Cross-modal cuing and selective attention

Austen Clark
Department of Philosophy
103 Manchester Hall U-2054
University of Connecticut
Storrs, CT 06269-2054

Presented at the University of Glasgow Conference on "Individuating the Senses", organized by Professor Fiona MacPherson, Department of Philosophy and Centre for the Study of Perceptual Experience, 4 and 5 December 2004. This is the final draft of May 2005, under review for publication as part of the volume of conference proceedings.

Experiments on cuing have long provided insights into the mechanisms of selective attention. A visual cue presented in a particular location can enhance subsequent visual discriminations at that location, making them faster, or more accurate, or both. The standard interpretation of such experiments is that the cue attracts attention. Subsequent stimuli at that location are then more likely to be attended as well, and thus they are more likely to receive quicker or more thorough processing.

Only recently has this paradigm been applied across modalities. In a cross modal experiment, the cue is in one modality, and the target in another. The results are profoundly interesting: across many pairs of modalities, a cue in one can help direct attention in the other. Some of these results will be described. They are interesting because they have profound implications about pre-attentive perceptual processing, about the mechanisms that direct selective attention, and about the character of the representations needed in order to do that directing successfully. The bulk of the paper will attempt to draw these implications from the results.

1. Some results

The argument will focus largely on the fascinating work done by Charles Spence, Jon Driver, and associates, starting in the 1990's. The experimental designs build on the "spatial cuing" paradigm of Michael Posner. A subject stares at a fixation point on a screen, and both cues and targets are presented on one or another side of that fixation point. Eye movements are controlled (perhaps tracked with an eye tracker), so that all of the orienting involved is "covert" orienting. Cuing can be of two sorts, which have come to be known as "endogenous" and "exogenous". In the "endogenous" form the occurrence of a cue is highly correlated with the subsequent presentation of a target on the same side. The subject is given time to learn this correlation, and then after each cue is presented, is given enough time to shift attention deliberately to the cued side. The deliberate shift of attention is "endogenous".

In contrast, "exogenous" cuing corresponds more closely to some salient or surprising stimulus grabbing one's attention, reflexively and involuntarily. These experiments try to prevent the subject from forming or acting on an expectation based on the cue. They do this in two ways. First, the experiment is set up so that the cue is non-predictive: all the conditional probabilities of target location, given the cue, are the same. Second, the target also appears more quickly, within 100-300 milliseconds after the cue terminates.

Dependent variables are reaction times and accuracy. Posner found that with a sufficiently salient cue, a target subsequently presented on the same side ("congruent") could be detected significantly faster than a target presented on the side opposite the cue ("incongruent"). Similarly, in the more leisurely and predictable endogenous cuing, a target presented on the same side as the cue could be detected more quickly, and discriminated more accurately, than targets presented on the opposite aide. The interpretation of both effects is the same: the cue attracts selective attention to its location. Once one is attending to that region, subsequent targets presented in the same region are more likely to be selected by selective attention for further, central processing. If targets receive more processing, they are discriminated more quickly, more accurately, or more thoroughly. So, the interpretation concludes, those targets are likely to be discriminated more quickly, accurately, or thoroughly than targets presented on the incongruent side.

Key features of this explanation have an even older lineage. Selective attention is thought to be a process that selects some representations but not others for further, "central" processing. To use various of the metaphors that have been suggested: it opens certain filters, or gates, or channels, or it allocates bandwidth, or focuses the spotlight, or adjusts the zoom lens, so that some stimuli are selected (and processed further), and others are not. Receipt of that further processing explains why some stimuli are processed more quickly, more accurately, or more thoroughly than others. In endogenous cuing the selection is deliberate, top-down, and voluntary; in exogenous cuing the cue reflexively attracts or grabs attention, and holds it to that region for a moment or two.

This same theoretical terminology is used throughout the reports of cross-modal cuing. The cross modal experiments are similar to Posner's, except that the cue is presented in one modality, and the target in another. For example, Spence and Driver (1996, 1997) have a subject staring a fixation point at the center of a screen. At each of the four corners of the screen there is a light, and behind that light, a loudspeaker. The experimental task is to discriminate whether the light or the sound came from an upper corner or a lower corner. If one presents a non-predictive visual cue on one side of the fixation point, a subsequent visual target on the same side is processed more quickly and accurately, just as Posner found. What is new is the finding that if one presents a non-predictive visual cue on one side, a subsequent auditory target on the same side is likewise processed more quickly and accurately. Targets at incongruous locations are discriminated with less speed and lower accuracy.

Cross modal cuing can work in the reverse direction as well (that is, from an auditory cue to a visual target). The mystery is: Which modality tells you how to do this? Audition cannot locate the light. Vision cannot place the sounds. The same problem recurs in all the cross modal effects. We seem to navigate effortlessly from one modality to another even though no one modality has the wherewithal to represent the entire route. So how do we do it?

The problem recurs across other pairs of modalities. Spence and Driver have demonstrated exogenous cross-modal cuing between vision and touch, in both directions. In one design, for example, subjects hold a sponge in each hand between the index finger and the thumb. There is a small vibrator (a vibro-tactile device) in the sponge near each digit, and also small LEDs that can light up near each digit. A non-predictive cue is given to one hand in one modality; then an up/down discrimination must be made in the other modality. Just as with the audio-visual links, Spence and Driver found that discriminations are faster if the cue was on the same side as the target. They have likewise found cross modal cuing between touch and audition.

It should be emphasized that some experiments have failed to demonstrate exogenous cross modal cuing; details of the task given to the subject and of the exact experimental paradigm used can make a big difference. Some cue-target pairs seem more delicate than others; visual cuing of auditory targets in particular has failed in a number of studies. Such cuing may only arise in specific experimental paradigms. Spence, MacDonald & Driver (2004) summarize the situation as follows:

> The repeated positive findings of cross modal exogenous spatial-cuing effects within paradigms that avoid many of the more obvious pitfalls inherent in the early research still provide existence-proof that such cross modal influences can (sometimes) arise, as also confirmed with the various other methods that we review below. The existing failures to observe exogenous spatial-cuing effects for some particular pairings of modalities might reflect specific details of the paradigms used.... (Spence, MacDonald & Driver 2004, 291).

