SPEECH AT THE INTERFACE

R.A. Bolt

Massachusetts Institute of Technology, Cambridge

ABSTRACT

In the life apart from the interface, we commonly speak not in some abstract context, but in the presence of persons and things. We touch and gesture as we speak, perhaps jot diagrams to aid our point. Further, language is eminently spatial its metaphorical content, and probably in its ontogenesis.

These observations suggest that speech interaction at the systems interface ought to benefit dramatically where the interface itself is also graphical, tangible and advisedly spacious.

Graphics enable imagery as the object of speech. Touch and position sensing afford a referencing modality independent of speech, but complementary to speech in a way that, say, the opposable thumb combines with fingers to create a hand. Spatiality provides referential context: "here", "there", "up", "down", "to the west of..."

Incorporated as interface dimensions, and orchestrated together, these modalities can form a highly plausible and especially natural matrix for interactive speech in human/systems communication.

RÉSUMÉ

Dans la vie courante, mis à part le dialogue au terminal, nous nous exprimons habituellement dans un contexte non abstrait, c'est-à-dire en présence de personnes et de choses. Lorsque nous parlons, nous gesticulons et nous touchons diverses choses; il nous arrive même de tracer des diagrammes pour mieux faire comprendre notre point de vue. De plus, le langage est éminemment spatial dans son contenu métaphorique et probablement dans son ontogénèse.

Il ressort de ce qui précède que le dialogue parlé, au niveau de l'interface d'un système, serait amélioré de beaucoup si l'interface présentait elle-même une dimension graphique, tangible et surtout spatiale.

La représentation graphique permet aux images d'être l'objet de la parole. Le toucher, la détection des positions fournissent un moyen de référence indépendant de la parole, mais qui lui est complémentaire de la même façon, par exemple, que le pouce opposable s'unit aux autres doigts pour créer la main. La spatialité offre un contexte de référence, qui permet de préciser "ici", "là", "en haut", "en bas", "à l'ouest de ...".

Ces modalités - intégrées sous formes de dimensions de l'interface et harmonisées entre elles - peuvent constituer une matrice hautement plausible et particulièrement naturelle pour le dialogue parlé dans le cadre des communications entre l'homme et l'ordinateur.

INTRODUCTION

There is something very compelling about being able to speak to a system in the presence of graphics, to be able to point at and to touch those graphics, where the system weighs your utterances in the light of those graphics, and your actions toward them.

The relationship postulated here between speech and graphics is mutual and reciprocal: speech is illuminated, modulated, interpreted in the light of the graphics the system is offering to the user, and about which the system "knows" certain things. And, in turn, the graphics are modulated in the light of what is said.

A similar relationship is postulated between speech and touch or gesture. Touch or gesture modulates the interpretation by the system of what you are saying, while what you say determines whether and how the manual input should be acted upon.

The three modalities taken together--speech, touch/gesture, a graphical presence--combine to form a total style of interaction at the systems interface. The discussion below assumes this three-way interdependence, while approaching it from the aspect of speech input.

SPEECH APART FROM THE INTERFACE

Whatever the implications in vendor literature that speech is the most "natural" input mode at the systems interface, speech apart from the interface occurs primarily in the world of common experience in the direct presence of sights, sounds, things, people. (The telephone, though ubiquitous, is herein excepted as being for our purposes an "interface."

Even when we are talking of abstract or absent topics, we talk to people before us. And speech is more than vocalization only. We gesture to animate and accent, rapping on tabletops, shrugging, chopping the air for emphasis, pointing to this or to that. Often, we reach for paper and pencil, or even for a stick to draw upon the sand, quickly to sketch to aid our intentions.

In both the life of the individual,and of the species, speech arises, and logically must have arisen in the context of daily experience, in the midst of sights and things, as part of the transactions between persons: i.e., in the presence of a palpable world, and as communicative acts between people about that world.

The exact beginnings of human speech are a deep and veiled mystery. Even to study its origins is difficult,perhaps impossible. There appears to be no human group that is not possessed of a fully developed language, and the most "primitive" of human groups speak in forms and symbology comparable to the most cultivated. There also appears to be no human "proto-language" constructable by comparative methods; the earliest reconstructable languages appear to have no less a degree of complexity than the languages of today. (Hockett, 1960).

