinguistic database (http://www.psy.uwa.edu.au/uwa_mrc.htm). We selected only nouns for which concreteness and familiarity scores were above average (mean values—and standard deviation (SD)—of concreteness and familiarity for the selected words according to the MRC database were 438 [120] and 488 [99], respectively). One hundred sixty nontarget (i.e., nonanimal) nouns were selected as well as 40 target (animal) names. Mean number (and SD) of syllables was 1.58 (0.67) for nontargets and 1.55 (0.64) for targets; mean familiarity was, respectively, 553 (39) and 558 (43); mean concreteness was, respectively, 596 (26) and 604 (17). All words were also highly imaginable (Mean Imaginability scores of 589 [32] and 591 [33] for nontargets and targets, respectively). Hence, animal targets should be distinguishable from nonanimal targets only by their semantic category, thus requiring relatively 'deep' processing of the word stimuli. Finally, given the large stimulus set, and the semantic specification of the targets, the task could not be solved by lipread information alone (unlike, say, a task of monitoring for one particular digit in a sequence of single-figure numbers).

Task and procedure

For each of the 12 scanning-blocks, a list of 50 words was created. Each list consisted of 40 nontarget words and 10 target (animal) words, randomly chosen from the pool of 160 nontarget items and 40 target items, respectively. Targets never appeared in immediate succession or as the first word in a block. For each block, a video was presented showing the face and moving lips of a man (author CS) as he read aloud the list of words. He produced words in pairs at a regular pace with a presentation rate of about 3 s per pair of words, thus resulting in a total block duration of 75 s. In the PET scanner room, the video was displayed on a TV monitor placed at a distance of 170 cm from the subject's eyes and at 8.3° to the right of their mid-saggital plane. The exact elevation of the TV screen was fixed individually to ensure comfortable viewing while each participant lay in the scanner and fixated the displayed mouth on the monitor; but the lateral angle was held constant at 8.3°. Speech sounds were delivered through either one of two loudspeakers in the free-field. One loudspeaker was placed just below the TV monitor (same location) while the other was placed symmetrically to the left of the subject (different location, 8.3° to the left of the subject's saggital midline, and thus 16.6° from the visual stimulus). In half of the scanning-blocks, the auditory and visual signals were delivered in synchrony (SYNC conditions), while in the other half, the two modalities were presented asynchronously (ASYNC conditions). In the latter case, the auditory signal preceded the visual signal by 240 ms, a temporal asynchrony between the two sensory streams to which human listeners should be sensitive (Dixon and Spitz, 1980), and which falls outside the usual temporal window for McGurk-like audiovisual integration in speech perception (Van Wassenhove et al., 2002).

Thus, each of the 12 blocks could be presented in one of four experimental conditions: with auditory and visual stimuli presented synchronously at the same location (SySm), synchronously at different locations (SyDf), asynchronously at the same location (AsSm), and asynchronously at different locations (AsDf). The order of presentation of the four experimental conditions was counterbalanced across subjects. White noise was presented from close to the right and left loudspeakers and remained switched on for the entire duration of the experiment to mask any spurious low-level background noise in the scanner.

Subjects were informed that they would see videos of a man reading a list of words, and that this video might sometimes "look like a badly dubbed foreign film" (asynchronous blocks). The subjects' task was to maintain fixation on the mouth seen on the TV screen (as confirmed by the experimenter watching the subject on video during scanning), and to press a button whenever they heard a target (animal) name. The sound level was set such that participants had to concentrate hard on the sound of the voice, while still allowing ceiling performance under all conditions to avoid differences in behavioral performance across conditions. While this precluded the detection of any behavioral effect associated with specific conditions (such as any spatial ventriloquism arising in the SyDf condition), our use of a task that was independent of the spatial and temporal manipulations, and could be performed at ceiling in all conditions, ensured that motor or difficulty-related processes should not confound our imaging results. The subjects' performance was scored manually during the scanning session, and this confirmed that subjects did indeed perform the semantic task correctly for all four types of audiovisual stimulation.

Positron Emission Tomography scan acquisition

Each subject had 12 scans of the distribution of $H_2^{15}O$ acquired with a Siemens/CPS ECAT EXACT HR^+ PET scanner (Siemens/CTI, Knoxville, TN) operated in high-sensitivity three-dimensional mode (1 scan per block of words). Subjects received a total of 350 MBq of $H_2^{15}O$ intravenously over 20 s. A Hanning filter was used to reconstruct the images into 63 planes, resulting in a 6.4-mm transaxial and 5.7-mm axial resolution (full width half maximum).