For the purposes of this paper, the important point is not whether cuing always occurs, but that it can sometimes occur. In all the pairwise combinations among vision, audition, and touch, and in both directions, exogenous cross modal cuing effects *sometimes* arise. To raise the conceptual issues, *once* would be enough.

a. Some initial unpacking

The first step in explaining these phenomena is to propose that the cue draws attention to its location: "our cuing effects must reflect a genuine improvement of localization in the cued region ... owing to the cue attracting covert attention there" (Driver & Spence 1998b, 1320). The next step: once drawn to a locale, attention "spreads". Within one modality it can spread from the "cued region" to regions in the vicinity. Even more powerfully, it can spread across modalities. At least that is the way Driver and Spence put it: "Our findings suggest that a strongly biased spatial distribution of endogenous attention in one modality tends to spread into other modalities as well, but at a reduced level" (Driver & Spence 1998b, 1325).

The locution is intriguing. What does it mean to say that attention in one modality tends to "spread" into other modalities? If one is spreading jam, for example, where can one find the toast that would connect sights to sounds?

To unpack this a bit, we need to know how to interpret talk about the "spatial distribution of attention", and of attention being "drawn" to a cue. We would like to spell out how attention is disposed at a given moment, so that we can then describe how that disposition changes over time, and in response to what. Attention at a given moment could be characterized by describing which representations it has selected, and which it has not. Per hypothesis there is some selectivity operative at that moment: some representations are selected for further, central processing, and some are not. What is this principle of selection, and how can it be described? In the perceptual domain, at least, one could catalog it exhaustively by listing all the stimuli that are possible at time *t*, and for each one, giving the probability that, at time *t*, it is selected for further, central processing. This would be a long list, and it would be impossible to complete; but the important point is that the list has to include some merely possible stimuli--stimuli that might have happened, but did not.

The current state of selectivity of the system is a dispositional property; it includes propensities to respond to various possible stimuli. One might be paying very close attention to whatever sounds come out of a room, but do nothing, since no sound ever emerges. Nevertheless sounds from that room would be processed with alacrity, were any to occur. What selectivity means is that one is disposed to process a certain subset of stimuli more thoroughly or quickly than others. To characterize that disposition adequately, the list must include some stimuli that in fact did not occur, but might have.

I will call such a list (of possible stimuli each associated with a probability of being selected) a "selection principle". It characterizes the current state of selectivity, or the principle operative in the system at a moment: how the system is disposed to select some stimuli, and not others, for further, central processing.

The idea gives a simple way to describe the difference between "location-based" and "object-based" selection principles. Scan all the stimulus arrays possible at time *t*, and collect all the stimuli whose probability of being selected for further processing is greater than some arbitrary cutoff value. If the selection is location-based then all those stimuli will be clustered in some region of space, whose boundaries do not necessarily correspond to any identifiable object, but are perhaps delimited in terms of spatial relations to the sense organs. Anything that might happen in that location is more prone to be selected subsequently. If the selection is object-based, then all those stimuli will cluster on some object, or objects, or proper parts thereof. Anything that might happen to some part of some selected object is more prone to be selected subsequently.

When the focus of attention changes, these odds change, but nothing else need change overtly. Told to pay attention to the lights, the odds of the subject performing quick visual discriminations go up, and of quick auditory discriminations go down. Faced with exactly the same stimulus array, but told to pay attention to the sounds, the odds of quick auditory discrimination go up, but those of visual ones go down. This change in selectivity can happen in the absence of any overt behavior; our subject sits motionless in both cases, but in one is paying attention to the sights, and in the other to the sounds.

Now the fundamentally interesting finding of the vast literature on cuing and priming is that perception of a cue can *change* the selectivity operative in the system. Even more exciting are the cross modal findings that it can change the selectivity operative in other modalities. How does it do this? More importantly for our purposes, what information must the system have available in order to implement those changing odds? It is vital to understand *what* the system needs to represent in order to manifest such selectivity, and *how* the system represents what it needs to represent to effect such changes in its own operating parameters.

b. How to be selective

An organism graced with selective attention can alter the flow of information within itself. With a shift in attention, some pathways are relatively enhanced, and others depressed. Some bits of information are now more likely to receive central elaboration, others less so; some channels have faster effects, and others are demoted. Something within the system must bring about these changes in the way information flows within it--changes in its selectivity--and let us call that thing the "Selectivity Effector". It is that which effects the changes in selectivity that are manifest when the focus of attention shifts. There are various models of what a Selectivity Effector might be, and many use metaphors of gates and switches, amplifiers and filters, pathways and channels, within electronic circuitry (see LaBerge 1995, 39). So a Selectivity Effector might be something like a "switch" in a computer network: a device that establishes different links between different parts of the network at different times, depending on the projected transmissions and the existing traffic. By opening and shutting gates and switches in different ways, such an Effector could alter the flow of information within the system, and thereby change the odds that some bits and not others receive further, central processing.

The job might get done in various ways, but however it is done, it is clear that the Effector needs certain critical bits of information in order to do its job. If our experimental subject is to pay attention to the lights and not the sounds, for example, the Selectivity Effector must accurately register which gates and switches control visual processing, and which ones control audition. It would be embarrassing, and perhaps disastrous, to confuse them. If a loud sound is to draw attention to its source, the Selectivity Effector must open just those gates and switches needed to allow processing of stimuli from locations *near* that source. Here too mistakes might be disastrous. Rewiring oneself on the fly is a risky business.

