NeuroReport 10, 2619–2623 (1999)

Response amplification in sensory-specific cortices during crossmodal binding

Gemma A. Calvert,[CA] Michael J. Brammer,[1] Edward T. Bullmore,[1] Ruth Campbell,[2] Susan D. Iversen[3] and Anthony S. David[1]

Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU; [1]Institute of Psychiatry, De Crespigny Park, Denmark Hill, London SE5 8AF; [2]Department of Human Communication Science, University College London, Chandler House, 2 Wakefield Street, London, WC1N 1PG; [3]Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK

[CA]Corresponding Author

INTEGRATING information across the senses can enhance our ability to detect and classify stimuli in the environment. For example, auditory speech perception is substantially improved when the speaker's face is visible. In an fMRI study designed to investigate the neural mechanisms underlying these crossmodal behavioural gains, bimodal (audio-visual) speech was contrasted against both unimodal (auditory and visual) components. Significant response enhancements in auditory (BA 41/42) and visual (V5) cortices were detected during bimodal stimulation. This effect was found to be specific to semantically congruent crossmodal inputs. These data suggest that the perceptual improvements effected by synthesizing matched multisensory inputs are realised by reciprocal amplification of the signal intensity in participating unimodal cortices. *NeuroReport* 10:2619–2623 © 1999 Lippincott Williams & Wilkins.

Key words: Auditory speech; Back-projections; Crossmodal integration; Multimodal

Introduction

The ability to detect relationships between different sensory events and to integrate them for perceptual gain is a prominent feature of brain function. For example, in noisy surroundings, sight of the speaker's lip and mouth movements can improve the perception of speech to a degree equivalent to altering the signal to noise ratio of the audible stimulus by 15–20 dB [1]. When confronted by multiple speakers, there is a powerful propensity to match the audible and visible components of speech to aid localization of each sound source so as to determine who is saying what [2].

Electrophysiological studies in lower mammals have shown that similar enhancements can be observed in the responses of single cells in the superior colliculus to spatially congruent multisensory inputs ('where' information) [3]. Although it is tempting to speculate that convergence of congruent unimodal inputs on to such multimodal cells may form a general physiological basis for polysensory integration, direct evidence to support this contention in higher mammals is lacking. Whether such a mechanism of crossmodal synthesis applies in humans, or for the integration of information relating to stimulus identity ('what' information), remains to be established [4].

Here, we report an fMRI study explicitly designed to investigate the neural mechanisms and sites for crossmodal integration during passive audio-visual speech perception and the perceptual enhancement obtained when speech is simultaneously perceived from two modalities rather than one. The strategy we adopted was to carry out two experiments involving both possible bimodal-unimodal contrasts and to identify those areas that were consistently activated irrespective of the nature of the unimodal contrast. In one experiment there was a continuous visual speech stream with periodic auditory input (visual *vs* audio-visual) and in the other, continuous auditory speech was paired with periodic visual speech (auditory *vs* audio-visual). Under these conditions we were able to detect cerebral activations coincident with combined heard and seen speech that were not ob-

served during speech perception from a single modality.

Subjects and Methods

Subjects: Five right-handed volunteers (three males; two females) with an average age of 35 years (range 24–49) participated in the study. All subjects gave informed consent to the protocol that had been approved by the local Research Ethics Committee.

Design: Both experiments employed a box-car design in which two subtasks (A and B), each of 30 s duration, alternated for a total scanning time of 5 min. Stimuli consisted of numbers between 1 and 10, spoken in random order by a female speaker at a rate of 1 every 3 s. These were presented either audio-visually (Task B in both experiments), heard whilst viewing a static image of the speaker's face (Task A, Experiment 1), or seen (i.e. lip-read) in the absence of auditory input (Task A, Experiment 2). Numbers were selected in preference to other lexical items (e.g. simple words or consonant-vowel syllables) as previous data [5] and pilot studies revealed that they were more closely matched for ease of comprehension in both modalities (auditory or visual) than common nouns. The latter proved considerably more difficult to discriminate by lip-reading than hearing. For visual presentations, only the lower half of the speaker's face was visible to minimize the influence of gaze or facial identity processing.