Whatever the details of its emergence, and they may never become known with certainty, speech must have developed as humans developed over past millenia in the context of a world and of each other. And, if it is untenable to postulate humanoids who could always speak-in-language, we must then postulate early humanoid groups within which speech developed, and hence brains which had at some point become sufficiently elaborated to support the initial stages of language development.

In this regard, the mathematical psychologist Roger N. Shepard has speculated convincingly that human conceptual and linguistic competencies are rooted in an evolutionally prior spatial competency. Arguing that purely syntactic approaches to psycholinguistics are rooted in transformational grammars that are insufficiently constrained (not unlike unrestricted Turing machines), and that the mastery of syntactic rules may depend strongly upon the availability of a semantic interpretation, Shepard suggests that the transformations underlying perception, imagination, and perhaps even thinking and language "...are subject to very strong semantically determined constraints corresponding, for example, to the constraints of projection and rigid motion in three-space." (Shepard, 1979). In effect, Shepard holds that the conceptual basis of language and the use of language can reasonably be supposed to rest upon cognitive abilities previously evolved to deal with the representation of objects and their transformations in space.

In formulating his position, Shepard tells of how he was struck by the way in which people would spontaneously use spatial metaphor when discussing relations of similarity: for example, that some color was "between yellow and orange," that certain music was "closer to Bach than to Handel," that the two political candidates were "far apart on the issues," and so on. The extension of propositions from specifically spatial meaning, e.g., "within prison walls," to non-spatial meanings, e.g., "within minutes of his arrival," "within a tenor's singing range," or "within their limited competence," and the ubiquity of such extensions, suggested the prior existence of a strata of spatial terms and forms as input to an abstracting, metaphor-making, developmental language stage.

This circumstantial evidence for the spatial origins of speech from the multiplicity of spatial-metaphoric roots in language may or may not point to deeper truths about the origins of language; the argument is intriguing, not conclusive. Yet, so much of speech and speech acts, including gesture, make a plausible and comfortable fit to a spatial setting that, constructively, an opportunity arises for mutual support between speech and events in space, such as gesture and graphics, that ought not to be missed.

Let us consider some such opportunities.

TALKING YOUR WAY AROUND

In the Architecture Machine Group's laboratory at MIT we have a special room, dubbed the "Media Room." (See Figure 1). The room places the user in a comfortable office chair before a wall-sized projection screen served by back-projection from a color TV "light valve" projector. Within easy reach are touch-sensitive color TV monitors, situated on either side of the user chair. Joysticks and tiny touch-sensitive pads in either arm complement the chair, and loudspeakers embedded in the room's walls permit surrounding the user with octophonic sound. An NEC (Nippon Electric Company) DP-100 Connected Speech Recognizer enables speech input.

The Media Room is the setting of,among other projects, our Spatial Data-Management System (SDMS). The underlying principle of managing data spatially is that, rather than retrieval of information on the basis of typed-in symbols on a keyboard, you retrieve information by going to where it is in a familiar graphical space: like finding your telephone or appointment book on your desktop.

The world of SDMS, called "Dataland," is a simple, yet commodious space with graphical "here's" and "there's," an "up," "down," "left," "right," "middle," and so on. Objects in that world have relationships definable with reference to each other and with regard to the greater spatial frame of Dataland.

In SDMS, the user can window about, via touch and/or joystick, the virtual world of Dataland amidst collections of tiny items--letters, maps, books, a calculator, movies, pictures, etc.--all depicted in TV color graphics on one of the side monitors. A "you-are-here" window, a small translucent rectangle of postage stamp size, demarks that area of the Dataland surface which currently appears magnified in scale upon the large, 13-foot diagonal screen before the user. The left-hand