The Selectivity Effector must then be guided, and guided well, when it does its job. If you're going to rewire yourself, you need accurate information about where the wires are, and where they need to go. Let us drop the electrical metaphors, and try instead to characterize that which provides the information needed to direct the Selectivity Effector when it does its job. Such a "Director" must extract and pass along the information needed in order for selective attention to make appropriate selections. It does not make those selections itself. It passes that information along to whatever parts of the system change the selectivity operative at a given moment. The focus here is not how the selections are implemented, but rather what sort of information is necessary for those selections to be appropriately sensitive to the environment.

Start with the very simplest case: exogenous location-based cuing within one modality. Cue *x* might be a flash of light, target *y* a flash nearby. The proferred explanation is that *x* attracts attention to its location, and, given the appropriate stimulus onset asynchrony (or "SOA") a subsequent stimulus at that same location or nearby is thereby more likely also to be selected. The information needed to specify how subsequent targets are selected is, simply, their location. The Selectivity Director needs tell the Selectivity Effector where the cue occurred, but nothing more.

Notice though that this description over-simplifies this simplest case. The mere fact that target *y* is in a region that cue *x* previously occupied will not suffice. We need to show that the system picks up and registers these facts. The cue must be perceived as occurring in region *R*, and the target must be perceived as occurring in, or near, the same region. If there is a cue in region *R*, but it is not perceived, then there is no cuing effect to explain. Conversely, it is not the actual location of the target stimulus that matters, but rather the location it is perceived to have. It must be perceived to be in the same general location as the cue; and as long as it is so perceived, we can get a cuing effect.

These "regions" are, therefore, perceptual; they are locations *as* perceived, to be specified by whatever terms the given modality musters when it manages spatial discriminations. The boundaries of such regions are more or less vague, depending on the accuracy of spatial discriminations possible in that vicinity (or in that direction). Some of these "regions" might not have limits in all three orthogonal axes; they may be delimited in only two, or even one, such direction. "On the horizontal plane" identifies a "region" in this sense: we know the elevation of any point in that region, but that is the only axis that is specified. Its azimuth and depth are left unspecified.

Audio-visual cuing provides examples of regions that have vague boundaries along some spatial axes and that totally lack such boundaries along others. Auditory discrimination of differences in the elevation of sounds is relatively poor, while discrimination of their azimuths is relatively good. So if a sound draws attention to a certain region, its horizontal extent is more precisely delimited than its vertical. Depth is very poorly discriminated, and might not be needed at all to manage cuing. The "region" of the sound can be thought of as the direction from whence it comes: a cone centered on azimuth theta, elevation epsilon, extending outwards in space. Like "on the horizontal plane" this region does not specify a depth at all. Any point within that cone would qualify. Such a region might be the R_m picked out by auditory spatial discrimination.

But with this amendment we get a possible location-based account. The cue attracts attention to a visually represented region R_m. Once attention is attracted to that region, it stays awhile, and perhaps spreads a bit. Then, given the appropriate SOA, a subsequent target which is perceived as occurring in or near region R_m is more likely to be selected by selective attention. That is, given the appropriate SOA, the only information needed to specify whether a subsequent target is likely to be selectively enhanced is whether or not it is perceived as occurring in or near that same region. This is a simple, "location-based", selection principle.

In contrast, if selective attention works in an object-based way, then to write down its selection principles we must at least sometimes use terms identifying objects. Selection is not solely dependent on perceived location; instead it sometimes ranges over elements that are perceived to be parts of the same object as things already selected. Of course the theorist here has the burden of specifying what is meant by "same object". It will in any case be an object as perceived: something represented by the system as the same object. But, with that proviso, there is no dearth of accounts of what might constitute these "perceptual objects". The central idea is that target *y* is more likely to be selectively enhanced if it is perceived as being part of the same object as was *x*.

2. Exogenous cuing without remapping

Now at last all the pieces are on the board and we can consider the conceptual implications of cross modal cuing. In this section and the next I will list a number of these implications. I think they follow from the models and results reported in Spence and Driver (2004).

In the cross modal case, the cue is in one modality, and the target in another. I will use subscripts to flag this fact. Cue x_m occurs in modality *M*. It is perceived to be in region R_m. The cue might be a sound, for example, and it draws attention to an acoustically discriminated region. The subsequent target y_n occurs in some distinct modality *N*. The region in which it is perceived to occur is delimited by a different modality than the one that delimits region R_m. The target might be a light, for example, and its location is discriminated visually. So the region the target is perceived to occupy will be dubbed S_n--a region as discriminated by modality *N*.

For many pairs *M*, *N*, cuing in modality *M* can improve the speed and reliability of discriminations in modality *N*. It has been demonstrated in both directions, and in every possible pairing, among vision, audition, and touch. The first inference drawn from this is that cuing in modality *M* can help to direct attention in modality *N*. In particular:

2.1 Cuing in modality *M* can help one select some subset of representations in modality *N* for further, central processing.

Given the theoretical assumptions sketched above, this inference is straightforward, since one finds that some discriminations in the target modality *N* are completed more quickly or more accurately given the cue in modality *M*. So the changes in selectivity given a cue are not confined to the modality of the cue. An implication from that:

2.2 A cue in modality *M* can provide information sufficient to yield some principle in terms of which some representations in modality *N* are selected.

For shifts in selective attention to be adaptive, the mechanism that does the shifting (the Selectivity Effector) must be at least somewhat sensitive to features and events in the immediate surroundings. The task of picking out the next thing to be attended itself requires perceptual information. If for example one is selecting targets found in some region, then information about the extent of that region must be accessible to guide selective attention. However it is done, there must be some route whereby some perceptual information can make it to the mechanisms that guide attention. Otherwise the selections made would have no reliable correlation with any features of the perceptible environment.