Procedure: All subjects were shown examples of the stimuli prior to scanning to ensure they were both familiar with the task instructions and able to lip-read the stimuli with close to 100% accuracy. For both fMRI experiments, subjects were instructed to rehearse silently each number as it was seen or heard. These instructions were designed to prevent intermittent voluntary silent articulation and to focus the subjects' attention to the stimuli. Visual stimuli were recorded on videotape and projected onto a screen located at the base of the scanner bed via a Proxima 8300 LCD projector. The stimuli were viewed through a mirror angled above the subject's head in the scanner. Auditory stimuli were presented from the audio output of a video recorder via a pneumatic headset designed to minimize interference from scanner noise. The sound level of the speech was ~95 dB with scanner noise attenuated to 80 dB.

Data acquisition: Gradient echo echoplanar MRI data were acquired with a 1.5 T General Electric Signa scanner retrofitted with advanced NMR operating console with a standard quadrature head coil. Head movement was minimized by positioning the subject's head between cushioned supports and by securing a headstrap. 100 T2*-weighted images depicting BOLD contrast [6] were acquired with an in-plane resolution of 3 mm (TR = 3 s, TE = 40 ms) at each of 10 near-axial non-contiguous 5 mm thick slices (with 0.5 mm interslice gap) positioned parallel to the intercommissural line to cover visual, auditory and frontal cortices. For each subject a 43 slice, high resolution inversion recovery gradient echo echoplanar image series of the whole brain was also acquired parallel to the intercommissural plane (TE = 40 ms, TI = 180 ms, TR = 16 s; in-plane resolution 1.5 mm; slice thickness 3 mm) in the same scanning session to aid in normalization of the individual datasets into standard stereotactic space [7] when generic activation maps were computed.

Data analysis: Following correction for movement during image acquisition using standard algorithms [8], analysis of the individual subject data by sinusoidal regression yielded estimates for the amplitudes of the sine and cosine components of the response at the frequency of alternation between the two alternating conditions. These estimates (γ and δ) were used to compute the standardised power (fundamental power quotient or FPQ) and phase of the response at each voxel [9]. γ and δ were then re-estimated 10 times at each voxel following random permutation of the time-series data. This facilitated construction of a distribution of FPQs under the null hypothesis of no experimentally determined response at the experimental design frequency. Tests for activation of any voxel could then be performed by obtaining the appropriate critical value from the distribution of randomized FPQs and accepting as activated any voxel whose FPQ exceeded this threshold (normally set at $p < 0.004$ in the current series of experiments).

The voxel-wise FPQ data obtained in each subject were then transformed into standard stereotactic space by methods described previously [8]. The data from the two experiments were then combined and analysed using a linear model to detect effects that were dependent on and independent of the nature of the unimodal contrast (auditory or visual).

The model can be expressed as $FPQ_{ij} = \alpha_{0i} + \alpha_{1i}G + \varepsilon_{ij}$, where FPQ_{ij} is the FPQ in the jth individual (i.e. across Experiment 1 or 2) at the ith voxel in standard stereotactic space, α_1 and α_0 are the parameters estimated from the model and ε_{ij} is the residual error at each voxel. The effect of group membership (G is the group classification parameter) is parameterized by α_{1i} at the ith voxel and the group-independent (overall mean) effect by α_{0i}.

This model was fitted to the Talairach transformed FPQ data obtained by random permutation of the time series (see above) as well as the FPQ data obtained by analysis of the observed time series. Fitting to the randomised FPQ data permitted construction of distributions of α_{1i} and α_{0i} under the null hypothesis that there was no experimentally determined response to periodic alternation of the bimodal and unimodal conditions in Experiments 1 and 2. The null distributions of α_{1i} and α_{0i} were then used to determine critical values of the two parameters for statistical significance at any required level of probability. As the main goal of the analysis was to identify voxels showing significant responses to bimodal or unimodal stimulation, regardless of the unimodal contrast condition, we were primarily interested in estimating and testing experiment-independent effects (α_{0i}). However, as α_{0i} is independent of α_{1i} a significant value of α_{0i} could arise principally due to a contribution from one of the two experiments. For example, a large response in one experiment and small one in the other may produce a mean value which is significant but does not imply any constancy of responses in the two experiments. The inclusion of the α_{1i} term in the model allows such responses to be identified and removed from activation maps. Following this conservative correction of the data, significant group effects were then rendered onto a morphological template obtained by transforming a high-resolution SPGR MRI image into standard space.

Results

We hypothesized that voxels activated consistently across the bimodal phases of the two experiments should correspond to those generically involved in crossmodal speech processing whereas the individual experiments should reveal elements restricted to a particular unimodal/bimodal contrast. Similarly, voxels activated consistently across the unimodal conditions in each experiment were hypothesized to have a specific role in the processing of speech from a single modality. Combined analysis of the two experiments revealed both experiment-dependent and experiment-independent brain activations (Table 1).

