

**Learning Directions of Objects Specified by Vision, Spatial Audition,
or Auditory Spatial Language**

Roberta L. Klatzky¹, Yvonne Lippa², Jack M. Loomis², Reginald G. Golledge³

¹Department of Psychology
Carnegie-Mellon University
Pittsburgh, PA 15213

²Department of Psychology
University of California
Santa Barbara, CA 93106

³Department of Geography
University of California
Santa Barbara, CA 93106

Running head: Learning directions of objects

key words: spatial representation, language, learning, audition, vision, azimuth

Authors' Note

This research was supported by National Eye Institute grant EY09740. We thank Andy Beall for software and Eyal Aharoni, Mike Kinsella, and Max Krasnow for help with data collection.

Abstract

The modality by which object azimuths (directions) are presented affects learning of multiple locations. In Experiment 1, participants learned sets of three and five object azimuths specified by a visual virtual environment, spatial audition (3-D sound), or auditory spatial language. Five azimuths were learned faster when specified by spatial modalities (vision, audition) than by language. Experiment 2 equated the modalities for proprioceptive cues and eliminated spatial cues unique to vision (optic flow) and audition (differential binaural signals). There remained a learning disadvantage for spatial language. We attribute this result to the cost of indirect processing from words to spatial representations.

Information about spatial layout can be conveyed by sensory cues, as from vision or spatial audition, or abstractly by spatial language (e.g., "1 o'clock, 6 feet"). While learning and memory for visually specified positions or object locations has been investigated (e.g., Chieffi and Allport 1997; Musen 1966; Naveh-Benjamin 1987; Pezdek et al. 1986; Postma and De Haan 1996; Tresch et al. 1993), little research compares learning and memory performance in 3-D space across modalities (e.g., Battacchi et al. 1981). Recently, Loomis, Lippa, Klatzky and Golledge (2002) demonstrated that spoken language could produce a spatial representation that functioned behaviorally like one derived from 3-D sound, despite the fact that the neural pathways to spatial representation are quite different across these input modalities. They showed that when a single location was specified and listeners walked to it directly or indirectly, without vision, their degree of convergence along the direct and indirect paths was comparable for the two modalities. This result indicates a functionally equivalent representation but does not indicate its code, which might be supra-modal or modality-specific, e.g., a visuo-spatial image activated by all modalities. However, Loomis et al. (2002) found that congenitally blind participants performed updating equivalently with 3-D sound and spatial language; arguing against visual recoding.

A functionally equivalent representation of location across modalities does not guarantee comparable processing demands for encoding. We asked here whether multiple azimuths could be learned from auditory spatial language as readily as from directly spatial modalities -- audition and vision. (Distance was not varied due to distortions in auditory distance perception, e.g., Loomis et al. 1998.) Participants were presented with a set of target objects, each at a specific azimuth, and then were probed in sequence with the object names and asked to indicate the corresponding azimuths, until a learning criterion was reached. To minimize effects due to intermodal differences in temporal and spatial resolution or access to verbal codes (Pezdek et al. 1986; Potter and Faulconer 1975), the stimuli were sequentially presented verbal object labels, the presentation times were equated across modalities and allowed ample time for stimulus encoding, and the presentation and response locations were restricted to a discrete array. The richness of the visual environment was constrained by using a virtual-reality display.

Experiment 1 compared the learning of three or five object azimuths specified visually with conditions in which the azimuths were specified by 3-D sound or by auditory spatial language. We collected data from young and middle-aged adults, because learning and memory for locations decreases with increasing age (e.g., Evans et al. 1984; Kirasic 2000; Light and Zelinski 1983; Naveh-Benjamin 1987; Pezdek 1983).

Thirty-six young adults (mean age 19.7, s.d. = 1.5) and eighteen middle-aged adults (mean age 49.0, s.d. = 6.1) were randomly assigned in equal numbers to three groups: Vision, Audition, and Language. Participants wore a head-mounted display (HMD – Virtual Research Model V8) and were placed in the center of a virtual world that had a black sky and a tiled floor extending to the horizon. Rotations of the head caused commensurate optic flow from the tile borders (eye movements were not monitored). The stimuli were five object labels (baby, bird, car, cat, and dog). The azimuths were 10, 11, 12, 1, and 2 o'clock relative to the participant, who faced 12 o'clock. Labels were presented as at a constant distance of 2 m and an elevation of 1.2 m (approximately eye height) in the Vision and Audition groups. To form an item set, three or five labels and azimuths were randomly paired without replacement.