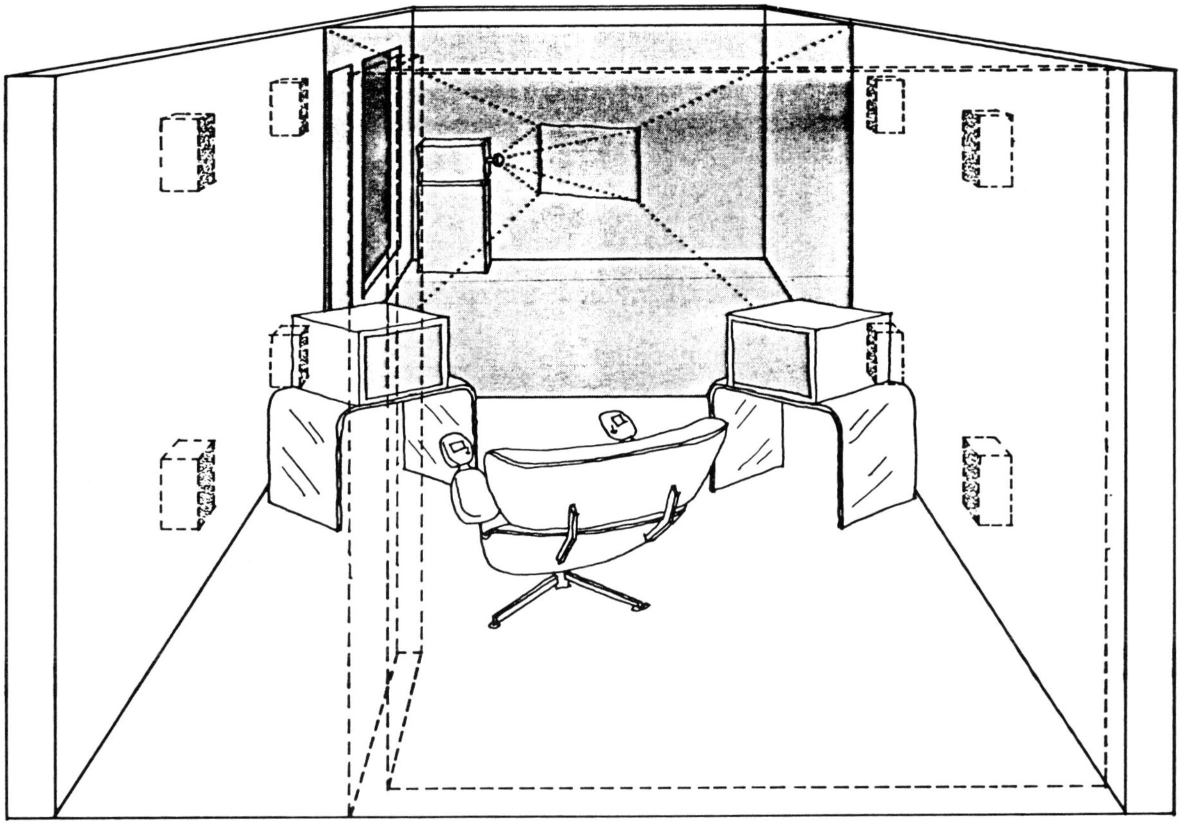

Figure 1
Sketch of Media Room

joystick permits zooming-in upon the area so demarked to peruse it or interact with some item that resides there. The features and operations of SDMS and its Dataland items have been described in detail elsewhere (Bolt, 1979). The point pertinent here is that the user can talk himself about the Dataland area: e.g., "Take me to the map above and to the right of the calculator."

The general forms of such voice-travel utterances include: addressing an object directly, as in "Take me to the Calculator;" making reference to the spatial frame of Dataland, as in "Go to the North;" making reference to other items, as in "Take me below and to the right of the Map area."

Travel directions also can be implicit: "I'd like to make a phone call..." takes you to the "telephone area," that is, brings up a telephone facility on one of the side monitors. An utterance of the form "Please call Mr. Frank Jones..." is even more implicit in that the call is placed directly by the system via auto-dialer, provided "Mr. Frank Jones" is an active vocabulary entry.

In this latter instance, access to a telephoning "facility" is not that of going to the "Telephone," a specific spot on Dataland where there resides a touch-sensitive telephone, but that of in effect causing the entire Media Room to travel in some abstract "state" space to a locus where it may perform the function of making the connection to the person you name, e.g., "Call Mr. Frank Jones." Similar in spirit is asking the system "What is the square root of 723?" instead of asking the system to "Take me to the Calculator" so that you may press upon its touch-sensitive graphical keys on the right-hand monitor to input your numbers and obtain your answer. Since such abstract "state" space locii, that of being a magic telephone operator or mathematical wizard, are, unlike real spaces, not mutually exclusive with other locii, you can readily be in one or more places (read "states") at one time. That is, there is nothing spatially discrepant about joysticking across Dataland while carrying on a telephone conversation via outside voice lines with Mr. Jones, during which you ask the system to extract a cube root for you.