Cross modal cuing shows decisively that information gleaned from one modality can serve to guide the selections applied within other modalities. The selection principle seems to be overlap of location: the cue draws attention to a region, and then one can discriminate stimuli in other modalities in that same region more quickly, or more accurately, or both. In common sense terms: a sound can draw your attention to a region, and thereafter, for a while, visible events in that vicinity are likely also to be perceived more quickly or more accurately, or both. Notice that any such common sense description treats the "region" as something that contains both auditory and visual stimuli. Events in that "region" can be perceived by any of various modalities; it is similar to a common sense "physical object", which is something that can be both seen and touched, for example. For this reason these regions or objects are sometimes called "distal", or "allocentric", or "external"; but the critical point is that these modifiers do not indicate that we have thereby identified some new kind of space. Instead, there is, and always has been, just one space in which anything that can be found is to be found. What these modifiers indicate are variations in the ways of identifying regions in that one space. A "visual location" is not a new kind of location, but simply a visual way of indicating a location in the one and only space.

Cross modal cuing requires that locations be indicated in ways not limited to any one modality; that (for example) a visual specifier of location and an auditory specifier of location can specify the very same place (or places that largely overlap), albeit in different manners.[1] It by no means a trivial matter to establish that two distinct specifiers both point to the same region, yet it must be done, and done correctly, if selective attention is to be guided in an adaptive way.

The task is to formulate, in modality *M*, a selection principle that would be operative in modality *N*. How might one pick out a visible region in acoustical terms? Or a visual region in tactile terms? In modality *M* one makes the spatial discriminations needed to delimit region R_m, but the target occurs in a different modality, and the extent of the region in which that target occurs is delimited by entirely distinct features S_n. So forging this link is the critical step for correctly guiding the formulation of a new selection principle. One needs to rely on the fact that the acoustic specifiers of location R_m pick out a region that largely overlaps the one indicated by the visual specifiers of location S_n. Such guidance must be something the system can represent and compute, with results that are mostly reliable. Only then could that which effects the selectivity of the system do so in an adaptive way.

One condition on possible solutions is relatively straightforward:

2.3 If the cross modal selection principle is location-based, then information about locations as represented in modality *M* must be commensurable with information about locations as represented in modality *N*.

By "commensurable with" I mean that one can be derived as a function of the other.[2] Unless they are commensurable we cannot use these "specifiers of location" to formulate a cross-modal selection principle. If there were not even a statistical association between them, cues in one modality would not enable one to make any predictions at all about events in another.

Consider the cuing from auditory to visual events as an example. As mentioned, the auditory "region" might be specified purely in terms of a direction: a cone, specified by a range of azimuths and elevations, giving the direction (relative to the two ears) from which the sound appears to emanate. Any point within that cone can qualify; depth is not even specified. Similarly, as long as we have a fixed eye position, many of the cuing tasks could be managed with nothing more than an analogous, visual "direction". In vision the azimuth and elevation might both be defined by angles from the visual fixation point, and both of them would be much more precisely delimited than is possible in audition. But if an acoustic cue could be massaged to yield such a visual direction, then the cuing results could be explained. Visual depth is not relevant to the task; as long as we could direct selective attention to select any visual events occurring in or near that cone, we would selectively enhance

[1] Identity would require the boundaries of the indicated regions to be exactly congruent, which would be rare in the cross modal case, given the differences in spatial discrimination in the different modalities. Overlap does not require congruent boundaries, but nevertheless implies a "partial" identification. That is, some of the places indicated by one modality are identical to some places indicated by the other.

[2] The function might be probabilistic and error-prone. At the very least we need some statistical association: one can be predicted from the other with odds that are at least somewhat better than chance.

the appropriate targets.

2.3 claims that if we observe auditory cuing of visual discriminations, then these visual directions must be commensurable with auditory directions. Region R_m is the auditory direction (or cone); S_n the visual. Commensurability is easy: given a fixed eye position, auditory azimuths and elevations can readily be converted into visual ones.

Notice that even if the specifiers of locations R_m and S_n happen to specify overlapping regions, it is only if the system somehow trades on that overlap that we can justifiably claim that the system represents them to be overlapping. The auditory system does not "know" that R_m picks out a region that also has an visual specification, and likewise for the visual system. So even though in fact these locations are (as always) locations within a "common space", to this point they are not yet represented *as* locations within a common space. To this point there is nothing to show that the system represents auditorally specified regions as being, even potentially, the same as visually specified ones.

But if cross modal selection principles are ever used, then the situation changes:

2.4 If the overlap of regions specified by R_m and S_n is used by the system to generate a selection principle applied to targets in modality N, then the system represents those regions *as* overlapping.

That is, if the system is using R_m to create instructions so as to guide the system to select S_n, then the system itself is trading on the identification. We find downstream consumers of the information, who rely on the world being as represented. Here the downstream consumer is selective attention itself.[3] It relies on the correct specification, in visual terms, of "the region" where the acoustic cue occurred. Specifically, the Selectivity Effector can do its job successfully only if the world is such that the region indicated by R_m overlaps that of S_n. So not only is it a region common to the two modalities; it is now also represented *as* a common region. Such talk is not justified until this stage, but here at last it is.

3. Remapping

The exogenous cross modal cuing experiments discussed so far give rather weak grounds for thinking that the system is trading on, or relying on, identification of locales. Those grounds are bolstered enormously when we consider the phenomena that Spence and Driver call "remapping". Remapping arises when we remove one of the constraints on the task assigned to the subject. So far all the experiments discussed require the subject to maintain a fixed posture. All the orienting is covert: the gaze does not shift, the head is fixed, the arms are immobile, held in a neutral posture. What happens if we allow our subjects to move?

[3] The methodology and terminology here both derive from "consumer" semantics, also known as "bio-" or "teleo-" semantics. See Millikan 1984, ch. 6; Millikan 2004.

