Multisensory perception: **Beyond modularity and convergence**

Jon Driver* and Charles Spence†

Recent research on multisensory perception suggests a number of general principles for crossmodal integration and that the standard model in the field – feedforward convergence of information – must be modified to include a role for feedback projections from multimodal to unimodal brain areas.

Addresses: *Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1E 6BT, UK. †Department of Experimental Psychology, Oxford University, University of Oxford, South Parks Road, Oxford OX1 3UD, UK.
E-mail: j.driver@ucl.ac.uk; charles.spence@psy.ox.ac.uk

Current Biology 2000, 10:R731–R735

0960-9822/00/$ – see front matter

Most textbooks on perception consider each sense — vision, hearing, touch, olfaction and so on — in isolation, as if each sensory modality was an entirely separate module. In many situations, however, our different senses receive correlated information about the same external objects or events, and this information is combined in our brains to yield multimodally determined percepts. Although there is a venerable literature on crossmodal integration [1,2], recent years have seen a renewed vigour in this field, for several reasons.

While research from the 1960s through to the 1980s was largely concerned with identifying separate modules in the mind/brain, there is an increasing realization [3–6] that understanding the interplay between components in an extended network is as important as fractionating that network into its component parts. Crossmodal integration is a paradigm case of the need to move beyond modularity in this way. There is also a growing awareness [4–7] that principles of crossmodal integration uncovered within one domain, such as speech perception may extend to many other domains, such as stimulus localization, and therefore reflect general architectural constraints. Several recent studies [8–10] now suggest that models of crossmodal integration must move beyond the notion of purely feedforward convergence between separate information sources, which has long been the dominant assumption in the field. Finally, the new methods of cognitive neuroscience, such as functional imaging, seem ideally suited for studying crossmodal integration [8,10–12] and have shed new light on fundamental issues.

In this dispatch, we shall provide examples of these recent developments, first considering cases where two or more modalities can provide information about the same property of the external world, which we term 'convergent' crossmodal integration.

Convergent information from multiple modalities

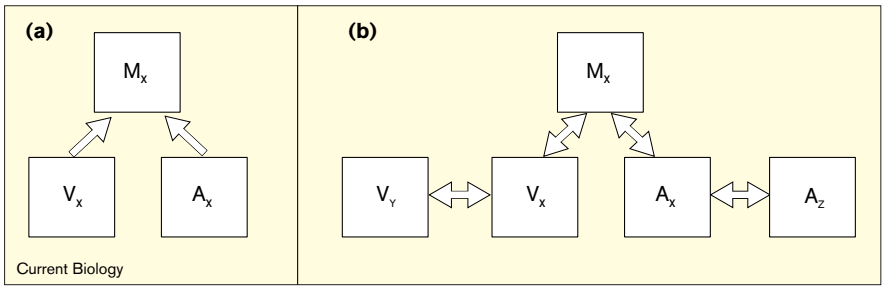
Particular events in the outside world often stimulate several of our senses simultaneously, as when we both hear and see someone speaking. Some of the most famous cases of crossmodal integration concern this particular example. For example, in the McGurk effect [13], seen lip-movements can alter which phoneme is heard for a particular sound [6]; while in the ventriloquism effect, they can alter the apparent location of speech sounds [4].

Such crossmodal effects on stimulus localisation or identification are not restricted to speech. Ventriloquist-like effects on auditory localization can arise whenever a transient visual event, such as a meaningless flash, is synchronized with a hard-to-localize sound, such as a meaningless tone [14,15]. Moreover, ventriloquism may also arise between tactile and visual stimuli [16]. Equally, McGurk-like effects can arise for non-speech stimuli, as when both hearing and seeing musical instruments [17], and for other pairs of modalities also. For instance, perturbing the sounds made as hands are rubbed together can affect the perception of skin texture [18], while changing the color of drinks or food can alter perception of their flavour [19]. The variety of such examples suggests that crossmodal integration may be the rule rather than the exception in real-world perception.

In the above 'convergent' cases, two or more modalities provide concurrent information about the same external property; for example, a spoken phoneme or stimulus location, even skin dryness or food quality. It seems entirely adaptive that the multiple sources of information — that is, different modalities — should be combined to yield the best estimate of the external property in such cases. From this perspective, it may become less surprising that vision can strongly influence auditory localization, as in the ventriloquist effect, given that vision is spatially more acute and so will usually provide helpful information regarding sound-source location.