Now, with respect to Dataland at large, there are certain words which are of general usage, such as "go," "take," i.e., global commands for getting about, the names of items on the Dataland surface, the "Calculator," the "Book Drop," the "Map Area," and the like, as well as general spatial reference words that stipulate relationships among items: "above," "below," "to the east of . . .", etc. These vocabulary entries are among the permanent entries of the Dataland vocabulary.

Other words are considered peculiar to specific sites within Dataland. For example, if one went over to and zoomed-in upon the "Calendar," then the down-loading of words such as "Monday," "Tuesday," etc., together with the names of the months, holidays, and so forth, would be appropriate. Conversely, words that would be relevant to getting about the Calculator, "add," "subtract," and the like could reasonably be over-written in the recognizer's active store as not now relevant.

This principal of constraining the active vocabulary on the basis of where you are has relevance not only to the problem of finite active storage for word reference patterns in speech recognizer memory, but for subsetting from total active resident vocabularies for purposes of matching optimization. If two word reference patterns, e.g., for "calendar" and for "colander" are very close rivals for matching to a just-input utterance, then it becomes helpful to know whether you are at the appointment book or in the kitchen. The system, of course, need not know the distinction between appointment book and kitchen in some "semantic" sense; it is sufficient selectively to load or activate vocabulary items on the basis of the x, y, (and possibly z) of where you are in your travels through virtual graphical space. This kind of spatial pragmatics in the service of speech interpretation in some interesting sense finesses semantics.

Pragmatics (and semantics) of course can conspire to produce the unintended. A grisly anecdote derives from the early Napoleonic Wars. After a rout of the opposition during his Egyptian campaign, Napoleon, on horseback with his officers, was confronting a contingent of prisoners. The young general was suffering from a cold, and after a fit of coughing exclaimed, "Ma sacrée toux!" ("My damned cough"). A nearby aide heard this as "Massacrez tous!"("Kill them all"), gave the order to fire, and a number of vollies were discharged into the unfortunate band of captives before the mistake was realized.

Needless to say, any command which has or could have permanent ("fatal") consequences for data had ought to require confirmation prior to being carried out. Voice commands are no exception, and pitfalls, witness the anecdote above, can arise in the most unanticipated and subtle ways.

SPEECH-DRIVEN GRAPHICS

Another laboratory project that resides in the Media Room is an exercise in speech-driven graphics which we have dubbed "Put-That-There" (Bolt, 1980).

In contrast to the general flavor and format of SDMS, one does not "window" about a graphics world that exists primarily off-screen, but interacts with the content of the screen through orchestrated speech and gesture. Eventually, when eye-tracking is installed in our Media Room, now estimated to be Summer 1981, the interactions will be three-way: speech/ eye/gesture.

Our initial ensemble of speech-commanded graphics were simple shapes, triangles, squares, and circles, displayed upon a neutral background.

Manipulable attributes of these items included: existence, location, size, color, shape. They could be called into existence by such statements as: "Put a large green circle...(gesturing) there." The gesturing was specifically a pointing action, there being a light-weight magnetic sensor worn on the speaker's wrist, which sensor supplied to the system both its attitude in space as well as its position. An "X" feedback cursor appeared on the screen indicating where the system sensed the sensor-cube to be aiming. Upon enunciation of "... there" in the command just described, the system would take the x,y coordinates of pointing contemporaneous with the utterance "there" as designating where the large green circle being called into existence was to appear.

Items, once created, could be moved about relative to one another ("Move the small blue diamond below and to the right of the yellow circle,") or relative to some cursor-indicated spot on the screen ("Put that [pointing to some item]...there [pointing to some spot]"). In this first example, notice that no pointing gesture is involved; both the item to be moved and the place to be moved to are completely specified by the words uttered. This use of language corresponds to Olsen's theory of reference "...in terms of a cognitive theory of semantics...[in which]...a semantic decision, such as the choice of a word, is made so as to differentiate an intended referent from some perceived or inferred set of alternatives." (Olson, 1970, p. 257). Specifically, when designating the item to be moved, the phrase "blue diamond" would be sufficient, given that there was only one blue diamond on view. If there were two or more, then additional words (in this case, small) need be uttered to uniquely designate the intended referent.

Of course, if there were several diamonds which were both small and blue, then additional information words would be necessary, such as "...the small blue diamond above the green square..." provided there was only one case of a small blue diamond situated above a green square.