For example, the visuo-tactile experiments thus far all proceeded with the subject's hands and arms in a neutral position, with the right hand on the right side of the body, left hand on the left side. But what happens if you cross your arms? The question seems a simple one, but only in retrospect: this experiment was first tried in 1996. If the selection functions were "hard wired" then a vibration on the right hand would still draw attention to some region on the right side of the body, no matter where the hand is. If the system were a little better designed, vibration on the right hand could draw attention to stimuli on the left side of the body. The latter is what is found.

Even more compelling is the converse result, using visual cues for tactile targets. In the neutral position the light on the right side will speed tactile discriminations on the right hand. If the arms are crossed, the light on the right side comes to speed tactile discriminations on the left hand, which is now located closer to that light than is the right hand. Notice that this "closer to" is a cross modal relative distance: with the arms crossed, tactile stimuli on the left hand are now closer to the light source than are tactile stimuli on the right hand. This "closer to" cannot be assessed by touch alone or by vision alone; it requires representation of the place of the light and the place of the hands in a common space. We must locate the light in a system of places and relative distances in which we can also locate the tactile stimuli on the two hands. Otherwise no sense could be given to the proposition that the light on the right side is now closer to the left hand. If the latter explains why attention spreads from the light to the left hand, then the system itself must have access to such "cross modal" or "common" distances. Its representation of and facility with these cross modal distances is revealed when it formulates a new selection principle, using the location of the light on the right side to direct attention to the left hand.

Similar results are found for audio-visual cross modal cuing. Label the loudspeakers *A*, *B*, *C*, *D*. Loudspeaker *B* is the second one from the left. In the first set of trials, the gaze is focused between speakers *B* and *C*. Sounds from loudspeaker *B* will speed processing of visual stimuli that occur to the left of the fixation point. The gaze is then shifted to a point between speakers *A* and *B*. After that shift in gaze, if the system were well designed, sounds from loudspeaker *B* should speed processing of visual stimuli in a region immediately to the right of fixation point. The latter is what was found.

We already knew the selection principles include multiple modalities. Now we learn that creating a new cross-modal selection principle is not a hard-wired affair, but is sensitive to and permuted by current bodily position. Selectivity is modified on the fly as posture changes.

This adds a new wrinkle to the explanation of how a cue can improve the processing of a target in a different modality. If the subject's head position is constant, then sounds from loudspeaker *B* will, both before and after the shift in gaze, still draw attention to the auditorally defined direction R_m--a sound near the sagittal plane, but a bit to the left. With the shift in gaze, attention drawn in that same auditory direction (R_m) will not spread to the same visually specified region (S_n). Previously it would have enhanced processing of visual stimuli seen to the left of the fixation point; now it enhances processing of some to the right. A different set of visual specifiers S_n' must now be used to give the visual location of the region that attracts attention.

What is the relation between the old S_n and the new S_n'? Or, to put it another way, after a shift in gaze, how do we determine the visual locations of the regions to which auditory cues will now draw attention? Common sense has a short answer: the loudspeakers have not moved, and so attention will continue to be drawn to the same place. As Driver & Spence put it:

> Sounds drew visual attention in the 'correct' direction with respect to external space, even when the eyes were deviated in the head. This entails that the mapping between auditory locations and retinal locations, which directed exogenous cross-modal attention, must have changed, to keep vision and audition in register as regards external space. (Driver & Spence 1998b, 1324)

It would again be a mistake for philosophers to read this as implying that "external space" is a kind of space to be contrasted with all the other kinds of space. Instead the issue is how the one and only space is represented. If it is represented *as* external, then the routes to identifying locations within it are not confined to any one modality. One can locate within it at least some of the things one hears, and some of the things one sees, and some of the things one feels. What remapping will add, critically, is that one can also identify some of the places within it proprioceptively, for some of those places overlap places where one's body is located. For now we can state remapping as follows:

3.1 When postures change but the task does not, cross-modal selection principles can change so as to continue to direct attention to the same place.

The same place in "external space", if you like, though I hope in this context the modifier is redundant.[4] Principle 3.1 summarizes some of the experimental results just reviewed, and gives us a simple method for determining the new visual specifier S_n' which is activated after a shift in gaze. To find it, find which visual specifier now specifies the same place that the old one did.

> the spatial mapping between modalities gets updated when different postures are adopted. ... The senses thus remain in useful register, with respect to each other and the outside world, despite changes in posture. (Driver & Spence 1998b, 1323)

Since we can do this, it follows that the system has the requisite means:

3.2 After successful remapping, the system has information sufficient, given representations of stimuli in modality *M*, to select representations in modality *N* that are representations of the same place.

Some theorists describe the earliest visual maps as "retinotopic", which

[4] The modifier is useful only if you want to contrast "external space" (i.e., space represented *as* external) from such things as "visual space" (i.e., space *as* represented visually). To avoid confusion, think of these modifiers as adverbs. They flag different ways of identifying places.

(strictly speaking) would imply that they are *about* the events on the retina. I urge that we think of them as retinocentric but *not* retinotopic. That is, they might use a retina-relative manner of identifying their subject matters, but that subject matter is resolutely distal. The maps are *about* events in front of the eyes, out there in external space. Remapping shows this. How? It shows that the system uses the visual specifiers of location (and auditory specifiers of location, and so on) to track regions of external space. Downstream consumers of the feature maps can ignore the differences among the retinocentric specifiers and instead use them to track the one distal location that they specify. The visual system in effect "sees through" the retinocentric differences (it remaps them) so as to keep a clear eye on their subject matter: the same unchanging region specified.

With each shift in gaze the system that directs attention unceremoniously throws out the old visual specifiers of location, and grabs new ones; and the new ones are precisely those that now specify the same external region that the old ones did. It manifests the utmost indifference to the differences among the specifiers; what matters is the region specified. Its loyalties are not retinotopic. This is precisely what shows us that those visual specifiers of location are used as specifiers; they are swapped in and out as need be to maintain a track on what they specify.

