The cognitive architecture of bimodal event perception: A commentary and addendum to Radeau (1994)

Brian D. Fisher and Zenon W. Pylyshyn

Rutgers Center for Cognitive Science

Cognitive architecture refers to aspects of human information processing that are relatively "hard wired" . These can be contrasted with higher-level cognitive processes that are affected by the subject's beliefs and intentions—i.e. are "cognitively penetrable". Dr. Radeau makes a compelling argument for a connection between the existing visual capture literature (much of it her influential work with Paul Bertleson) and current issues in cognitive architecture. She points out that in a complex world of unimodal and multimodal events, the first step in any information processing sequence may be to partition information from different sensory channels into one or more multimodal perceptual events. Two aspects of this discussion merit special consideration:

1) Pairing of visual and auditory stimuli within the cognitive architecture using the bottom-up matching cues of and temporal spatial separation. These are discussed in terms of the Gestalt principles of common fate and proximity. Higher level matching cues, such as category fit of stimuli (presumably relative to stored exemplars of multi-modal events) seem to play a lesser role.

2) A unique spatial integration module for visual and auditory stimuli. Developmental and neurophysiological data cited in the target article suggests that the superior colliculus may accomplish this function.

Our group is also interested in the application of bimodal information integration tasks to the study of human cognitive architecture. Some of our recent investigations (Fisher, 1992a, 1992b; Fisher & Pylyshyn, 1993) are in general agreement with Dr. Radeau's data and conclusions. We

differ on some minor points, however, and will highlight these as they arise.

The hypothesis that stimulus pairing and integration of information between vision and hearing take place within an architectural module as understood according to Fodor's (1983) criteria, follows from the large body of ventriloquism and bimodal speech studies Dr. Radeau describes. Evidence from individual studies by a number of researchers support the case for cognitive impenetrability of pairing in both tasks. Our laboratory has further tested this hypothesis by carrying out phoneme perception and auditory localization tasks simultaneously, using the same stimuli. If both of these task modules are informationally encapsulated and cognitively impenetrable, we would expect that stimulus matching will not covary. Thus we would predict that errors in auditory localization induced by a visual distractor at a different location would not be correlated with errors in auditory speech perception caused by the phonemic category of that distractor.

Similarly, domain specificity would imply that each module would conduct matching based only on the type of information that falls within the task domain of that particular module. Thus, when two modules process the same bimodal event (for a location estimation and a phonological judgement of a bimodal speech act, for example) there is the distinct possibility that the two modules performing the different tasks would not have access to the same matching cues. This leads us to predict that there should be a minimal effect of spatial separation on matching for phoneme perception, and a minimal effect of phonemic discrepancy on visual capture. This should be seen here on a trial by trial basis, just as it was implied by the experiment to experiment comparisons in the target article.

In order to test these hypotheses both phonemic and spatial discrepancies were introduced to bimodal speech stimuli. Subjects were asked to perform simultaneous phoneme identification and auditory localization tasks using these targets. In the phoneme judgement task, we predicted that

modality matching itself, as well as phoneme perception, should be driven primarily by linguistic factors (cognitive fit), rather than the bottom-up factors of proximity in space and time. Similarly, the hypothetical location estimation module should be unable to access to linguistic information contained within the phoneme perception module. Finally, we predict no correlation between matching in the two tasks, since neither cognitive level nor intra-modular information about the number and sensory composition of events should be able to affect the matching process in the other module. This led us to the somewhat counter-intuitive prediction that observers may maintain multiple "views of the world" within different input modules that conflict in terms of the number and sensory composition of events.

Our experimental apparatus consisted of a video display projected on a hemicylindrical screen which concealed an array of speakers. The image of a man's face pronouncing a syllable could be displayed at any position along the subject's horizon line in the hemicylinder. For the studies we discuss here, we used visual BA or DA syllables synchronized with an auditory stimulus emanating from one of the 15 speakers in our concealed array. The auditory target was a synthetic syllable that varied in a continuum from BA to DA. Subjects were asked to report the syllable that they heard with a button press, and the location of the sound using either a pointing response or a verbal location estimation. This setup allowed us to vary independently both the physical distance between visual and auditory sources and the phonological "distance" between the visual and auditory speech components.

The results of these studies documented a clear dissociation between fusion of visual and auditory information in the two tasks. As predicted, each task module matched stimuli using information that fell within its task domain. An error analysis gave no indication that breakdown of fusion in one task correlated with breakdown in the other. These results can be taken as strong support for the integration of information within computationally specialized and informationally isolated processing modules, and thus for Radeau's position. While computational isolation in these

particular tasks is not unexpected, it seems a good demonstration of the utility of multiple simultaneous information integration tasks in establishing modularity of processing.