Experiment-dependent activations: In the contrast between audio-visual and auditory speech perception (Experiment 1) activation in phase with the bimodal condition was localised predominantly in bilateral extrastriate cortex in the region of the occipito-temporal junction (BA 19/37) previously shown to correspond to the functional visual motion area V5 [10]. A smaller cluster of activated voxels was also evident in secondary auditory cortex (BA 42) during the audio-visual condition. The right insula cortex was activated in phase with the unimodal (auditory) condition. Activations identified in the contrast between audio-visual and visual (silently mouthed) speech perception (Experiment 2) were situated primarily on the lateral edge of Heschl's gyrus (BA 41) in both hemispheres. These activations were in phase with the bimodal condition. A further cluster of (5) activated voxels was detected in extrastriate cortex (BA 18) in the region of V2 in phase with the unimodal (seen speech) condition.

Experiment-independent activations: Brain areas activated specifically during bimodal speech included a large bilateral section across the occipito-temporal junction corresponding to visual motion cortex V5 (BA 19/37) and primary and secondary auditory cortex (BA41/42; Fig. 1). The power and extent of these cortical activations was substantially greater than those detected during the audio-visual phases in the independent analyses. Activation in primary auditory cortex was more extensive in the left hemisphere, presumably reflecting the use of verbal stimuli. There was no additional contribution from any other brain area (representing a possible and specific sensory integration site) during the bimodal condition that was not also activated by unimodal speech. Activation specific to both unimodal conditions was localised in the right insula-claustrum (Fig. 1; Table 1).

Table 1. Experiment-independent effects identified by linear modelling of combined experimental results

Talairach coordinates			Cluster size	FPQ	Side	Anatomical definition	BA	Task
x	y	z						
−43	−69	3	17	1.7	L	Middle occipital/middle temporal gyral border	37/19	Bimodal
−46	−69	8	15	1.9	L	Middle occipital/middle temporal gyral border	37	Bimodal
49	−63	3	14	2.4	R	Middle temporal gyrus	37	Bimodal
49	−56	8	9	1.7	R	Middle temporal gyrus	37/21	Bimodal
57	−22	13	8	2.2	L	Heschl's gyrus	41	Bimodal
−46	−25	13	6	1.6	R	Superior temporal gyrus	42/22	Bimodal
29	−6	3	4	1.7	R	Claustrum	–	Unimodal

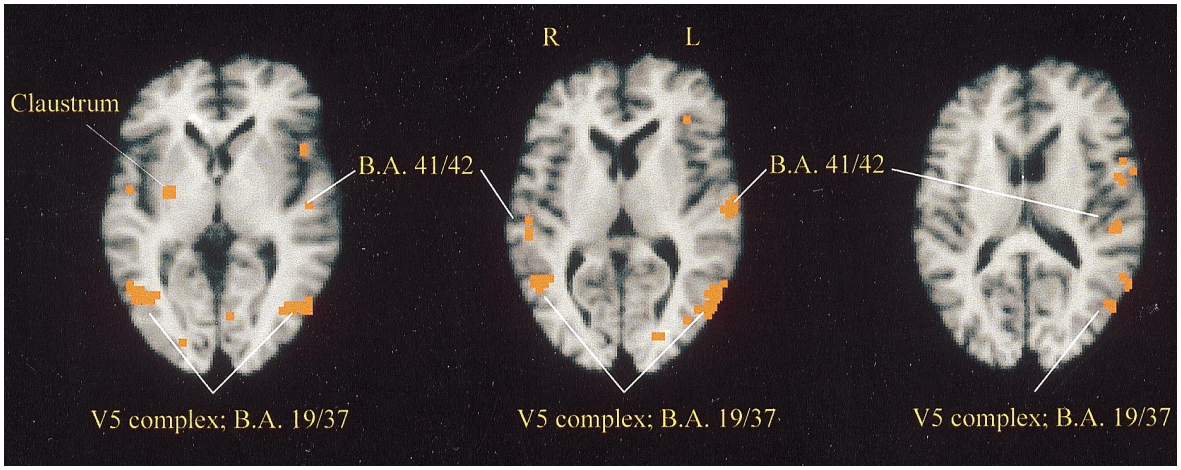

FIG. 1. Experiment-independent effects: brain areas showing significant group activations across both bimodal-unimodal speech contrasts. Three axial slices at Talairach z coordinates +3, +9 and +14 containing >95% of the voxels activated with a probability of type I error <0.001 are shown. Images are displayed in radiographic convention.

