"Put-That-There": Voice and Gesture at the Graphics Interface

Richard A. Bolt

Architecture Machine Group
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

Recent technological advances in connected-speech recognition and position sensing in space have encouraged the notion that voice and gesture inputs at the graphics interface can converge to provide a concerted, natural user modality.

The work described herein involves the user commanding simple shapes about a large-screen graphics display surface. Because voice can be augmented with simultaneous pointing, the free usage of pronouns becomes possible, with a corresponding gain in naturalness and economy of expression. Conversely, gesture aided by voice gains precision in its power to reference.

Key Words: Voice input; speech input; gesture; space sensing; spatial data management; man-machine interfaces; graphics; graphics interface.

Category Numbers: 8.2, 6.9.

The work reported herein has been supported by the Cybernetics Technology Division of the Defense Advanced Research Projects Agency, under Contract No. MDA-903-77-C-0037.

INTRODUCTION

Recently, the Architecture Machine Group at the Massachusetts Institute of Technology has been experimenting with the conjoint use of voice-input and gesture-recognition to command events on a large format raster-scan graphics display.

Of central interest is how voice and gesture can be made to inter-orchestrate, actions in one modality amplifying, modifying, disambiguating, actions in the other. The approach involves the significant use of pronouns, effectively as "temporary variables" to reference items on the display.

The interactions to be described are staged in the MIT Architecture Machine Group's "Media Room," a physical facility where the user's terminal is literally a room into which one steps, rather than a desk-top CRT before which one is perched.

The Media Room Sketched in Figure 1, is the size of a personal office: about sixteen feet long, eleven feet wide, and about eight feet from floor to ceiling. The floor is raised to accommodate cabling from an ensemble of mini-computers which drives displays and devices resident in the Media Room. The walls, finished in dark brown pile fabric, house banks of loudspeakers on either side of a wall-sized, frosted-glass projection screen, and on either side and a bit to the rear of the user's chair.

The user's chair is a vinyl-covered Eames-type chair, exactly as comes from the furniture store, except for two types of instumentation based in its arms. Either arm bears a small, one-inch high joystick, of the non-displacing variety, sensitive to pressure and direction. Nearby each joystick is a two-inch on edge, square-shaped touch sensitive pad.