The target article also discusses the possibility that the superior colliculus implements auditory-visual pairing for location, and thus mediates visual capture. While this area may directly control eye movements to visual and auditory cues, recent evidence suggests that limb motor control is more likely to involve posterior parietal cortex (Goodale and Milner, 1992) . This area accepts input from the colliculus, and displays similar overlaid visual and auditory receptive fields (Ungerleider & Mishkin, 1982). It is also known to synapse with motor control neurons, and has been proposed as the site of sensory integration for motor activity (Stein, 1992). If this hypothesis is correct, the data-driven spatial integration of visual and auditory stimuli that resulted would be expected to give rise to a high level of visual capture for motor performance measures. It is interesting to note that this motor performance module (if it is indeed modular) does not fall within the traditional "horizontal modules" of perception, cognition, and response. Rather it seems as if it may combine perception and response stages, to a large extent bypassing cognitive processing.

A different situation may exist for other localization measures, however. While PP cortex may mediate motor response to bisensory events, a number of researchers in visual perception have found evidence for a division of processing within the visual modality based upon the response required of the subject (Bridgeman 1992, Goodale and Milner 1992). They conclude that a separate "ventral stream" of processing that terminates in inferotempral cortex may process visual location information for cognitive task performance. Unlike the dorsal stream, there is no compelling evidence for data-driven intersensory integration here. This difference led us to predict less visual capture for cognitive location estimation than for motor performance. An experiment was run to test this hypothesis using the apparatus described above. Subjects were asked to localize the auditory target using either a pointing or a verbal location estimation

measure, as well as to report the auditory phoneme.

As predicted, we found a higher level of visual capture for motor performance than for apparent (cognitive) location of an auditory target in the presence of a visual distractor. In neither task was there any influence of linguistic variables or correlation between matching in the phoneme perception and localization tasks. These results have led us to hypothesize that visual/auditory matching in motor performance behaves differently from matching for cognitive perception, implying that the "two visual systems" hypothesis may be better described as "two perceptual systems", and suggesting that at least two cross-modal localization modules may exist.

In summary, we feel that our results compliment those presented in the target article. We are in agreement with the importance of understanding the matching and fusion of sensory channels in event perception, and find support for the modular nature of the cognitive architecture in our own work, as well as in the studies cited in the target article. We further suggest that matching may take place in a number of modules, each with different matching rules. Continued investigation of the way in which information is cognitively partitioned in different tasks may provide important clues to help us map out human cognitive architecture and better understand the way in which observers process information from rich sensory environments.

This project was supported in part by the Air Force Office of Scientific Research grant 90-0095 to Bruce Bridgeman and an Institute for Robotics and Intelligent Systems grant to Zenon Pylyshyn.Correspondence can be sent to: Brian Fisher, Rm. A104 RuCCS/Rutcor Bldng, Rutgers U., Piscataway N.J. 08854 Ph: (908) 445-6118, e-mail: salmo@ruccs.rutgers.edu

References:

Bridgeman, B. (1992). Conscious vs. unconscious processes: The case of vision. Theory and Psychology 2(1), 73-88.

Fisher, B.D. (1992a). Integration of visual and auditory information in the perception of speech events. (Doctoral Dissertation, University of California at Santa Cruz, 1991). Dissertation Abstracts International, 52, 3324B.

Fisher, B.D. (1992b, June). Cognitive architecture and bimodal integration. Paper presented at the 2nd Annual Meeting of the Canadian Society for Brain Behaviour and Cognitive Science.

Fisher, B. D. & Pylyshyn, Z. P. (1993). Mental objects and real events: Task differences in the integration of visual and auditory speech components are predicted by FINST and ANCHOR based processing module characteristics. (Technical Report Series Cogmem #62) London, Ontario: University of Western Ontario, Centre for Cognitive Science.

Fodor, J. A. (1983). The modularity of mind : an essay on faculty psychology Cambridge, Mass.: MIT Press.

Goodale, M. A. & Milner A. D. (1992). Separate visual pathways for perception and action. Trends in Neurosciences 15(1), 20-25.

Stein, B. (1992). Posterior parietal cortex and egocentric space. Behavioral and Brain Sciences, 15, 691-700.

Ungerleider, L. G. & Mishkin, M. (1982) Two cortical visual systems. in D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield (Eds.), Analysis of Visual Behavior Cambridge: M.I.T. Press.