Thus far we have the capacity to remap, and the means to do so. What then are those means? What sort of information does a subject use in order to remap?

When bodily posture changes, the selection principle that relates the cue in modality *M* to the target in modality *N* must be modified, so that the senses remain in register with regards to external space, and attention can be directed to the correct place. Evidently the subject perceives the changes in bodily position, and those perceptions can be used to effect the remapping. So (sometimes) proprioceptive input is used to change the cross-modal selection principle, and to change it in such a way that it continues to select representations that are *of* the same place. Put another way, the selection principles operative between two modalities can be changed, given information arising from a third one.

3.3 Proprioceptive information can be used successfully to modify the operative selection principle, even in cases when the two modalities involved do not include proprioception.

Remapping is particularly straightforward in the audio-visual experiments. It is easy to see how an auditory direction is commensurable with visual ones. With a shift in gaze, an auditory direction must be associated with a new visual direction--a new visual specification of the region to which attention continues to be drawn. But a shift in gaze is a shift in the *direction* of gaze; and this direction is just another direction, commensurable with the other two. So a change in the direction of gaze could be used to remap auditory directions onto new visual ones. Here then is a particular instance of 3.3:

3.3.1 The auditory specifiers of location and visual specifiers of location are both commensurable with a third set: the proprioceptive indicators of the direction of gaze. The latter can be used to adjust the link between the first two.

Oddly enough, even though the link that needs to be reforged connects audition to vision, the information needed to reforge that link is found in neither modality, but instead in proprioception. Proprioceptive information about the direction of gaze can do that job, and sometimes it seems to be the only information available.

The same holds for all the visual-tactile experiments in which subjects are prevented from seeing their hands or arms. To determine which buzzing feeling is closer to the seen light, those subjects must employ proprioception:

> the spatial mapping from particular retinal activations in vision, to somatic activations in touch, gets updated when the hands adopt different postures. This is presumably owing to an influence from proprioceptive signals specifying the current hand position...Thus a third modality (here proprioception) can apparently influence the attentional interactions between two other modalities (here vision and touch). (Driver & Spence 1998b, 1322)

Why proprioception? Spence and Driver suggest a simple answer. Sensory receptors are spread across the body. Changes in posture are important in sensory terms when they change the spatial relations between one set of receptors and another. For remapping purposes, the important information is not about bodily posture per se, but rather what it indicates about the altered spatial distribution of receptors. We need remapping when (and only when) a change in bodily posture changes the spatial distribution of one set of receptors relative to another.

3.4 The proprioceptive information needed in order to effect a remapping specifies the location of parts of one's own body; in particular the spatial disposition of one's receptors.

A overly fancy way to put it: those receptors are represented as located within the same space as the auditory and visual stimuli that one senses.[5] Less fancy, but more accurate: information specifying the spatial disposition of receptor surfaces is commensurable with information specifying the location of stimuli in the various modalities. In these early, stimulus-driven, pre-attentive processes we seem already to find means of identifying locations that can comprise all the various classes of auditory and visual and tactile stimuli, as well as the various receptor surfaces stimulated by them.

Remapping thus adds another sense in which the "common" space is common. Now within it one must also be able to locate one's own body, or in particular, the spatial array of one's own active receptors. It includes not only the things one senses but also the receptors used to sense them. The latter

[5] This is too fancy partly because of the suggestion that the receptors might be located in some space other than the one in which stimuli are found, but also because the representations involved do not employ the concepts of "receptor" or of "space".

must be included if we are to keep straight all the relations among the former. The point provides a simple, sensory analog for a feature of our conceptual framework that was noted by Kant and re-emphasized by Strawson. Here is how Strawson described it:

> It is a single picture that we build, a unified structure, in which we ourselves have a place, and in which every element is thought of as directly or indirectly related to every other, and the framework of the structure, the common, unifying system of relations is spatio-temporal. By means of identifying references, we fit other people's reports and stories, along with our own, into the single story about empirical reality; and this fitting together, this connexion, rests ultimately on relating the particulars which figure in the stories in the single spatio-temporal system which we ourselves occupy. (Strawson 1963, 17)

I suggest that something analogous is found even in humbler surroundings, where we do not yet have thoughts, or identifying reference, or reports or stories, or even words. It is found among the sensory systems of representation employed in cross modal cuing. They too must fit both the things sensed and the receptors used to sense them into one common framework of spatial relations. But we don't need a conceptual framework of "material bodies" for this: it is built into our sensory mechanisms.

Even in that simpler, sensory setting, there are good reasons to think that the system uses its specifiers of location in something like the way that referring terms are used. I have argued elsewhere (Clark 2000) that the way they are used in solving the problem of feature integration (or "property binding") gives us grounds for thinking that the specifiers of location have at least a quasi-referential role. They serve to indicate the subject matter of the representations in ways that are independent of what the representations say about that subject matter.

But cross modal cuing, and in particular remapping, provide further and independent substantiation for this idea. The challenge is to guide selective attention in such a way that its selections will be adaptive. Cross modal cuing shows that when a cue changes the state of selectivity operative in the system, the system relies upon the region of the cue overlapping the region of the target. It has to do so if a cue in one modality can selectively enhance the processing of targets in another. Furthermore, remapping shows quite vividly that the use of different specifiers of location is not merely idle, but that the system is sensitive to, and relying on identifications of, what those specifiers specify. Relations between sets of specifiers in two different modalities depend upon bodily position. Remapping shows that the system is exquisitely sensitive to the proprioceptive registration of that position. It uses proprioceptive information to alter the relation from a cue in one modality to a target in another. When posture changes, the system immediately swaps an old indicator for a new one; and the swap is made so that the new indicator specifies the same region the old one did. Such use demonstrates that all three sets of specifiers must be commensurable with one another, and that they are being used to track locations in space. Otherwise it very is hard to understand why the shifts in selective attention proceed as they do, and how sensory mechanisms could possibly guide those shifts so that they remain adaptive.