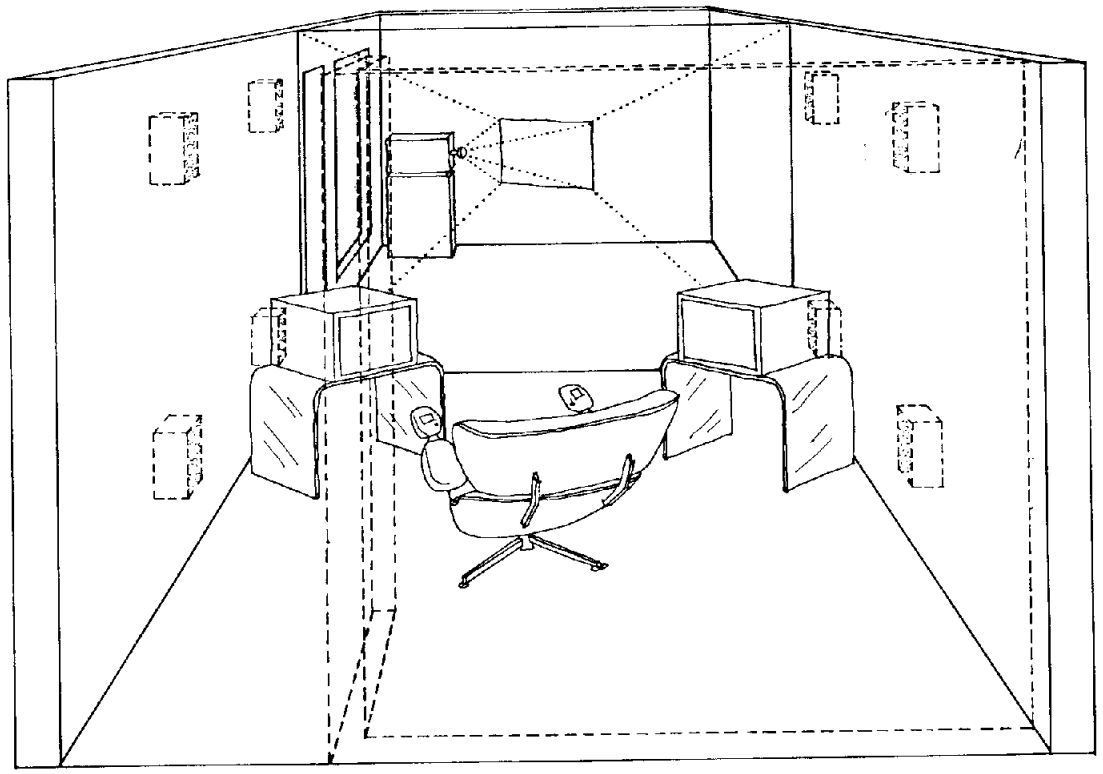

Figure 1

Sketch of Media Room

The wall-sized screen, about eight feet to the user's front, is served by back-projection from a color TV light-valve projector situated in an adjoining room. Color TV monitors are situated on either side of the user's chair, each with its tube face overlain with a transparent, touch-sensitive pad.

Apart from its role as an embodiment of the user terminal as an "informational surround [1]," the Media Room with its user chair has played a key role in our researches into a "Spatial Data-Management System, or SDMS [2]."

The specific rationale for spatially indexing data derives from our everyday experience of retrieving items, say, from our desktop: the phone to the right and above the blotter; the appointment calendar in the lower right; the "in-box" nearby the ashtray at the lower left, and so forth. Retrieval is natural and automatic for these items, with even an apparently "messy" desk having a spatial logic well-known to its creator and user, the knowledge of where this and that item are located being encoded conjointly in mental and motor models of the layout of the desktop.

The user of SDMS retrieves information not by typing names, i.e., alphanumeric strings on a keyboard terminal, but instead uses joystick and, occasionally, touch controls to navigate about in a helicopter-like manner to where specific caches of information reside in a rich graphics world of color and sound.

The world of information in SDMS, dubbed "Dataland," appears in its entirety upon one of the color TV monitors near-by the user chair. A small transparent rectangular overlay, a "you-are-here" marker, can be moved and positioned about Dataland by the user's managing of the chair's right-hand joystick (or by direct touch on the TV screen, if desired). That sub-portion of the Dataland surface indicated by the "you-are-here" rectangle is portrayed with increased detail on the large screen, effectively a magnifying window onto Dataland. The left-hand joystick on the user chair enables the user to zoom-in upon information to get a closer look at any of a number of multi-media data-types (e.g., maps, electronic "books," videodisc episodes), and perhaps to peruse them with the aid of an associated touch-sensitive "Key-map" which comes up on the other TV monitor by the user chair.

The Media Room setting, in addition to its power to generate a convincing impression of interacting with an implicit, "virtual" world of data behind the frame of the physical interface, implies yet another realm or order of space rife with possibilities for interaction: the actual space of the Media Room itself.

The sheer extent of the Media Room's physical interface creates a "real-space" environment. The user's focal situation amidst an ensemble of several screens of various sizes creates a set of geometrical relationships quite apart from any purely logical relationship between any one screen's content and that of any other.

Properly orchestrated, the two spatial orders, virtual graphical space, and the user's immediate real space in the Media Room, can converge to become effectively one continuous interactive space.

User awareness of this common space is implicit: the user points, gestures, references "up," "down," "...to the left of...," and so on, freely and naturally, precisely because the user is situated in a real space. Tapping this interactive potential is rooted in two new technical offerings in the areas of: 1) connected speech recognition; 2) position sensing in space.

SPEECH AND SPACE: THE TECHNOLOGIES

Two broad categories of currently commercially available speech recognizers may be distinguished: those which recognize discrete or isolated utterances, and those which recognize connected speech.

With those speech recognition systems restricted to discrete utterances, parsing of the speech signal into word-by-word tokens, is not done. The human speaker must talk to the system in a "clipped" or word-by-word style.

The recognition of connected speech has been a classic challenge in the field of speech recognition generally [3]. The DP-100 Connected Speech Recognition System (CSRS) by NEC (Nippon Electric Company) America, Inc. is capable of a limited amount of recognition of connected speech [4]. No pause between words is necessary, and up to five words or "utterances" are permitted per spoken sentence.