Statistical analysis

Data were analyzed using statistical parametric mapping (SPM99) software from the Wellcome Department of Imaging Neuroscience (London; http://www.fil.ion.ucl.ac.uk/spm). After realignment, scans were normalized to a standard stereotactic space. T1 structural MRIs from each subject were coregistered into the same space. PET data were also smoothed with a Gaussian kernel of 12-mm full-width half maximum and adjusted to a global mean of 50 ml/dl min^{-1}. A blocked (by subject) analysis of covariance (ANCOVA) model was fitted to the data at each voxel, with condition effects for the four experimental conditions, plus global cerebral blood flow (CBF) as a confounding covariate. Contrasts of condition effects at each voxel were assessed by *t* statistic, transformed to *Z* statistics to provide statistical parametric maps.

We used conjunction analyses (Friston et al., 1999; Price and Friston, 1997) to isolate brain areas showing specific patterns of activation consistent with our three postulated effects (see last paragraph of Introduction) as follows: (1) Brain areas affected by temporal synchrony of the multisensory stimulation *but irrespective of stimulus location*. These were identified using the conjunction of the two simple effects of 'Synchronous minus Asynchronous', under the two levels of stimulus location (i.e., SySm minus AsSm and SyDf minus AsDf). (2) Brain areas responding selectively to synchronized audiovisual speech, but more so specifically when the two sensory streams originated from the *same external location*.

Table 1
Effects of stimulus location during perception of synchronous audiovisual speech

Area	Coordinates	*Z*	*P*
(a) Brain regions activated during synchronous audiovisual speech irrespective of the relative locations of the two sources			
Right fusiform gyrus	28, −58, −18	3.36	0.0004
Right medial lingual gyrus	12, −44, 8	4.84	0.05*
Left fusiform gyrus	−28, −56, −16	3.87	0.0001
Left superior temporal sulcus	−64, −58, 0	3.35	0.0004
(b) Brain regions selectively activated during synchronous audiovisual speech when spatially coincident (vision and audition in the same hemifield)			
Right lateral occipital cortex	34, −68, 20	4.71	0.0001
Right dorsal occipital cortex	12, −68, 44	3.43	0.0003
Left lateral occipital cortex	−36, −86, 0	3.71	0.0002
Left dorsal occipital cortex	−24, −88, 20	4.20	0.0001
(c) Brain regions activated during synchronous audiovisual speech specifically when visual and auditory streams are in opposite hemifields (ventriloquist condition)			
Right inferior parietal cortex	40, −44, 32	3.65	0.0002

* Corrected *P* value.

These areas were highlighted with the conjunction between the main effect of temporal synchrony and the simple main effect of 'same-location' in the context of synchronized stimulation (Sync minus Async and SySm minus SyDf). (3) Finally, we tested for brain areas responding to synchronous stimulation particularly when auditory and visual stimuli came from different locations (i.e., the condition classically associated with ventriloquist-like phenomena). For this, we used the conjunction of the main effect of temporal synchrony with the simple main effect of 'different-location' in the context of synchronized stimulation (Sync minus Async and SyDf minus SySm). Given previous results indicating involvement of particular visual areas and multisensory regions at the temporo-parietal junction in audiovisual integration for speech processing (e.g., see Calvert et al., 2000), we report activation within these regions at a level of P uncorrected < 0.001. Other activations are reported only if they survived correction for multiple comparisons across the whole brain (P corrected < 0.05).

Results

Overall, the synchronous conditions resulted in the activation of a network of brain regions including posterior dorsal and lateral extrastriate regions, plus fusiform gyrus (see Table 1). At lower threshold, activation in left superior temporal sulcus (STS) was also seen. This pattern of activation is consistent with previous studies manipulating temporal aspects (and/or identity-related congruence) of audiovisual bimodal speech stimuli (e.g., Calvert et al., 1999, 2000).

The central aim of the present study was to examine the role of relative stimulus location in crossmodal audiovisual integration, either independently of, or in relation to, temporal synchrony. Using the conjunction analyses specified above (see Methods), we tested for three different patterns of response. First, brain regions showing greater activity to synchronous than asynchronous audio-visual stimuli *independently* of their relative location. This revealed activation of right and left fusiform gyri, plus left STS (Table 1a and Fig. 1). The signal plots in Fig. 1 show that activity in these areas was greater for synchronized stimulation not only when the two stimulus streams were at the same location (bar 1 vs. bar 3 in each histogram), but also when audition and vision were stimulated in opposite hemifields (bar 2 vs. bar 4). Thus, activity here depended on audiovisual synchrony, but not on relative location.