4. Locations and objects

A third and final nugget can be extracted from these results. They have some interesting implications for the contrast between "location-based" and "object-based" models of selective attention. The difference between these is easy to describe in terms of selection principles. If selective attention were purely location-based, then all the principles of selection employed by selective attention could be written down with terms whose reference is confined to the locations of stimuli. Whereas if selective attention is object-based, then to formulate the selection principles one must employ terms that refer to objects of some sort. They might be "visual objects" or "proto-objects", but in any case these entities must be distinct from mere locations. One simple distinction is that objects can move, while locations cannot.

It is quite clear that there are powerful object-based effects on selective attention. What I want to consider is the more radical thesis that *all* of the selection principles are object-based; that selection principles *always* employ terms that refer to some kind of object. The system never uses location-based selection. This thesis is suggested in places by Zenon Pylyshyn (2003), who argues

> that focal attention is typically directed at *objects* rather than at *places,* and therefore that the earliest stages of vision are concerned with individuating objects and that when visual properties are encoded, they are encoded as *properties of individual objects*. (Pylyshyn 2003, 181).

I think the cross modal effects pose a difficulty for this claim. A scheme that always and only selects objects runs into problems representing what it needs to represent to make cross-modal links. If for example the representations of "visual objects" represent just their visible features, then it becomes very difficult to explain how a visual object can serve as a cue that directs attention towards some sounds (and not others) or towards some tactile stimuli (and not others).

Part of the issue hangs on what one means by a "pure object-based" scheme, and what the representations thereof may or may not include. For Pylyshyn, for example, a "visual object" is just anything to which a visual index can be attached; its representation may be purely demonstrative, and it might not include any representation of the location of the referent. But if we do represent locations, this scheme treats them exclusively as properties of the indexed items. We have a small number of indexed "visual objects", and each such object might have a location that is stored in its "object file".

One critical question is whether the representation of visual objects allows for their location to be represented in anything other than visual terms. As far as I can tell, this is an open question for Pylyshyn; nothing commits him to an answer either way. Visual objects themselves typically have lots of non-visual properties, since as Pylyshyn notes visual objects typically turn out to be ordinary physical objects, and ordinary physical objects are more than merely visual. But these are properties of the objects themselves; the question is still open whether visual *representations* of such objects do or do not represent anything other than visual features of them.

Both choices face difficulties. One might naturally assume that visual representations of visual objects represent only their visible features. The locations stored in a visual object file would be purely visual. In audition we would likewise have "acoustic objects" whose locations are represented purely auditorally; and "proprioceptive objects" whose locations are purely proprioceptive. The problem is that in no one of these systems could we represent, or even express, a cross modal identification: that this visually identified object is in the same place as that tactile one is not expressible in purely visual terms or in purely tactile terms. It needs both.

The problem has been a familiar one since Berkeley's *New Theory of Vision*:

> The extension, figures, and motions perceived by sight are specifically distinct from the ideas of touch, called by the same names; nor is there any such thing as one idea, or kind of idea, common to both senses. (Berkeley 1709, §127).

I take the claim here to be inarguable: that visual ideas are not the same ideas as ideas of touch.

This problem is not insuperable; one can simply suppose that there is some extra-modal or perhaps supra-modal mechanism that *can* correlate the two kinds of objects. The layered maps in the superior colliculis and multi-modal neurons in parietal cortex provide possible candidates. But there remains a deeper problem if we insist that all the representations involved are object-based: that if information about locations or directions is ever stored, it is always stored as locations or directions of objects. To explain cross modal cuing, we need to represent cross modal relative distances in a much more fulsome manner than seems possible in a small set of object files.

Consider the visual-tactile experiments, for example. A light goes on near the subject's left hand. Subsequent tactile stimuli on the left index finger or thumb are discriminated more quickly than those on the right hand. The location-based explanation is that visual attention is drawn to a region that is visually identified; and then tactile stimuli in or near that region are more likely to be selected for central processing. So tactile discriminations and reaction times in that vicinity improve. An object-based account would propose that attention is drawn to a visual object (the light); and then it spreads to tactile objects that are nearby.

But notice this last step requires that "nearby" must be assessed cross-modally. The left hand is relatively closer to the light than is the right hand. The system must have access to such cross-modal relative distances in order for attention to "spread" appropriately from visual objects to the appropriate tactile ones. The step from one sensory object to "nearby" ones in another modality is essential for the selection. Put another way: given the visual object, the system must have preattentive access to the relative distances between it and any of the discernible tactile objects, the discernible acoustic objects, the discernible proprioceptive objects, and so on. Otherwise attention could not spread to the "closer" ones of the discernible candidates.

Unfortunately, once this problem has been recognized, it spreads everywhere. We cannot simply store one location per object, since objects have disparate parts, spread out in space. Stimuli that attract attention might be associated with any such part. It won't do, for example, to store exactly one location for the entirety of the felt left hand; the up/down discriminations require one to discriminate the vibrations on the tip of the index finger from the ones on the side of the thumb. One can imagine innumerable variations of this experiment employing tactile stimuli on many different places on the hand. For these we would need to represent relative distances among the possibly distinct places of all the different parts of the hand. If these are each distinct "tactile objects", then tactile objects essentially become equivalent to any discriminably distinct place: any place of any part of the hand that one can feel to be distinct from other such place. That would be a lot of places.

It becomes impracticable to store all these relative distances as properties within each object file.[6] Each such file would have to include distances from each part of the object to all current visual, proprioceptive, or auditory stimuli that might attract attention. Finally, the specifiers of location must be cast in terms that all three modalities can employ, so that the cross modal relative distances could be expressed. I submit that the upshot is essentially equivalent to a location-based scheme. That a vibration on the hand can cue auditory attention to a particular region requires a preattentive capacity to compare locations and distances across the two modalities. It can be described rather awkwardly as relative distances between tactile objects and auditory objects, but the scheme of relative distances required is one that cannot be confined to any one modality, so the talk of "tactile objects" and "auditory objects" becomes otiose. Much simpler then is to acknowledge that the specifiers of location specify locations, and they do so independently of what particular objects, or what kinds of object, occupy those locations.