The recognition response time at the end of each sentence is about 300 milliseconds. Output is a display of the text of the utterance on an alphanumeric visual display, and/or a set of ASCII codes (numbers or letters) to be received by a processor interfaced with the NEC system.

The device's vocabulary, held in the recognizer's active memory as a set of word reference patterns, is a maximum of 120 words. With an optional "discrete utterance" mode, the size of the active vocabulary in the system's memory may be larger, about 1000 words. Except for the digits "one" to "ten," which must be spoken twice by the user when "training" the machine, each word in training mode need be spoken only once. The standard system comes with a lightweight, head-mounted microphone. We look forward to eventual use of a "shotgun" microphone in the Media Room, remote from but aimed at the speaker.

A space position and orientation sensing technology suitable for our intentions was found to be made by Polhemus Navigation Science, Inc., of Essex, Vermont. This system, called ROPAMS (Remote Object Position Attitude Measurement System) is based on measurements made of a nutating magnetic field. Essentials of the system are as follows.

Three coils are epoxied into a plastic cube, their mountings mutually orthogonal to correspond to x,y, and z spatial axes. Two such cubes are involved: one, about 1.5 inches on edge which acts as a transmitter, and another, 0.75 inches on edge, which functions as a sensor. The arrangement of coils in either cube essentially creates an antenna that is sensitive in all three orientations.

The transmitter cube radiates a nutating dipole field pointed at the sensor cube. When the pointing vector is correct, the field strength received will be constant. When it is not, there will be an error signal consisting of the nutation frequency. This error is used to generate the output pointing angles, and to re-aim the transmitter.

The orientation in space of the sensor cube is determined by transforming the differential signals from the three individual orthogonal coils in the sensor cube. The sensor cube's distance from the transmitter cube is computed by the $1/R^3$ fall-off of signal strength from the radiating dipole, or by triangulation with an additional radiator.

The sensor cube is very lightweight, and although it has a small cord running out of it, it is not an especially troublesome item to handle. Such sensors can readily be wrist-mounted, worn as finger rings, mounted on the visor of baseball caps, or put on a sort of "lab jacket" in lieu of cuff and collar buttons or epaulets.

COMMANDS

Suppose the user seated before the Media Room's large screen, with a space-sensing cube attached to a watchband on his wrist, and that the system's microphone is ready and listening. Some commands from the system's current repertoire illustrative of voice and pointing in concert are the following:

"Create . . ."

In our demonstration system, the large screen is initially either clear, or bears some simple backdrop such as a map. Against this background, simple items are called into existence, moved about, replicated, their attributes altered, and then may be ordered to vanish.

The items used are basic shapes: circles, squares, diamonds. They are non-representational in that the thing is the shape. Variable attributes are: color (red, yellow, orange, green, blue. . .), and size (large, medium, small).

For example, the user points to some spot on the large screen. A small, white "x" cursor on the screen provides running visual feedback for pointing. The user then says:

"Create a blue square there."

The size of the square is not given explicitly in this example command; the default size, "medium," is used. A blue square appears on the spot where the user is pointing. There is no default color; some color from the pre-programmed parent ensemble of color names must be given. The same is true for shape.

Where the feed-back cursor is residing on the screen at the time the spoken "there" occurs becomes the spot where the to-be-created item is placed. The occurrence of the spoken "there" is thus functionally a "when"; that is, it serves as a "voice button" for the x,y cursor action of the pointing gesture.

Accordingly, a considerable pause before the occurrence of the "there" is permissible, i.e.:

"Create a blue square . . . there."

The complete utterance in effect is a "call" to a Create routine, which routine expects certain parameters to be supplied. Before the user recites "there," the routine is parameter hung. The awaited parameter is input, completing the conjunction of x,y pointing input from the wrist-borne space sensor with the utterance "there."

Figure 2 shows the user having created a number of items on the screen before him.

"Move . . ."

The user can readily move items about the screen, and has available a variety of ways in which to express the complete "move" command.

Consider the user command:

"Move the blue triangle to the right of the green square."

This example command relies on voice mode only. Should, for example, there exist only one triangle on the screen at the time the command is given, the adjective "blue" bears no information, and could be omitted; the same logic applies for the qualifier green in "green square."