Each participant learned three sets of three items and three sets of five items; the two set sizes occurred in alternation. For each set, the task alternated between learning and test trials. On learning trials, each item of the set was presented once. In the Vision group, the object labels appeared in the virtual world as words on simulated cards. Participants slowly turned the head left or right to scan between 10 and 2 o'clock. When the head orientation coincided with a stimulus, the word appeared for 2 sec, followed by an inter-trial interval constrained to at least 4 s (further depending on scan time). In the Audition and Language groups, the participants faced 12 o'clock, and the object labels occurred with an inter-item interval of 4 s, spoken by a synthesized male voice and repeated twice with a delay of 500 ms. (In unreported data, we found that a 5-s period was sufficient to encode verbal azimuth and distance and begin walking to the designated location.) In the Language group, the stimuli were presented through headphones attached to the HMD. Participants heard the digit indicating the clock position before the repeated object label. In the

Audition group, the headphones were moved aside, and the object labels emanated from loudspeakers at the target azimuths.

During the test trial, each object label was cued visually, as letters on a card at a virtual location 3 m in front of the subject on the floor plane, until the participant responded. The visual onset coincided with a synthesized voice speaking the label. Responses were made with a custom-fabricated pointing device, consisting of a handle attached to a box with indentations corresponding to the five possible azimuths. When the handle snapped into one of the indentations, a linear potentiometer connected to it was read by the computer, which recorded the handle position. After each test trial, participants received visual feedback in the format “You got x out of y.” The learn/test sequence repeated until all items were correct in two test trials (not necessarily successive) or until eight trials were completed, whichever came first.

The first two item sets, one of each size, were eliminated from data analysis to avoid practice effects. Across the remaining four sets there were two dependent measures -- trials to criterion (maximum = 8) and percentage correct after two learning trials, shown in Figure 1. Using these variables, a multivariate analysis of variance (MANOVA; Wilks' Lambda criterion) was conducted on age, modality, and set size. Age effects were not significant, although there was a trend toward superior performance by young adults. Participants learned 3-item sets more easily than 5-item sets, $\lambda = .74$, $F(2,47) = 8.37$, $p < .001$. There was a significant effect of modality, $\lambda = .80$, $F(4, 94) = 2.81$, $p < .05$, which interacted with set size, $\lambda = .73$, $F(4, 94) = 4.07$, $p < .01$. To test the interaction, two orthogonal contrasts were performed within each set size. The first, Language vs. Audition and Vision combined, showed a disadvantage for Language in the 5-item sets only ($F_s(1,51) = 12.43$ for trials to criterion and 8.41 for % correct, $p_s < .01$). The second, Audition vs. Vision, showed an advantage for Vision in the 3-item sets only ($F_s(1,51) = 4.48$ for trials and 5.40 for % correct, $p_s < .05$).

Experiment 1 showed that auditory spatial language leads to slower learning of object locations than vision or audition, when memory load is relatively high. This indicates that perceptually based spatial learning was more effective than linguistically mediated learning. In

addition, at low memory load, vision was superior to audition. Experiment 2 provided a more stringent test of the differences across modalities, in which the cues were more nearly equated. In Experiment 1, the perceptual cues had included optic flow for vision and differential binaural cues for 3-D sound. In addition, because Vision participants turned the head to scan for targets, additional proprioceptive factors (kinesthetic and vestibular) and possibly efference copy were involved in their performance (Lewald & Ehrenstein, 2000). In Experiment 2, all inputs were accompanied by proprioceptive cues, and other perceptual spatial cues -- optic flow in the Vision group and binaural asymmetries in the Audition group -- were eliminated.

The design replicated Experiment 1. The virtual world was now black to eliminate optic flow. In the Vision condition, facing an object location triggered the viewed label. In the Audition condition, facing the location triggered the object label from the loudspeaker; thus, binaural cues always indicated that the target was straight ahead. In the Language condition, participants were instructed to turn and face the direction specified by the clock term after it occurred. Sixty young adults participated (mean age 18.6, s.d. = 1.1), 20 in each group; none had participated in Experiment 1.