Second, we tested for brain regions whose response depended on the relative spatial *location* of *synchronous* auditory and visual sources. We hypothesized that some regions should be maximally responsive when audiovisual inputs were not only temporally synchronous but also spatially co-localized (cf. Stein and Meredith, 1993; Wilkinson et al., 1996), whereas others might be most active for synchronous but spatially discrepant audiovisual inputs (which can classically produce ventriloquist-like phenomena). Because either type of effect is contingent upon vision and audition being synchronous, we tested for the conjunction of the

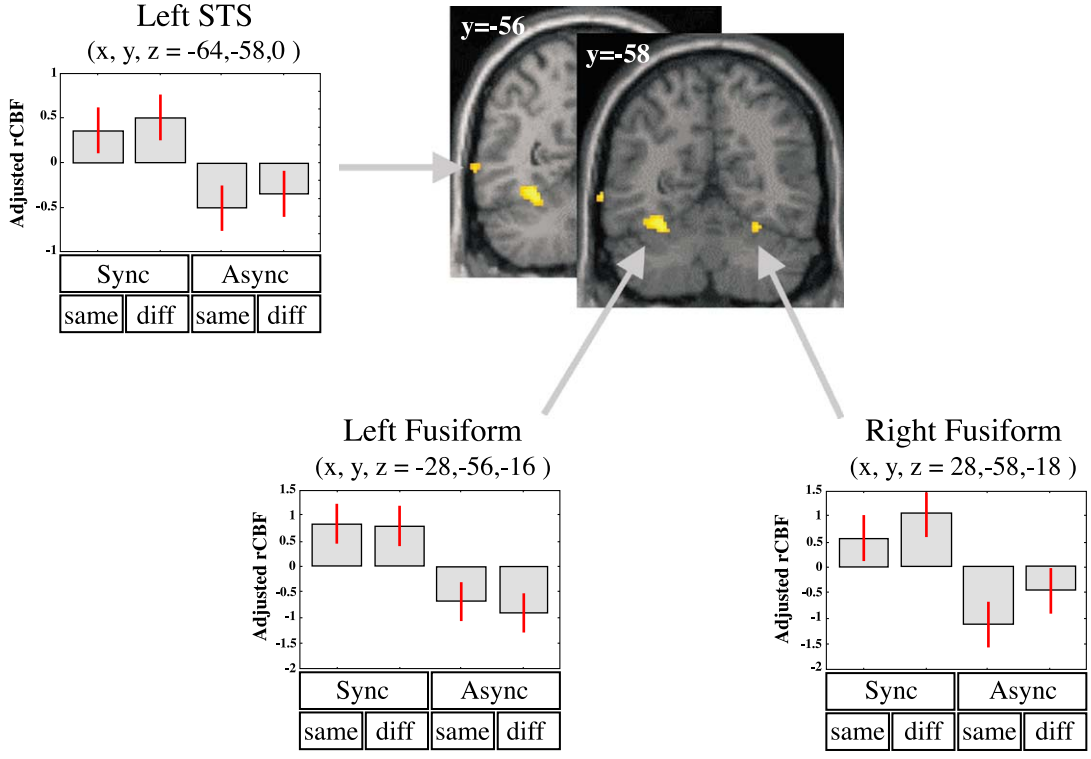

Fig. 1. Location-independent effects of temporal synchrony between auditory and visual speech streams. Anatomical location and signal plots for the three regions showing higher activity for synchronous audiovisual speech (bar 1 and 2) compared to asynchronous speech (bar 3 and 4), irrespective of the relative location of the sources (same or different hemifield). The level of activity in each condition corresponds to the regional blood flow (rCBF) at the maxima, normalized to whole brain global activity of 50 (ml/dl per min), and mean adjusted to zero (±SEM; same notation for all subsequent figures). Sync/Async: synchronous/asynchronous; same/diff. auditory and visual sources in the same or different hemifield.