We note in passing that in the phrase ". . . the green square," the attribute "green" as voiced is treated simply as part of the name of the item as originally created. That is, the color name is used in a nominal sense, as in Moscow's "Red Square," where "Red" is functionally part of a proper name, not a signal that we should expect a city square to be painted all in red.

Apropos of color, a more ambitious "interpretive" approach might be to map the utterance "green" to pixel values, the matching mediated through the classical CIE color space, partitioned into a number of referenceable regions. The partitioning of the CIE color space on the basis of an ensemble of color names could be programmer determined on an ad hoc basis, or the partitioning might involve a quite sophisticated calibration on the basis of having subject observers name or classify displayed colors. The essential point is that the mapping from attribute-name to item-attribute can be well defined, even though it may be as complex as one cares to attempt.

In any event, the result of the above command is that the blue triangle upon "hearing" its name, de-saturates as immediate feedback that it has been "addressed," disappears from its present site to re-appear centered in a spot to the right of the green square.

The exact positioning "to the right" is programmer determined in our version; some reasonable placement is executed. The meaning, intent and interpretation of relational expressions in graphic space is a complex issue [5,6]; the

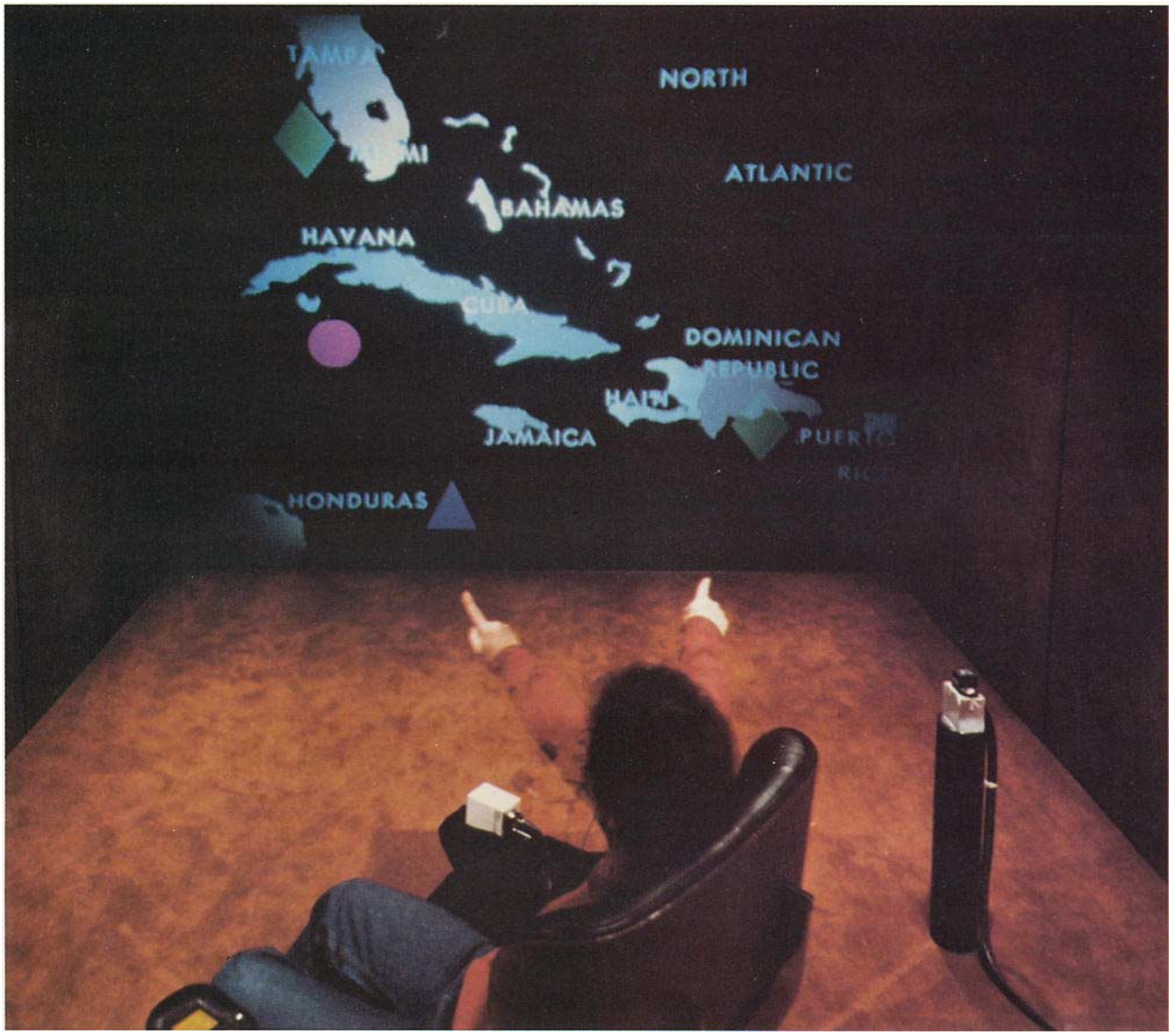

Figure 2

Talking and pointing to items on the Media Room's large screen. Here, the items are circle and diamond shapes being moved about against a backdrop of a Caribbean map. A double exposure effect catches two images of the user's right arm, strapped to which is the smaller of the pair of space-sensing cubes (covered by the user's cuff). On a pedestal to the right of the user chair is a lucite block, and to the top of this block is attached the larger transmitter cube.

important thing is that the item is now where the user has ordered it to be, and he can make minor modifications in position later.

Now, in our example action, the user might equally well have said:

"Move that to the right of the green square."

In this option, the user employs the pronoun "that," simultaneously pointing to what is intended, the pointing act being a motor analogue to the speech string: ". . . the blue triangle . . . "

Notice that in this mode of giving the command, the user may not only omit the words "blue" and "triangle," he need not even know what the thing is, or what it is called. In our simple graphics world, what anything is, is in a subtle and interesting sense, where it is.

"That" is thus defined as whatever is pointed out; effectively, it is "ostensively defined" [7]. For the namer, at least, the process is not unlike that of telling a small child what things "are": for example, pointing at a cat, and saying "cat" or "kitty." The meaning of the word is given by indicating what is the intended referent in the context of alternatives, namely, whatever else is in the scene.