Results are in Figure 1. A MANOVA was performed on modality and set size. Again, learning three items was easier than learning five items, $\lambda = .77$, $F(2, 56) = 8.23$, $p < .01$, and there was a modality x set size interaction, $\lambda = .84$, $F(4, 112) = 2.54$, $p < .05$. The same contrasts as before showed a disadvantage for Language relative to Vision and Audition with 5-item sets only (for trials, $F(1,57) = 4.37$, $p < .05$; for % correct, $F(1,57) = 3.50$, $.05 < p < .10$). Vision and Audition did not differ at either set size. An additional MANOVA on modality, set size, and experiment was conducted to compare the two experiments, incorporating only the young adult participants. There were significant effects of set size and modality x set size, but none of the terms involving experiment was significant, indicating a comparable deficit for spatial language.

To summarize the principal findings, in Experiment 1, when modality-specific spatial cues were provided for vision and audition, perceptually based learning was superior to auditory spatial language under high memory load. When, in Experiment 2, modality-specific spatial cues were

eliminated and proprioceptive cues were provided in all conditions, a deficit for language at high memory load was again observed. This result suggests that relative to intrinsically spatial modalities, the specification of egocentric direction by language deterred the formation of a representation of spatial layout for multiple azimuths. This cannot be due to a failure to translate the verbal azimuth into a spatial direction, because Experiment 2 imposed a head turn that verified each such translation. The cost must instead arise from the relatively indirect neural and/or cognitive pathways from language inputs to a spatial level of representation.

The spatial representation operative in the present experiments conveyed egocentrically defined azimuths. A plausible location for that representation is posterior parietal cortex (PPC), which has been implicated in visuo-motor guidance and complex spatial processing such as mental imagery (Kolb and Whishaw 1996). Parietal sites for location retention have been distinguished from object-retention sites by patterns of slow cortical potentials (Bosch et al. 2001), although occipital sites also appear to be involved in rehearsing locations (Awh et al. 1999). Visual inputs are directed to PPC by the dorsal processing stream (Ungerleider and Mishkin 1982). The auditory spatial processing stream also projects to PPC, which has been found to be activated by virtual auditory spatial stimuli (Rauschecker 1998a, b). PPC also appears to provide a head-centered reference frame that combines proprioceptive cues to eye and head position (Andersen 1995). These links would be used in the Audition and Vision conditions.

With respect to the Language condition, pathways from verbal azimuth labels to PPC have not been identified, but presumably they are less direct and involve general semantic processing along with spatial interpretation. Using PET, Phillips, Humphreys, Noppeney and Price (2002) found that when people judged whether a particular action occurs with a named or pictured object, both formats activated temporal and frontal areas associated with a general semantic system, but words also uniquely activated a left anterior fusiform area indicating additional semantic processing relative to pictures. Also with PET, perceptual judgments about named objects were found to activate different regions depending on the judged attribute, along with common temporal and frontal areas (Kellenbach et al. 2001).

It is possible that the inferior learning in the Language condition reflects not indirect cortical pathways per se, but the cognitive conversion of language to spatial content. The directly spatial conditions would also likely suffer if conversion were required, e.g., if the target azimuth was a constant rotation from the one presented. Of course, then the direct perceptual pathway to location representations would also be eliminated.

The present results also indicated that there is no intrinsic advantage for the learning of visually specified locations relative to 3-D sound, when vision is restricted to content only; that is, proprioception is equated and modality-specific cues (optic flow, binaural cues) are eliminated. Vision was superior to audition for learning a small set of locations, if modality-specific spatial cues were present and audition lacked accompanying proprioceptive cues. A visual presentation might be advantageous even in the reduced-cue situation, if the target azimuths varied along a continuum and multiple targets were presented simultaneously. A continuous angular error measurement might also reveal greater spatial-modality differences under the present methodology.