Massaro [6] has for many years advocated a 'fuzzy logic' model of perception, encompassing many different cases of crossmodal integration within the same general framework of optimally efficient information combination. One modality can be weighted more heavily than another in this combination, to the extent that it produces a greater reduction in uncertainty (as when visual localisation is less variable than auditory localisation). Vision does not always

Figure 1



(a) Schematic illustration of the conventional notion of feedforward convergence of information from different sensory modalities; here, auditory (A_x) and visual (V_x) information about the same external property (x) is combined to generate the multimodal representation M_x. **(b)** Schematic illustration of the idea that multimodal levels of representation may feedback to influence levels traditionally considered as 'unimodal'. Furthermore, by means of this 'vertical' feedback, combined with 'horizontal' connections between the coding of different properties within one modality, crossmodal interactions concerning one stimulus property (x) could affect unimodal coding of a different property (y or z).

dominate; for instance, audition can dominate vision for temporal properties, because of its higher temporal acuity, as when auditory flutter drives the perception of visual flicker [20,21]. Fuzzy logic provides one influential version of a more general idea: that crossmodal integration must be produced by convergence of separate information sources, with suitable weighting of each as they are combined (Figure 1a).

Simple heuristics for integration

In the above examples, different modalities can provide convergent information about the same external event or property. Clearly the nervous system has to discriminate such cases from those where stimulation in the different modalities is entirely unrelated. Several simple heuristics have been suggested for this, to date all depending on spatio-temporal correlations. Thus, stimuli in different modalities occurring at the same (or similar) time and/or place will tend to be treated as referring to the same external event. Temporal correlations dominate some crossmodal effects, spatial correlations dominate others [21,22].

In a series of pioneering studies, Stein and colleagues [3] uncovered spatio-temporal heuristics for integration, at both the behavioural and neural level. Recording from single-cells, they and others found multimodal neurons in the superior colliculus that show interactions between stimulation in two or more modalities. The receptive fields of such neurons typically fall in approximate spatial register across the different modalities that drive them. Moreover, responses to multimodal stimulation from the same location, at roughly the same time, can be facilitated in an over-additive (or 'multiplicative') fashion, as compared with responses to weak stimulation in either modality alone.

Concurrent stimulation at different locations in multiple modalities can lead to suppressive interactions (again of a non-linear kind) as compared with the unimodal baselines. Although some of these non-linearities could reflect floor or ceiling effects in neural responding [23], they might also serve as a cellular instantiation of the multiplicative crossmodal combinations envisaged by Massaro [6]. Similar principles have now been observed in cortical multimodal neurons [24]. Stein and colleagues [3] report that an animal's overt orienting behaviour can follow the same heuristics, while we [5] have reported related findings for covert spatial attention in humans.

In a recent *Current Biology* article, Calvert *et al.* [7] sought to extend such principles of integration in two ways; first, by arguing that similarly non-additive crossmodal interactions can arise in the human cortex; and second, by claiming that this arises for crossmodal influences on stimulus identification, just as for stimulus localization. In their functional magnetic resonance imaging (fMRI) study, participants were presented with speech sounds, visual lip-movements, or congruent *versus* incongruent combinations of the two. Relative to the sum of the two unimodal baseline activations, a region in the left superior temporal sulcus showed over-additive response enhancement during congruent bimodal stimulation, and apparently sub-additive 'suppression' for incongruent bimodal stimulation.

These findings were considered analogous to the non-linear spatial interactions observed by Stein and colleagues. Note, however, that the observed 'suppression' was not strictly equivalent to that found by Stein for spatial incongruence, as activity never fell below the baseline of stimulation in a single modality alone. As with any single experiment, one can quibble with the interpretation of these fMRI results; for example, is the activated area really the site of the effective crossmodal integration, or instead some 'downstream' area responding to the quality of the final speech percept?. Nevertheless, the study is noteworthy for seeking to generalise simple principles of crossmodal integration, and for testing these with fMRI. Indeed, functional imaging seems a particularly suitable

method for addressing crossmodal issues. Much of the groundwork of characterising sensory responses to individual modalities has already been done, and one can now exploit this to test for commonalities versus differences across modalities in terms of the activated brain areas [11,12].