The audio-visual cuing experiments provide vivid examples of what I mean by saying that visual objects and auditory objects would become otiose. Such cuing could be accomplished with nothing more than specifiers of visual and auditory *directions.* Suppose, as in the later experiments, we use designs in which differences in elevation drop out as irrelevant. Then we need the direction of the gaze relative to the head, the angle between the focal point of the gaze and the visible stimulus, and the direction of the sound relative to the head. The task reduces to converting an auditory azimuth to a visual one, or vice-versa. So the audio-visual cross modal cuing effects require that one compare directions of sounds to those of visual stimuli, but nothing more than

[6] It might be objected that they do not need to be stored, since the distances are out there in the world, and to update them one need merely probe the world perceptually. The problem with this response is that such relative distances have pre-attentive effects. Directing a perceptual probe at the world requires attention already to be directed correctly.

that.

The parsimonious hypothesis is that we have what we need, but no more. We need visual directions and auditory directions to solve this problem, but we do not need visual objects or auditory objects. Even if we had visual objects and auditory objects, they would not gain employment when this problem needs to be solved. It can be solved on a skimpy basis.

In short, starting with a pure object-based model, we seem forced to embrace a system of specifiers of location that would have these properties: the specifiers must be applicable to any discriminable place of any part of any object; they must be comparable across all possible pairs of modalities; and the comparisons can proceed without regard to which "objects" include those locations. And this is all done preattentively. If we have all that we basically have variables that range over locations, and do so independently of what objects are found at those locations.

5. Conclusions

These fall in three broad categories. First, the evidence for cross modal cuing gives support to the view that a number of different modalities transact significant business in a location-based way. Information about locations in one can be transformed into a principle selecting stimuli in another.

Second, the remapping studies in particular give support to the view that the locations represented by the various modalities that can participate in cross modal cuing are locations in external space, and that they are represented as such. They are not locations within mind-dependent "sensory" spaces; nor are they places on the retina. They are at best retinocentric, but not retinotopic. With a change in posture, specifiers of location within an affected modality are swapped out so that new specifiers continue to specify the same region. This is the critical finding needed to show that the system is using those specifiers *as* specifiers, and what they specify are locations in external space.

Finally, a pure object-based scheme does not seem to have the wherewithal to express the content of the identifications needed in order to secure cross modal mapping. We have to represent information about locations in a fuller way than is therein allowed.

References

Berkeley, George (1709) *An Essay towards a New Theory of Vision.* Reprinted in M. R. Ayers (ed), *George Berkeley: Philosophical Works*. London: Dent, 1975. (References in the text are by paragraph number.)

Clark, Austen (2000). *A Theory of Sentience.* Oxford: Oxford University Press.

Driver, Jon & Grossenbacher, P. G. (1996). Multimodal spatial constraints on tactile selection attention. in T. Innui & J. L. McClelland (eds.), *Attention and Performance XVI. Information integration in perception and communication.* Cambridge, MA: MIT Press, 209-235.

Driver, Jon & Spence, Charles (2000). Multisensory perception: beyond modularity and convergence. *Current Biology 10*: R731-R735.

Driver, Jon & Spence, Charles (1998a). Attention and the cross-modal construction of space. *Trends in Cognitive Sciences 2*: 254-262.

Driver, Jon & Spence, Charles (1998b). Cross-modal links in spatial attention. *Philosophical Transactions of the Royal Society of London B 353*: 1319-1331.

Driver, Jon & Spence, Charles (2004). Crossmodal spatial attention: evidence from human performance. In Spence and Driver 2004, 179-220.

Kennett, S., Spence, C., & Driver, J. (2002). Visuo-tactile links in covert exogenous spatial attention remap across changes in unseen hand posture. *Perception and Psychophysics 64:* 1083-1094.

LaBerge, David. (1995). *Attentional Processing.* Cambridge, Mass.: Harvard University Press.

Ladavas, E. (1987). Is the hemispatial deficit produces by right parietal lobe damage associated with retinal or gravitational coordinates? *Brain 110:* 167-180.

Millikan, Ruth (1984). *Language, Thought, and Other Biological Categories.* Cambridge, MA: MIT Press.

Millikan, Ruth (2004). *Varieties of Meaning.* Cambridge, Mass: MIT Press.

Pylyshyn, Zenon (2003). *Seeing and Visualizing: It's Not What You Think*. Cambridge, MA: MIT Press.

Spence, Charles and Driver, Jon (1994). Covert spatial orienting in audition: exogenous and endogenous mechanisms facilitate sound localization. *Journal of Experimental Psychology: Human Perception and Performance 20:* 555-574.

Spence, Charles and Driver, Jon (1996). Audiovisual links in endogenous covert spatial attention. *Journal of Experimental Psychology: Human Perception and Performance 22:* 1005-1030

Spence, Charles and Driver, Jon (1997). Audiovisual links in exogenous covert spatial orienting. *Perception and Psychophysics 59:* 1-22.

Spence, Charles and Driver, Jon (eds.) (2004). *Crossmodal Space and Crossmodal Attention*. Oxford: Oxford University Press. (ISBN 0-19-852486-2).

Spence, Charles, MacDonald, John, & Driver, Jon (2004). Exogenous spatial-cuing studies of human crossmodal attention and multisensory integration. In Spence and Driver 2004, 277-320.

Strawson, Peter F. (1963). *Individuals*. New York: Anchor Books.

Strawson, Peter F. (1974). *Subject and Predicate in Logic and Grammar*. London: Methuen & Co. Ltd.