References

- Andersen, R.A. 1995. Coordinate transformations and motor planning in posterior parietal cortex. In *The cognitive neurosciences* (ed. M.S. Gazzaniga), pp. 519-532. MIT Press, Cambridge, MA.
- Awh, E., Jonides, J., Smith, E.E., Buxton, R.B., Frank, L.R., Love, T., Wong, E.C., and Gmeindl, L. 1999. Rehearsal in spatial working memory: Evidence from neuroimaging. *Psychological Science* 10:433-437.
- Battacchi, M.W., Franza, A., and Pani, R. 1981. Memory processing of spatial order as transmitted by auditory information in the absence of visual cues. *Memory and Cognition* 9:301-307.
- Bosch, V., Mecklinger, A., and Friederici, A.D. 2001. Slow cortical potentials during retention of object, spatial, and verbal information. *Cognitive Brain Research* 10:219-237.
- Chieffi, S., and Allport, D.A. 1997. Independent coding of target distance and direction in visuo-spatial working memory. *Psychological Research* 60:244-250.
- Evans, G.W., Brennan, P.L., Skorpanich, M.A., and Held, D. 1984. Cognitive mapping and elderly adults: Verbal and location memory for urban landmarks. *Journal of Gerontology* 39:452-457.
- Kellenbach, M.L., Brett, M., and Patterson, K. 2001. Large, colorful, or noisy? Attribute- and modality-specific activations during retrieval of perceptual attribute knowledge. *Cognitive, Affective and Behavioral Neuroscience* 1:207-221.
- Kirasic, K.C. 2000. Age differences in adults' spatial abilities, learning environmental layout, and wayfinding behavior. *Spatial Cognition and Computation* 2:117-134.
- Kolb, B., and Wishaw, I.Q. 1996. *Fundamentals of human neuropsychology, 4th Edition*. W.H. Freeman, New York, NY.
- Lewald, J., and Ehrenstein, W. H. 2000. Visual and proprioceptive shifts in perceived egocentric direction induced by eye-position. *Vision Research*, 40: 539-547.

Light, L.L., and Zelinski, E.M. 1983. Memory for spatial information in young and old adults. *Developmental Psychology* 19:901-906.

Loomis, J.M., Klatzky, R.L., Philbeck, J.W., and Golledge, R.G. 1998. Assessing auditory distance perception using perceptually directed action. *Perception and Psychophysics* 60:966-980.

Loomis, J.M., Lipka, Y., Klatzky, R.L., and Golledge, R.G. 2002. Spatial updating of locations specified by 3-D sound and spatial language. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28:335-345.

Musen, G. 1996. Effects of task demands on implicit memory for object-location association. *Canadian Journal of Experimental Psychology* 50:104-113.

Naveh-Benjamin, M. 1987. Coding of spatial location information: An automatic process? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13:595-605.

Pezdek, K. 1983. Memory for items and their spatial locations by young and elderly adults. *Developmental Psychology* 19:895-900.

Pezdek, K., Roman, Z., and Sobolik, K.G. 1986. Spatial memory for objects and words. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 12:530-537.

Phillips, J.A., Humphreys, G.W., Noppeney, U., and Price, C.J. 2002. The neural substrates of action retrieval: An examination of semantic and visual routes to action. *Visual Cognition* 9:662-684.

Postma, A., and De Haan, E.H.F. 1996. What is where? Memory for object locations. *The Quarterly Journal of Experimental Psychology* 49A:178-199.

Potter, M.C., and Faulconer, B.A. 1975. Time to understand pictures and words. *Nature* 253:437-438.

Rauschecker, J.P. 1998a. Parallel processing the auditory cortex of primates. *Audiology and Neurootology* 3:86-103.

Rauschecker, J.P. 1998b. Cortical processing of complex sounds. *Current Opinion in Neurobiology* 8:516-521.

Tresch, M.C., Sinnamon, H.M., and Seamon, J.G. 1993. Double dissociation of spatial and object visual memory: Evidence from selective interference in intact human subjects.

Neuropsychologia 31:211-219.

Ungerleider, L.G., and Mishkin, M. 1982. Two cortical visual systems. In *Analysis of visual behavior* (eds. D. J. Ingle, M.A. Goodale, and R.J.W. Mansfield), pp. 549-586. MIT Press, Cambridge, MA.

Figure Caption

Figure 1. Trials to criterion and % correct recall for middle-aged adults in Experiment 1 (left panel) and young adults in Experiment 1 (center panel) and Experiment 2 (right panel). The error bars indicate 1 standard error of the mean.