Orthogonal crossmodal influences on 'unimodal' perception

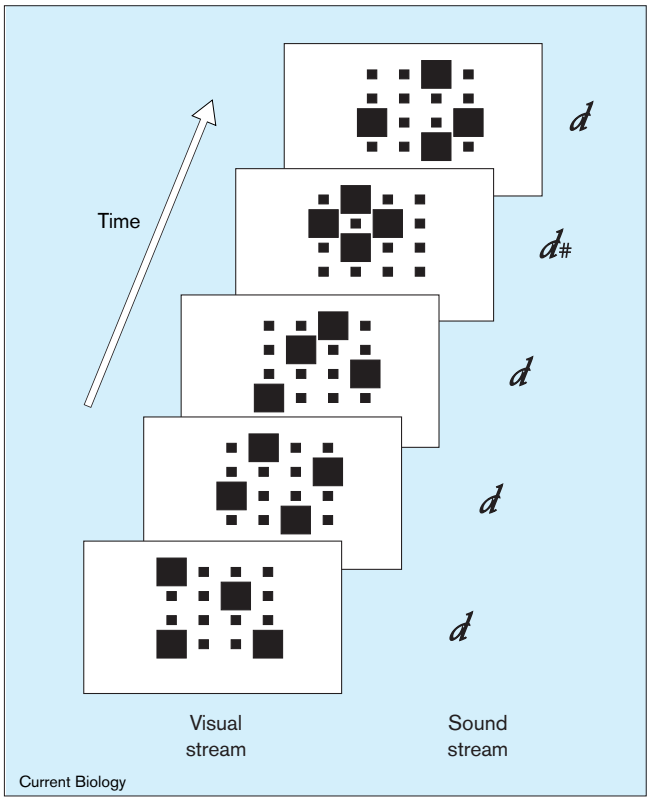
In the above examples, different modalities provide convergent evidence about some particular external property. It seems only natural that these sources of evidence should therefore be combined, and perhaps equally natural to model this in terms of feedforward convergence. But in other crossmodal interactions, judgements for one modality are influenced by a second modality, even when the latter modality can provide no information about the judged property itself (and so is truly 'orthogonal' to this property).

A recent example comes from Vroomen and de Gelder [9]. They presented participants with a rapid sequence of visual displays. A series of tones was also presented from a fixed location, one concurrently with each visual display (Figure 2). The task was to detect and then localise a diamond-shape that could appear among the visual stream. One of the sounds in the task-irrelevant auditory sequence could be unique, for example, a high tone amongst low tones. Subjectively, the visual display that coincided temporally with the unique sound appeared to segregate from the other visual displays. Objectively, the visual target was better detected and localised if it coincided with the unique sound, even when the latter was entirely non-predictive and could provide no information about the shape of any visual event (thus producing an 'orthogonal' crossmodal effect in our terms).

There are now several further examples of crossmodal effects that are 'orthogonal' in this sense. For instance, one of us [8] showed that ventriloquism elicited by visual lip-movements could lead to an objective improvement in the identification of spoken words during selective listening, even though the varied location of the lip-movements could provide no further information about word identity. Such 'orthogonal' crossmodal effects may require some further explanation beyond the now conventional idea of feed-forward convergence of information from different sources (as in Figure 1a).

One possible account for the situation studied by Vroomen and de Gelder [9] is as follows. Visual events might become grouped with concurrent auditory events by temporal contiguity, achievable within a standard convergent feedforward architecture. 'Popout' of the unique sound among its auditory stream could then lead to accompanying popout of the concurrent visual display from its visual stream. Crucially, this must cascade to affect processing of visual properties, such as shape, that bear no relation whatsoever to the auditory stream itself. Such cascading from one property to another is reminiscent of findings on 'object-based' attention in unimodal visual studies. One way to achieve this in crossmodal situations would rely on feedback from a level where multimodal convergence arises (Figure 1b), to affect 'earlier' levels of unimodal (visual) processing.