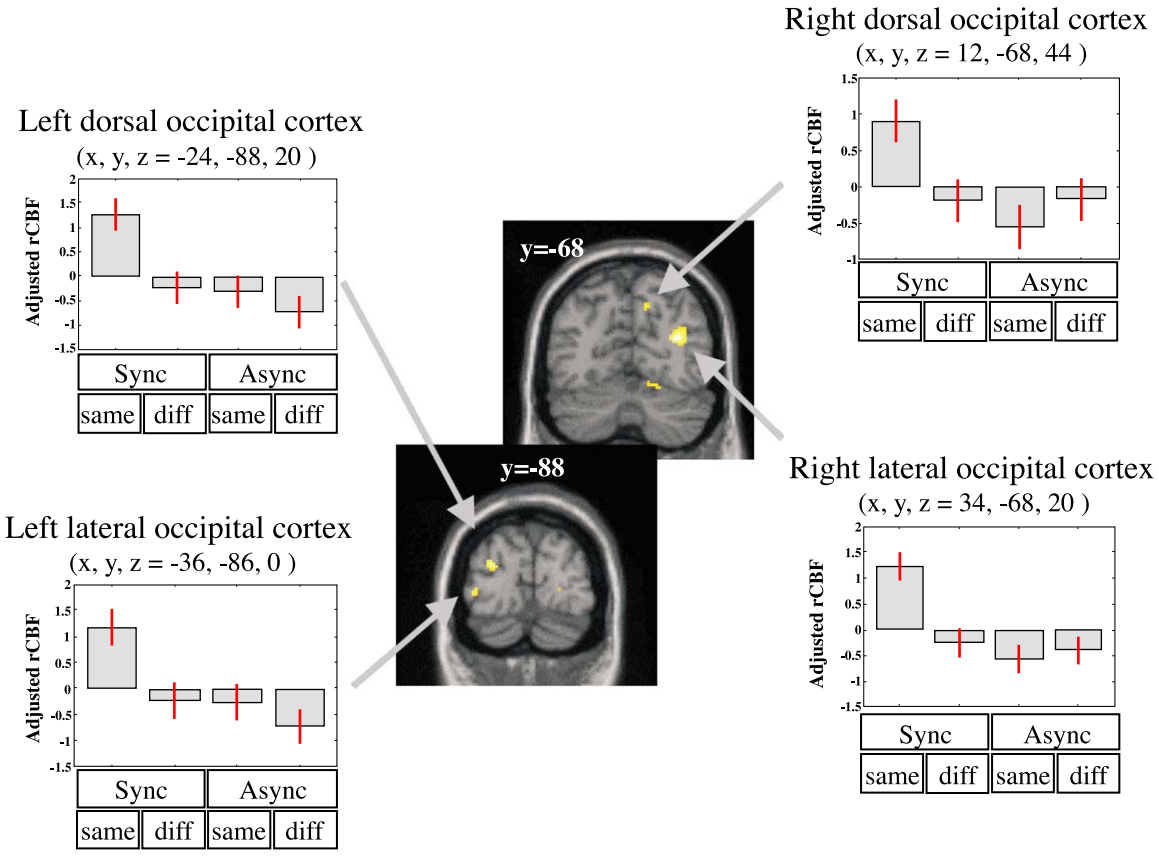

Fig. 2. Areas responding more strongly to synchronized audiovisual stimuli only when presented from the same location. Lateral and dorsal occipital regions showed selective activation when visual and auditory stimuli were not only synchronous but also at the same location (see bar 1, in all signal plots). Peak activations for the left hemisphere were located slightly more posterior than for the right hemisphere, but in both hemispheres two distinct activations could be found in dorsal and lateral occipital cortex.

main effect of synchronous versus asynchronous audiovisual stimulation, together with either simple main effect of spatial location during synchronous stimulation (i.e., concordant minus disparate location, or disparate minus concordant location; see also Methods).

The comparison testing for modulatory effects of temporal synchrony only when the auditory and visual stimuli were in the *same* hemifield yielded activations of lateral and dorsal occipital regions (Table 1b and Fig. 2). In the left hemisphere, the two maxima were found in the posterior part of the superior occipital gyrus and the middle occipital gyrus. In the right hemisphere, the maxima were located more anteriorly, and included the lateral occipital sulcus (extending into the occipito-temporal junction) and the medial part of the superior occipital gyrus (in proximity to the

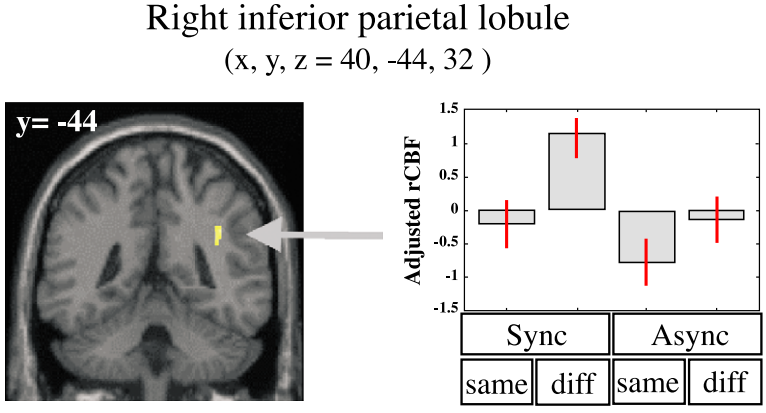

Fig. 3. Selective responses to synchronized auditory and visual stimulation from discordant spatial locations. A region in the right inferior parietal cortex responded selectively to the synchronized condition only when auditory and visual sources were in opposite hemifields (see bar 2 in the plot). Psychologically, this specific condition is classically associated with the ventriloquist effect, a displacement of the perceived source of the auditory stimuli toward the visual location. The present activation in inferior parietal cortex might provide one possible neural substrate for this crossmodal effect.