Alternatively, the user could simply point at the item: "Put that (pointing at the small blue diamond)...below and to the right of the yellow circle." This is defining the item ostensively: "An ostensive definition is given, not simply by pointing to a referent, but by indicating the referent relative to a set of alternatives." (Olson, p. 264).

The ultimate in verbal economy is achievable in this latter instance by defining the spot to where the item is to be moved by pointing (ostensively) as well: "Put...that... there." The use of the pronouns "this," "that," "there," etc., combined with the possibility of pointing can dramatically abbreviate the burden of utterance, as well as generate a natural and flexible set of options for expression the self-same target state of affairs. Gesture has been termed a "motor analogue" to speech. (Sondheimer, 1976). Within speech itself there are various ways to "ask" the green square to re-locate itself somewhere; with gesture and speech, there is so much redundancy in the available means of expressing the same intention, that the user can concentrate more upon what work he wants to accomplish in graphic layout rather than upon how to express his commands.

We have been talking about the manual side of things as being gesture rather than touch. This is, of course, because our setting has been the Media Room's large screen beyond the immediate reach of the user. Direct tactile touch on a small nearby screen would mesh with the same ease with speech acts.

Consider a near-at-hand display screen which is touch-sensitive, but has no speech input facility. Provision might be made for touch to "pick up" some indicated item, to "place" it where one next touches. However, we would still need some independent channel to specify to the system that we are now in "move-item" mode, i.e., that our next touches are to be interpreted so that we move things rather than, say, activate some graphic "ink" upon putting finger to screen.

As channels for this type of "mode-setting," either keyboard commands or touch-menuing come to mind. More efficient, however, is to maintain one modality (touch) assigned to a cognitively homogeneous set of actions (e.g., "manipulation"), and another (speech) to a distinctly different set of actions. In our example, such words as "move," "delete," "that," "there," modulate the interpretation the system is to place upon what we do with our hands. This cross-modality modulation is, as we said earlier, reciprocal; the actions of our hands influence how the system interprets speech.

A striking example of this last point is the following. I look about on the large screen before me, wave my arm slowly and broadly about, and ask "What's that?" The system begins to describe via synthesized voice the global scene before me, e.g., "This is a map of the Caribbean Sea area. There are a number of cruise ships shown enroute to destinations, in simulation of the logistics of a ship-chartering business..." and so on.

In contrast, I look steadily at some specific item on the screen, extend my hand toward it (with very little side-to-side sway occurring), and say "What's that?" The self-same words as before, but now the voice-synthesized output of the system is: "That is the liner Island Princess, 400 passengers aboard Capt. Jones commanding, two days out from..." etc.

CONCLUSION

The meaning of speech is very much a function of where and how we say it, and the interface situation we have described (and which we have prototyped) is no exception. The provision of a graphical/spatial context for speech, together with the system capacity to capture pointing and/or touch, permits the user to speak at the interface in ways he would spontaneously do apart from the interface, and allows utterances to have the same validity at the interface as elsewhere.

The root issue is the negotiability of the person and the integrity of personal style across what has traditionally been the intractable boundary between "life" and "the interface."

ACKNOWLEDGEMENT

The work discussed herein has been supported by the Cybernetics Technology Division of the Defense Advanced Research Projects Agency, U. S. Department of Defense, under Contract Number MDA-903-77-C-0037.

REFERENCES

BOLT, R.A. "Put-That-There": Voice and Gesture at the Graphics Interface. SIGGRAPH '80 Conference Proceedings, 14(3), July 14-18, 1980, Seattle, Washington, 262-270.

BOLT, R.A. Spatial Data Management. DARPA Report. M.I.T. Architecture Machine Group, Cambridge, Massachusetts, March, 1979.

HOCKETT, C.F. The origin of speech. Scientific American, September, 1960, 203(3), 89-96.

OLSON, D.R. Language and thought: aspects of a cognitive theory of semantics. Psychological Review, 1970, 77(4), 257-273.

SHEPARD, R.N. Psychophysical complementarity. In H. Kubovy and J.R. Pomerantz (Eds.), Perceptual organization. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1979.

SHEPARD, R.N. Parallels between spatial and nonspatial uses of English prepositions. M.I.T. Workshop on Mental Representation, 21 January 1978.

SONDHEIMER, N.K. Spatial reference and natural-langiage machine control. Internal Journal of Man-Machine Studies, 1976, 8, 329-336 (Cf. p. 330).