Figure 2

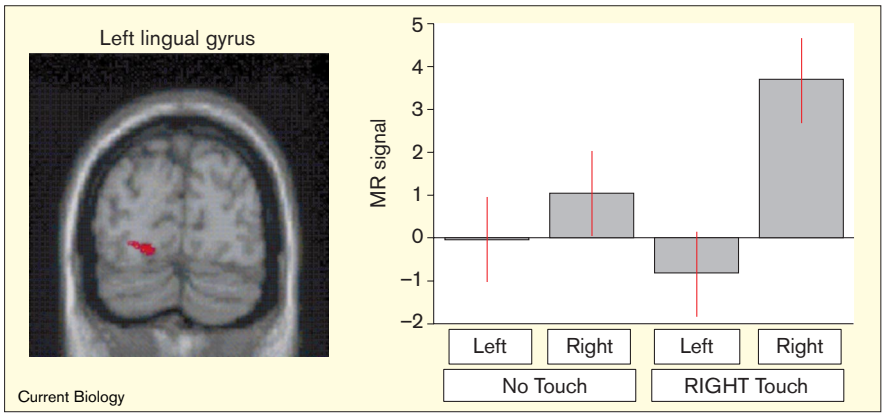


Schematic illustration of a representative sequence of events in the experiments of Vroomen and De Gelder [9]. A rapid series of visual events is presented, each visual event concurrent with an auditory event (illustrated here with the 'musical' notation). The task was to detect (and report the location of) a diamond-shaped visual target (present in frame 4 of the illustration). This was performed better when the visual target coincided with a unique sound in the auditory stream.

Feedback from multimodal to 'unimodal' brain areas

Recent work provides some preliminary evidence for this idea that levels of processing where multimodal convergence arises can feedback to influence levels traditionally considered 'unimodal'. One example comes from the literature on crossmodal spatial attention. Several studies have shown that a spatially nonpredictive tactile cue can lead to enhanced judgments for visual targets presented near to that tactile cue, relative to those presented elsewhere, even for visual properties that are entirely

Figure 3



Illustrative results from the fMRI study of Macaluso *et al.* [10]. The left lingual gyrus (activation shown in red) is traditionally considered a unimodal region of retinotopic visual cortex. Unsurprisingly, it responded more strongly to a contralateral (right) visual stimulus than to an equivalent stimulus in the ipsilateral (left) visual field (see leftmost two bars in histogram). The novel crossmodal observation was that the visual response to a light in the right visual field was enhanced by concurrent touch at the same right location (rightmost bar), even though the lingual gyrus did not respond to touch *per se* (see second bar from right).

orthogonal to the nature of the tactile cue [5,25]. The conventional account of such crossmodal links in spatial attention would invoke purely feedforward convergent pathways (Figure 1a), but a recent fMRI study by Macaluso *et al.* [10] shows that a spatially congruent tactile cue can enhance neural responses to a visual target within 'unimodal' visual cortex in the lingual gyrus (Figure 3; see also [25] for analogous evidence from an event-related potential (ERP) study).

In a related, albeit non-spatial, vein, Calvert *et al.* [26] reported that visual lip movements can activate primary auditory cortex (see also [37]), again suggesting a feedback influence from multimodal levels of representation. Finally, Giard and Peronnet [28] report ERP evidence that tones synchronized with a visual stimulus, as in the study by Vroomen and De Gelder [9], may influence early visual responses.

Thus, brain areas traditionally considered as 'unimodal', may only be so in terms of their afferent projections. Back projections from multimodal convergence areas could result in responses to the primary modality in 'unimodal' brain areas being modulated by stimulation in a second modality, as shown by Macaluso *et al.* [10]. This may relate to one of the most fascinating aspects of crossmodal interactions, namely that subjective experience within one modality can be dramatically affected by stimulation within another. The finding that lip-movements can affect identification of a spoken word [13] may be less remarkable, given that the lips add information, than the fact that the same speech actually *sounds* different when different lip movements are seen. This subjective fact may relate to the presence of feedback pathways from convergence zones [6,10]. Such feedback could produce a multimodally determined percept which nevertheless has the unimodal qualia associated with the activation of brain areas receiving afferent input from only one primary modality.

Coda

In this brief article, we have been able to consider only a few of the current issues in research on crossmodal integration, leaving other important issues — such as development, plasticity and the effects of brain damage — entirely untouched [23,29,30]. We hope we have said enough to illustrate that significant progress is being made in the field. Moreover, recent methodological breakthroughs, such as the development of event-related fMRI, offer much for the future in this field.

Acknowledgements

The authors' crossmodal research is supported by a Programme Grant from the Medical Research Council (UK).