precuneus). These regions showed increased responses selectively for audiovisual signals that were not only temporally synchronous but also spatially co-localized. The critical effects can be seen in the signal plots of Fig. 2, which shows that in these areas, greater activity was observed only for the 'Synchronous–Same Location' (SySm) condition (bar 1 in each histogram).

By contrast, the comparison of spatially discordant minus co-localized synchronous stimuli (i.e., greatest activity for synchronous vision and audition presented from opposite hemifields) yielded selective activation of the right inferior parietal cortex (see Table 1c and Fig. 3). The signal plot for this area again shows the specificity of the responses, with maximal activity for the 'Synchronous–Different Location' (SyDf) condition (bar 2 in the histogram of Fig. 3). This is classically the condition in which any ventriloquist-like phenomena should arise (see Bertelson, 1998; Bertelson and de Gelder, in press), since ventriloquism depends on spatial discrepancy between temporally synchronous auditory and visual inputs (see also Caclin et al., 2002).

Discussion

While previous imaging studies on audiovisual speech processing primarily investigated audiovisual synchrony or congruency, here we manipulated both temporal *and* spatial aspects of audiovisual speech stimuli orthogonally to examine how these factors may conjointly affect brain responses. We found that temporally synchronous (minus asynchronous) audiovisual speech activated a network of brain areas, including ventral and dorsolateral occipital cortex, plus left superior temporal sulcus (STS). Activity in ventral occipital areas and STS increased during synchronous audiovisual speech, irrespective of the relative location of the auditory and visual input (i.e., same or opposite hemifields, see Fig. 1). By contrast, dorsolateral occipital cortex was most responsive for audiovisual stimuli that were not only synchronous but also spatially concordant (see Fig. 2). Finally, a region in right inferior parietal cortex was selectively activated when auditory and visual streams were synchronous but spatially discordant (see Fig. 3).

Temporal synchrony of audiovisual signals provides a powerful cue for linking multisensory inputs (e.g., Bertelson, 1998; Bushara et al., 2001; Calvert et al., 2000; McGurk and MacDonald, 1976; Meredith et al., 1987; Van Wassenhove et al., 2002). Several previous imaging experiments demonstrated that such synchrony (and/or phonological and semantic congruence between auditory and visual speech signals) can influence brain activations (e.g., Bushara et al., 2001, 2003; Calvert et al., 2000). These studies indicated that audiovisual synchrony can affect activity not only in multisensory areas receiving convergent input from different sensory-specific cortices (e.g., superior temporal sulcus), but may also modulate activity in visual and/or auditory unimodal cortices.

Here, we found that audiovisual synchrony produced increased activity in the superior temporal sulcus (a region of multisensory convergence; Bruce et al., 1981; Jones and Powell, 1970), and also in visual regions of the ventral occipital cortex. These findings agree with previous reports of multisensory influences on these areas (e.g., Calvert et al., 2000), when congruent versus incongruent audiovisual speech combinations were presented, but always from slightly different locations in those prior studies (i.e., auditory stimulation over headphones, but the visual display on a screen in front of the subject, as in Calvert et al., 2000 and related studies). Here we establish for the first time that these particular activations do not depend on relative stimulus location at all, but instead only upon the temporal synchrony between matching audiovisual inputs. Since temporal synchrony can strongly influence whether lipread information contributes to the identification of speech sounds (e.g., see Van Wassenhove et al., 2002), whereas relative auditory and visual location apparently does not (Bertelson et al., 1994; Colin et al., 2001), this may be consistent with the multisensory effects observed in these regions relating to stimulus identification (see also Calvert, 2001). This would accord with behavioral studies documenting that audiovisual interactions related to stimulus identification can occur even with auditory and visual sources placed at different locations (Bertelson et al., 1994, 1995; Colin et al., 2001).