This process of "pronomialization" can readily be extended in our simple graphical example. The intended target spot to which the item is to be moved can be rendered as

"Put that there"

where there, now indicated by gesture, serves in lieu of the entire phrase ". . . to the right of the green square." The power of this function is even more general. The place description ". . . to the right of the green square" presupposes an item in the vicinity for reference: namely, the green square. There may be no plausible reference frame in terms of already extant items for a word description of where the moved item is to go. The intended spot, however, may readily be indicated by voice-and-pointing: there. In this function, as well as others, some variation in expression is understandably a valuable option; thus, a mini-thesaurus of common synonyms, such as "move," "put," "place," etc., is built into the vocabulary.

"Copy . . ." as a command is simply a variant of the move action, except that the image of the item to be moved also remains in place at the original spot.

"Make that . . ."

The attributes of any item in this graphic mini-universe that the user has called into existence by voice and gesture can be modified. Here, the attributes are those of color and size.

For example, the utterance:

"Make the blue triangle smaller"

causes the referenced item to become reduced in size. The mode of reference in this instance is via voice alone, but the user could as well have said, pointing to the desired item:

"Make that smaller . . ."

The command:

"Make that a large blue diamond"

uttered while the user points at a small yellow circle causes the indicated transformation.

Extrapolations readily suggest themselves, e.g., the command line:

"Make that (indicating some item) like that (indicating some other item).

The second "that" is, functionally, a when to read the x,y coordinate of pointing. The item indicated when the second "that" is uttered becomes the "model" for change, and internally, the action is an expunging of the first referenced item, to be replaced in a "copy"-like fashion by the second referenced item.

"Delete . . ."

The "delete" command (synonyms: "erase; expunge; take out. . . ," etc.) allows the user to drop selected items from display.

As before, the "operand" of the command can be:

". . . the large blue circle"

or

". . .that" (pointing to some item).

Again, variations and extrapolations of the basic notion suggest themselves: global expunging, "clear" or "delete everything," in order to wipe the graphical slate clean; or "Detete everything to the left of this (drawing a line vertically down the face of the screen)."

NAMING

Consider a blue square that is present upon the screen. The user points to it, saying:

"Call that . . . the calendar"

with the intention of later somehow elaborating the blue square at that node into a graphical "appointment book."