Discussion

The aim of these experiments was to study the neural mechanisms engaged during the integration of auditory and visual speech signals. The major finding was enhanced activation of auditory (BA41/42) and visual sensory (V5 complex) cortices during bimodal speech perception. This suggests that the perceptual gains experienced during multimodal signal are realised by amplification of the neuronal response in participating sensory cortices (in the present study, auditory and visual). Such a formulation is consistent with the subjective experience of an improvement in hearing when the speaker can be both seen and heard and enhanced visual attraction towards the sound source (the lip and mouth movements whose dynamic frequency/amplitude patterns correlate with the stream of audible speech).

To establish that the enhancements observed in this study do not generalise to all experimental conditions where there is coincident audio-visual input, we have also analysed the results of an experiment in which unrelated auditory (talking book) and visual (flashing checkerboard) stimuli were presented at different block lengths (auditory 30 s ON/OFF; visual 39 s ON/OFF) permitting the independent characterization of activation due to each input modality [11]. One consequence of this design is that the two modalities overlap at various times during the experiment yet coincident presentation of the checkerboard and the talking book did not produce levels of activation in the primary auditory or visual cortex over and above those observed with the individual stimuli. We therefore conclude that the enhancements seen in the present study are specific to multisensory inputs that are semantically congruous, concordant with behavioural findings [12–14].

In view of the anatomical separation and lack of direct connections between the auditory and visual cortices and, in the present study, absence of any additional contribution from a possible intersensory region during the bimodal condition, the question arises as to how the auditory and visual signals are combined and response enhancements mediated. Because lesions of putative polymodal cortex (e.g. superior temporal and intra-parietal sulci, the lateral prefrontal cortex and the amygdala) in primates have not produced consistent impairments of crossmodal performance [15] and recent human PET data have failed to implicate these areas in tasks of crossmodal matching [16], alternative models of intersensory synthesis have been proposed. In an exhaustive review of the relevant primate literature, Ettlinger and Wilson [17] suggested a system whereby the senses could access each other directly via an interconnecting structure such as the claustrum, which both receives and gives rise to projections from the different sensory cortices. Concordantly, Hadjikhani and Roland [16] reported claustral activation during the crossmodal transfer of visuo-tactile information. In the current study, however, claustral activation was detected specifically during the unimodal phases of the bimodal–unimodal contrasts. These findings suggest that whilst the claustrum may be involved in crossmodal processes, its role may be complex and requires further characterisation. A different model which eliminates the requirement for a dedicated intersensory convergent site, proposes instead a more parallel and distributed system whereby crossmodal binding is achieved through synchronization of the activity of neuronal assemblies in modality-

specific cortices [18]. Although not inconsistent with our data, communication between these unimodal areas must be necessary to facilitate synchronization.

One hypothesis is that the auditory and visual speech signals are combined in polysensory regions of the superior temporal sulcus (STS) and the subsequent response amplification in A1 and V5 effected by means of back-projections. Several sources of evidence support this formulation. First, primate studies have shown that the STS receives convergent inputs from the auditory and visual cortices [19] and contains cells responsive to auditory and visual stimulation [20]. Although homology between primate and human STS cannot automatically be assumed, functional neuroimaging studies in humans have also implicated this region in auditory and visual tasks of a communicative nature including the perception of eye gaze, mouth movement [21] and phonetic perception [22]. The latter finding is particularly compelling in view of psychophysical evidence indicating that the audible and visible components of speech are combined at the phonetic level [23]. Furthermore, in a previous fMRI study comparing brain areas activated during heard speech and by silent lipreading (each contrasted against rest), coincident activations were detected in primary and association auditory cortex, including in the superior temporal sulcus [24]. Finally, in a study of audio-visual speech using MEG, Sams and colleagues [25] demonstrated that the characteristic response (M100 wave) of the auditory cortex to sound could be modified by simultaneous visual speech, with the appearance of a second wave 220 ms after the M100. Such a delay is compatible with the indirect (back-projected) entry of information from the visual stream into primary auditory cortex. In the light of these data, it may seem surprising that no differential activation of this area was detected in the current study in the bimodal condition. However, the absence of a response enhancement in so-called heteromodal areas does not necessarily imply lack of involvement but may simply reflect the small proportion and widely dispersed nature of multimodal cells in these areas [20]. Increased activity amongst such neurons may not therefore be accompanied by a substantial enhancement of the BOLD effect. One way to probe the role of the STS in audio-visual integration might be to use paradigms in which the response to matched and mismatched stimuli are contrasted.