Further imaging studies may be required to assess any different role for the activity in ventral occipital cortex versus superior temporal sulcus, which were commonly activated here for synchronous audiovisual inputs regardless of relative stimulus position. The hierarchical level of these areas within the visual system may provide some clues. The ventral occipital areas may be involved primarily in processing visual attributes from the seen face. By contrast, STS may not only be sensitive to specific biological movements such as lip movements (Bonda et al., 1996), but its higher position within the cortical hierarchy also provides this region with convergent connectivity from both visual *and* auditory areas (Bruce et al., 1981; Jones and Powell, 1970). Convergence of visual and auditory signals to STS might make this area well placed to identify multisensory signals, and any coherence in the identities suggested by each modality. The finding that activity in STS, but not in ventral occipital cortex, was lateralized to the left hemisphere in the present experiment, and in other studies that similarly used linguistic material (see Calvert, 2001), may be consistent with this.

Unlike activity in ventral occipital cortex (and STS), responses in dorsal occipital cortex were affected not only by audiovisual synchrony, but also by the relative location of visual and auditory sources. In both the lateral and superior occipital gyri, activity was selectively boosted for synchronized audiovisual signals when these also originated from the same external location. Multisensory audiovisual interactions in dorsal occipital areas are reported here for the first time. A possible explanation for why previous human imaging studies on audiovisual speech failed to detect multisensory effects upon these particular occipital regions lies in the characteristic responses found here for these regions, which apparently require both audiovisual synchronization and spatial co-localization for strong activation. As mentioned in the Introduction, none of the previous imaging studies on audiovisual speech integration used the same external location for visual and auditory inputs, thus potentially explaining why none previously activated these particular occipital areas.

Such a dependence on synchronized but also spatially coherent multisensory stimulation seem analogous in some respects to the experimental paradigms used in single-unit multisensory research (e.g., Stein and Meredith, 1993). Studies of that type have shown that activity in single neurons during synchronized and spatially congruent multisensory stimulation can exceed the sum of the responses during unimodal stimulation (Stein and Meredith, 1993), while spatially discordant bimodal stimulation can result in suppression instead. Within our design, we were unable to

assess any such nonlinear effects, or to determine any possible contribution of deactivation for spatially discordant stimulation (i.e., vision and audition in opposite hemifields). Assessing this would have required numerous further conditions, including unimodal stimulation of vision and of audition, plus possibly a rest (no stimulation) condition to test for deactivations. Given the limited number of scans allowed with PET, we chose to focus instead on the central issue of any spatial and temporal interactions during multisensory stimulation. For the dorsal occipital regions under discussion, we found that both relative stimulus position and temporal synchrony are important determinants of activity. The dorsal location of these spatially specific effects fits with the general theme that more dorsal areas of visual cortex may be involved in representing spatial aspects of stimuli, while more ventral areas may be mainly concerned with stimulus discrimination and identification (e.g., Haxby et al., 1994). Moreover, the observation of crossmodal but spatially specific interactions in superior and lateral occipital cortex accords with some previous reports of spatially specific crossmodal effects in such regions using combinations of stimuli from other sensory modalities (e.g., vision and touch, see Macaluso et al., 2002; Misaki et al., 2002).

The third main result to emerge here concerned the condition where the two streams were synchronized but spatially discordant. Behaviorally, this type of condition has often been associated with 'ventriloquist' phenomena (i.e., the displacement of the perceived source of the auditory stimuli toward the visual location; see Bertelson, 1998, plus Bertelson and de Gelder, in press). In principle, this type of phenomenon might be associated with at least two different types of neural processes. One possibility would be that auditory location is primarily coded within just auditory areas, and that any ventriloquist-like shift would therefore affect activity only in such regions. A second hypothesis would implicate higher-order areas that are known to be involved in auditory space perception (Griffiths et al., 2000; Pavani et al., 2002), and that have also been associated with spatial representations across sensory modalities (e.g., Andersen et al., 1997; Colby and Goldberg, 1999; Graziano and Gross, 1995; Macaluso and Driver, 2001). The results of the present study may accord with the second type of account, highlighting activation in the right inferior parietal lobule specifically for the condition with synchronous but spatially discordant audiovisual streams. Moreover, activation of these higher-order areas may be consistent with the crossmodal nature of ventriloquist phenomena, which depends on binding information across sensory modalities (see also Bushara et al., 2003). The lateralization of the inferior parietal activation to the right hemisphere might in principle relate to the dominant role of this hemisphere in spatial cognition. However, the asymmetry of our experimental set up (spatially discordant sounds were always presented on the left side) might also play a role.

One limitation of the present study was that although we included the stimulus condition that should classically induce ventriloquism (i.e., synchronous but spatially discordant audiovisual stimulation), we did not directly measure ventriloquism behaviorally during scanning. Moreover, it is possible that the inferior parietal activation might actually reflect subject's awareness that visual and auditory stimuli were at different locations. But if so, this activation should presumably have been as great in the asynchronous spatially discordant condition, where no ventriloquism should arise. Yet the inferior parietal activation was specific to *synchronous* spatially discordant stimulation, which classically causes ventriloquism.