References

1. Urbantschitsch V: **Über den Einfluss einer Sinneserregung auf die übrigen Sinnesempfindungen**. *Archiv f d gesch Physiol* 1880, **42:**155 ff.
2. Welch RB, Warren DH: **Intersensory interactions.** In *Handbook of Perception and Performance, Vol. 1, Sensory Processes and Perception*. Edited by Boff KR, Kaufman L, Thomas JP. John Wiley and Sons: New York; 1986.
3. Stein BE, Meredith MA: *The Merging of the Senses*. Cambridge, MA: MIT Press; 1993.
4. Bertelson P: **Ventriloquism: a case of crossmodal perceptual grouping.** In *Cognitive Contributions to the Perception of Spatial and Temporal Events*. Edited by Ashersleben G, Bachmann T, Müsseler J, Elsevier Science, B.V.: Amsterdam; 1999:347-362.
5. Driver J, Spence C: **Attention and the crossmodal construction of space.** *Trends Cognit Sci* 1998, **2:**254-262.
6. Massaro DW: **Speechreading: illusion or window into pattern recognition.** *Trends Cognit Sci* 1999, **3:**310-317.
7. Calvert GA, Campbell R, Brammer MJ: **Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex.** *Curr Biol* 2000, **10:**649-657.
8. Driver J: **Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading.** *Nature* 1996, **381:**66-68.
9. Vroomen J, de Gelder B: **Sound enhances visual perception: Cross-modal effects of auditory organization on visual perception.** *J Exp Psychol* 2000, in press.
10. Macaluso E, Frith C, Driver J: **Modulation of human visual cortex by crossmodal spatial attention.** *Science* 2000, **289:**1206-1208
11. Macaluso E, Frith C, Driver J: **Selective spatial attention in vision and touch: unimodal and multimodal mechanisms revealed by PET.** *J Neurophys* 2000, **83:**3062-3075.

12. Downar J, Crawley AP, Mikulis DJ, Davis KD: **A multimodal cortical network for the detection of changes in the sensory environment.** *Nat Neurosci* 2000, **3:**277-283.
13. McGurk H, MacDonald J: **Hearing lips and seeing voices.** *Nature* 1976, **264:**746-748.
14. Radeau M: **Auditory-visual spatial interaction and modularity.** *Curr Psychol Cogn* 1994, **13:**3-51.
15. Spence C, Driver J: **Attracting attention to the illusory location of a sound: reflexive crossmodal orienting and ventriloquism.** *NeuroReport* 2000, **11:**2057-2061.
16. Pavani F, Spence C, Driver J: **Visual capture of touch; out-of-the-body experiences with rubber gloves.** *Psychol Sci* 2000, in press.
17. Saldaña HM, Rosenblum LD: **Visual influences on auditory pluck and bow judgments.** *Percept Psychophys* 1993, **54:**406-416.
18. Jousmäki V, Hari R: **Parchment-skin illusion: sound-biased touch.** *Curr Biol* 1998, **8:**R190.
19. DuBose CN, Cardello AV, Maller O: **Effects of colorants and flavorants on identification, perceived flavor intensity, and hedonic quality of fruit-flavored beverages and cake.** *J Food Sci* 1980, **45:**1393-1399.
20. Welch RB, DuttonHurt LD, Warren DH: **Contributions of audition and vision to temporal rate perception.** *Percept Psychophys* 1986, **39:**294-300.
21. Regan D, Spekreijse H: **Auditory-visual interactions and the correspondence between perceived auditory space and perceived visual space.** *Perception* 1977, **6:**133-138.
22. Bertelson P, Vroomen J, Wiegeraad G, de Gelder B: **Exploring the relation between McGurk interference and ventriloquism.** *Proc Int Conf Spoken Lang Process* 1994, **2:**559-562.
23. King AJ, Schnupp JWH: **Sensory convergence in neural function and development.** In *The New Cognitive Neurosciences.* 2nd Edition, Edited by Gazzaniga MS. Cambridge, MA: MIT Press; 2000:437-450.
24. Wallace MT, Meredith MA, Stein BE: **Integration of multiple sensory inputs in cat cortex.** *Exp Brain Res* 1992, **91:**484-488.
25. Kennett S, Eimer M, Spence C, Driver J: **Tactile-visual links in exogenous spatial attention under different postures: convergent evidence from psychophysics and ERPs.** *J Cogn Neurosci* 2000, in press.
26. Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, Woodruff PW, Iversen SD, David AS: **Activation of auditory cortex during silent lipreading.** *Science* 1997, **276:**593-596.
27. Sams M, Aulanko T, Hamalainen H, Hari R, Lounesmaa OV, Lu DT, Simola J: **Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex.** *Neurosci Lett* 1991, **127:**141-145.
28. Giard MH, Peronnet F: **Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study.** *J Cogn Neurosci* 1999, **11:**473-490.
29. Lewkowicz DJ: **The development of intersensory temporal perception: An epigenetic systems/limitations view.** *Psychol Bull* 2000, in press.
30. Ladavas E, di Pellegrino G, Farne A, Zeloni G: **Neuropsychological evidence of an integrated visuotactile representation of peripersonal space in humans.** *J Cogn Neurosci* 1998, **10:**581-589.