The initial portion of this utterance, "Call that . . .," when processed by the recognizer unit results in codes being sent over to the host system signalling that a "naming" command has been issued. The x,y coordinates of what item is singled out by pointing are noted by the host system.

The host system then immediately directs the speech recognition unit to switch from "recognition mode" to "training mode" so that the recognizer will add the latter part of the utterance, ". . . the calendar," as a new entry in its file of word reference patterns. Upon completion of this action, the recognizer is directed to go back into recognition mode, to be ready for the next verbal input.

As the communications for switching the recognizer under host-system control between recognition and training modes currently takes a finite amount of real-time, a brief pause (indicated in the command above by three dots) must occur in the spoken command line to accommodate the time taken for the mode shift. However, the user tends to pause at precisely that point in the command line anyway, waiting for momentary desaturation of the blue square. This quick desaturation of an addressed item was noted earlier in this paper as being the system's way of giving visual feedback that the user has indeed "contacted" the item.

This spontaneous pause for feedback fortunately operates in this context to "mask" for the user the system's need for a pause in input. However, the obligation to pause represents to the system designer something of a break-down in the general convenience of continuous vs discrete speech input. An eventual strategy for relieving the necessity for a user pause in speech is the augmentation of the "intelligence" resident in a speech recognizer unit so that it to some extent interprets as well as recognizes.

For example, upon the recognition of certain "key" words or phrases within the input utterance, the recognizer itself switches directly from recognition to training mode so that sub-portions of the input utterance are handled appropriately.

In the case of the "Call that . . ." or naming command, the action of the now "intelligent" recognizer would be in effect to truncate-off from the "front-end" of the original input speech signal that span of signal corresponding to successive recognized words of the command, the non-recognized residue of the speech line to be then assumed as the new name to be assimilated by the recognizer to its internal reference pattern lexicon. In order to maintain overall coordination with the host system, the recognizer would of course simultaneously transmit ASCII codes for recognized or learned words, together with any relevant "control" codes.

While such a strategy may eliminate the need for a within-sentence speaker pause, the general problem of "coarticulation" remains: the phonemic properties of the speech signal for any word are influenced by what words are spoken with it, what particular words precede or follow the word in question (Cf. reference 3, p. 518). Thus, while not required to pause, the speaker yet must enunciate very clearly, particularly when about to utter the new name to be added.

SUMMARY

The foregoing rudimentary set of commands, concerning themselves with the simple management of a limited ensemble of non-representative objects, is intended to suggest the versatility and ease of use that can enter upon the management of graphic space with voice and gesture. More real-life examples of commanding about "things" in a more meaningful space come readily to mind: moving ships about a harbor map in planning a harbor facility; moving battalion formations about as overlays on a terrain map; facilities planning, where rooms and hallways as rectangles are tried out "here" and "there."

The power of the described technique is that indications of what is to be done with these visible, out-there-on-view items can be expressed spontaneously and naturally in ways which are compatible with the spirit and nature of the display: one is pointing to them, addressing them in spoken words, not typed symbols.

Further, the pronoun as verbal tag achieves in the graphical world the same high usefulness it has in ordinary discourse by being pronounced in the presence of a pointed to, visible graphic which functionally defines its meaning.

ACKNOWLEDGEMENTS

The programming and systems expertise of Chris Schmandt and Eric Hulteen underlay the implementation and development of the concepts described in this paper. Their efforts are duly appreciated.

REFERENCES

1. Negroponte, N The Media Room. Report for ONR and DARPA. MIT, Architecture Machine Group, Cambridge, MA, December 1978.
2. Bolt, R.A. Spatial Data-Management. DARPA Report. MIT, Architecture Machine Group, Cambridge, MA, March 1979.
3. Reddy, D.R. Speech recognition by machine: a review. Proceeding of the IEEE, 64, 4 (April 1976), 501-531.
4. Robinson, A.L. More people are talking to computers as speech recognition enters the real world. (Research News) (First of two articles) Science, 203, (16 February 1979), 634-638.
5. Sondeheimer, N.K. Spatial reference and natural-language machine control. International Journal of Man-Machine Studies, 8, (1976), 329-336.
6. Winston, P. Learning structural descriptions from examples. MIT Project MAC, TR-76, 1970.
7. Olson, D.R. Language and thought: Aspects of a cognitive theory of semantics. Psychological Review, 77, (1970), 257-273.