Future studies might attempt to correlate behavioral measurement of perceived auditory shifts during multisensory stimulation that can induce ventriloquist effects, with the level of activity in inferior parietal cortex (see Zatorre et al., 2002), possibly on a trial-by-trial basis using event-related fMRI (Bushara et al., 2003). However, note that this would require the use of sounds in virtual space, given that fMRI scanning does not allow free-field situations unlike the PET scanning method used here. Moreover, motor confounds (such as might arise in a task such as pointing to apparent auditory locations) would have to be avoided.

In conclusion, the present study indicates some dissociation between distinct brain areas responding to different aspects of audiovisual relations during speech processing. Ventral occipital areas and the superior temporal sulcus were activated in all conditions that involved synchronized auditory and visual stimuli, suggesting the involvement of these areas in discrimination and identification of speech signals for multisensory integration. By contrast, dorsolateral occipital regions selectively responded to synchronized bimodal stimulation only when originating from a common external location. We tentatively related, these activations to the spatial aspect of crossmodal integration involved in allocating both visual and auditory signals to a single external event. Finally, we found that the condition traditionally associated with ventriloquism, comprising synchronous but spatially discrepant audiovisual stimulation, led specifically to increased activity in the right inferior parietal lobule, highlighting the possible role of this region in the crossmodal construction of space.

Acknowledgments

This research was supported by a Medical Research Council (UK) Programme grant to J.D., who also holds a Royal Society-Wolfson Research Merit Award. We thank Peter Aston for his technical help.

References

Andersen, R.A., Snyder, L.H., Bradley, D.C., Xing, J., 1997. Multimodal representation of space in the posterior parietal cortex and its use in planning movements. Annu. Rev. Neurosci. 20, 303–330.

Bertelson, P., 1998. Starting from the ventriloquist: the perception of multimodal events. In: Sabourin, M., Craik, F., Roberts, M. (Eds.), Advances in Psychological Science: Biological and Cognitive Aspects, vol. 1. Psychology Press, Hove, UK, pp. 419–439.

Bertelson, P., de Gelder, B., in press. The psychology of multimodal perception. In: Spence, C., Driver, J. (Eds.), Crossmodal space and crossmodal attention. Oxford Univ. Press, Oxford, UK.

Bertelson, P., Vroomen, J., Wiegeraad, G., de Gelder, B., 1994. Exploring the relation between McGurk interference and ventriloquism. International Congress on Spoken Language Processing (Yokohama), pp. 559–562.

Bertelson, P., Vroomen, J., de Gelder, B., 1995. Interaction of auditory and visual data in speech recognition and voice localization: McGurk interference vs. ventriloquism. Paper presented at the meeting of the Experimental Psychology Society, July 11, Birmingham, UK.

Bonda, E., Petrides, M., Ostry, D., Evans, A., 1996. Specific involvement of human parietal systems and the amygdala in the perception of biological motion. J. Neurosci. 16, 3737–3744.

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

Bushara, K.O., Grafman, J., Hallett, M., 2001. Neural correlates of auditory-visual stimulus onset asynchrony detection. J. Neurosci. 21, 300–304.

Bushara, K.O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., Hallett, M., 2003. Neural correlates of cross-modal binding. Nat. Neurosci. 6, 190–195.

Caclin, A., Soto-Faraco, S., Kingstone, A., Spence, C., 2002. Tactile "capture" of audition. Percept. Psychophys. 64, 616–630.

Calvert, G.A., 2001. Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb. Cortex 11, 1110–1123.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., David, A.S., 1997. Activation of auditory cortex during silent lipreading. Science 276, 593–596.

Calvert, G.A., Brammer, M.J., Bullmore, E.T., Campbell, R., Iversen, S.D., David, A.S., 1999. Response amplification in sensory-specific cortices during crossmodal binding. NeuroReport 10, 2619–2623.

Calvert, G.A., Campbell, R., Brammer, M.J., 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr. Biol. 10, 649–657.

Colby, C.L., Goldberg, M.E., 1999. Space and attention in parietal cortex. Annu. Rev. Neurosci. 22, 319–349.

Colin, C., Radeau, M., Deltenre, P., Morais, J., 2001. Rules of intersensory integration in spatial scene analysis and speech reading. Psychol. Belg. 41, 131–144.

Dixon, N.F., Spitz, L., 1980. The detection of auditory visual desynchrony. Perception 9, 719–721.