It is also perhaps worthwhile to note that although several researchers [15,16] have drawn the conclusion that heteromodal cortex plays little role in crossmodal matching or transfer (where information perceived from different modalities and relating to two distinct objects is matched along some shared dimension, e.g. size, shape, intensity), it is by no means clear that similar neural mechanisms are operative for tasks of crossmodal integration (like that reported in the current study) which involve determining whether two or more distinct sensory inputs arise from the same object. Further experimentation into the neural basis of crossmodal matching, transfer and binding is required to clarify these contradictory findings.

Conclusion

Processing of bimodal audio-visual speech leads to enhanced activation in auditory (BA 41/42) and visual (V5 complex) cortices. The auditory and visual activations may provide a basis for the perceptual gains elicited when speech information is perceived simultaneously from both sensory modalities. These unimodal sites are, however, unlikely to be the loci for signal combination, the elucidation of which will require further experimentation but may instead be mediated by feedback connections from higher-level heteromodal areas.

References

1. Sumby WH and Pollack I. *J Acoustic Soc Am* **26**, 212–215 (1954).
2. Driver J. *Nature* **381**, 66–68 (1996).
3. Stein BE and Meredith MA. *The Merging of the Senses*. Cambridge, MA: MIT Press, 1993.
4. Calvert GA, Brammer M and Iversen SD. *Trends in Cog Sci* **2**, 247–253 (1998)
5. Campbell R and Dodd B. *Q J Exp Psychol* **32**, 85–99 (1980).
6. Kwong KK, Belliveau JW, Chesler DA *et al. Proc Natl Acad Sci USA* **89**, 5675–5679 (1992).
7. Talairach J and Tournoux P. *Co-planar Stereotactic Atlas of the Human Brain*. Stuttgart: Thieme, 1988.
8. Brammer MJ, Bullmore ET, Simmons A *et al. Magn Reson Imaging* **15**, 763–770 (1997).
9. Bullmore E, Brammer M, Williams SC *et al. Magn Reson Med* **35**, 261–277 (1996).
10. Watson JD, Myers R, Frackowiak RS *et al. Cerebr Cortex* **3**, 79–94 (1993).
11. David AS, Woodruff PW, Howard R *et al. Neuroreport* **7**, 932–936 (1996).
12. Breeuwer M and Plomp R. *J Acoust Soc Am* **79**, 481–499 (1986).
13. Easton RD and Basala M. *Percept Psychophys* **32**, 562–570 (1982).
14. Stein BE, Meredith MA, Huneycutt WS *et al. J Cogn Neurosci* **1**, 12–24 (1989).
15. Ettlinger G and Wilson WA. *Behav Brain Res* **40**, 169–192 (1990).
16. Hadjikhani N and Roland PE. *J Neurosci* **18**, 1072–1084 (1998).
17. Sherk H. The claustrum and cerebral cortex. In: Jones EG and Peters A, eds. *Cerebral Cortex* Oxford: Oxford University Press, 1986: 467–499.
18. von Stein A, Rappelsberger P, Sarnthein J and Petsche H. *Cerebr Cortex* **9**, 137–150 (1999).
19. Jones EG and Powell TP. *Brain* **93**, 793–820 (1970).
20. Baylis GC, Rolls ET and Leonard CM. *J Neurosci* **7**, 330–342 (1987).
21. Puce A, Allison T, Bentin S *et al. J Neurosci* **18**, 2188–2199 (1998).
22. Binder JR, Frost JA, Hammeke TA *et al. Brain* **119**, 1239–1247 (1996).
23. Green KP and Kuhl PK. *J Exp Psychol Hum Percept Perform* **17**, 278–288 (1991).
24. Calvert GA, Bullmore ET, Brammer MJ *et al. Science* **276**, 593–596 (1997).
25. Sams M, Aulanko R, Hamalainen M *et al. Neurosci Lett* **127**, 141–145 (1991).

ACKNOWLEDGEMENTS: This work was supported by a Pump Priming grant from the University of Oxford and the Bethlem and Maudsley Research Fund. G.A.C. is supported by the Medical Research Council of Great Britain and E.T.B. by the Wellcome Trust.

Received 9 June 1999;
accepted 22 June 1999