Driver, J., Spence, C., 1994. Spatial synergies between auditory and visual attention. In: Umilta, C., Moscovitch, M. (Eds.), Attention and Performance XV: Conscious and Nonconscious Information Processing. MIT Press, Cambridge, USA, pp. 311–331.

Driver, J., Spence, C., 1998. Cross-modal links in spatial attention. Philos. Trans. R. Soc. Lond., B. Biol. Sci. 353, 1319–1331.

Driver, J., Spence, C., 2000. Multisensory perception: beyond modularity and convergence. Curr. Biol. 10, 731–735.

Eimer, M., in press. ERP studies of crossmodal spatial attention. In: Spence, C., Driver, J. (Eds.), Crossmodal space and crossmodal attention. Oxford Univ. Press, Oxford, UK.

Friston, K.J., Holmes, A.P., Price, C.J., Buchel, C., Worsley, K.J., 1999. Multisubject fMRI studies and conjunction analyses. NeuroImage 10, 385–396.

Graziano, M.S., Gross, C.G., 1995. The representation of extrapersonal space: a possible role for bimodal, visuo-tactile neurons. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences. MIT Press, Cambridge, USA, pp. 1021–1034.

Griffiths, T.D., Green, G.G., Rees, A., Rees, G., 2000. Human brain areas involved in the analysis of auditory movement. Hum. Brain Mapp. 9, 72–80.

Haxby, J.V., Horwitz, B., Ungerleider, L.G., Maisog, J.M., Pietrini, P., Grady, C.L., 1994. The functional organization of human extrastriate cortex: a PET-rCBF study of selective attention to faces and locations. J. Neurosci. 14, 6336–6353.

Jones, E.G., Powell, T.P., 1970. An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. Brain 93, 793–820.

King, A.J., Calvert, G.A., 2001. Multisensory integration: perceptual grouping by eye and ear. Curr. Biol. 11, 322–325.

Macaluso, E., Driver, J., 2001. Spatial attention and crossmodal interactions between vision and touch. Neuropsychology 39, 1304–1316.

Macaluso, E., Frith, C.D., Driver, J., 2000. Modulation of human visual cortex by crossmodal spatial attention. Science 289, 1206–1208.

Macaluso, E., Frith, C.D., Driver, J., 2002. Directing attention to locations and to sensory modalities: multiple levels of selective processing revealed with PET. Cereb. Cortex 12, 357–368.

McDonald, J.J., Ward, L.M., 2000. Involuntary listening aids seeing: evidence from human electrophysiology. Psychol. Sci. 11, 167–171.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.

Meredith, M.A., Nemitz, J.W., Stein, B.E., 1987. Determinants of multisensory integration in superior colliculus neurons: I. Temporal factors. J. Neurosci. 7, 3215–3229.

Misaki, M., Matsumoto, E., Miyauchi, S., 2002. Dorsal visual cortex activity elicited by posture change in a visuo-tactile matching task. NeuroReport 13, 1797–1800.

Pavani, F., Macaluso, E., Warren, J.D., Driver, J., Griffiths, T.D., 2002. A common cortical substrate activated by horizontal and vertical sound movement in the human brain. Curr. Biol. 12, 1584–1590.

Price, C.J., Friston, K.J., 1997. Cognitive conjunction: a new approach to brain activation experiments. NeuroImage 5, 261–270.

Recanzone, G.H., 1998. Rapidly induced auditory plasticity: the ventriloquism aftereffect. Proc. Natl. Acad. Sci. U. S. A. 95, 869–875.

Spence, C., Driver, J., in press. Crossmodal space and crossmodal attention. Oxford Univ. Press, Oxford, UK.

Spence, C., Pavani, F., Driver, J., 2000. Crossmodal links between vision and touch in covert endogenous spatial attention. J. Exp. Psychol. Hum. Percept. Perform. 26, 1298–1319.

Spence, C., Baddeley, R., Zampini, M., James, R., Shore, D.I., 2003. Crossmodal temporal order judgments: when two locations are better than one. Percept. Psychophys. 65, 318–328.

Stein, B.E., Meredith, M.A., 1993. The Merging of the Senses. MIT Press, Cambridge, USA.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 226, 212–215.

Van Wassenhove, V., Grant, K.W., Poeppel, D., 2002. Temporal integration in the McGurk effect. J. Cogn. Neurosci. 14, 146 (Suppl).

Wilkinson, L.K., Meredith, M.A., Stein, B.E., 1996. The role of anterior ectosylvian cortex in cross-modality orientation and approach behavior. Exp. Brain Res. 112, 1–10.

Zatorre, R.J., Bouffard, M., Ahad, P., Belin, P., 2002. Where is 'where' in the human auditory cortex? Nat. Neurosci. 5, 905–909.