CONF-9510197--1

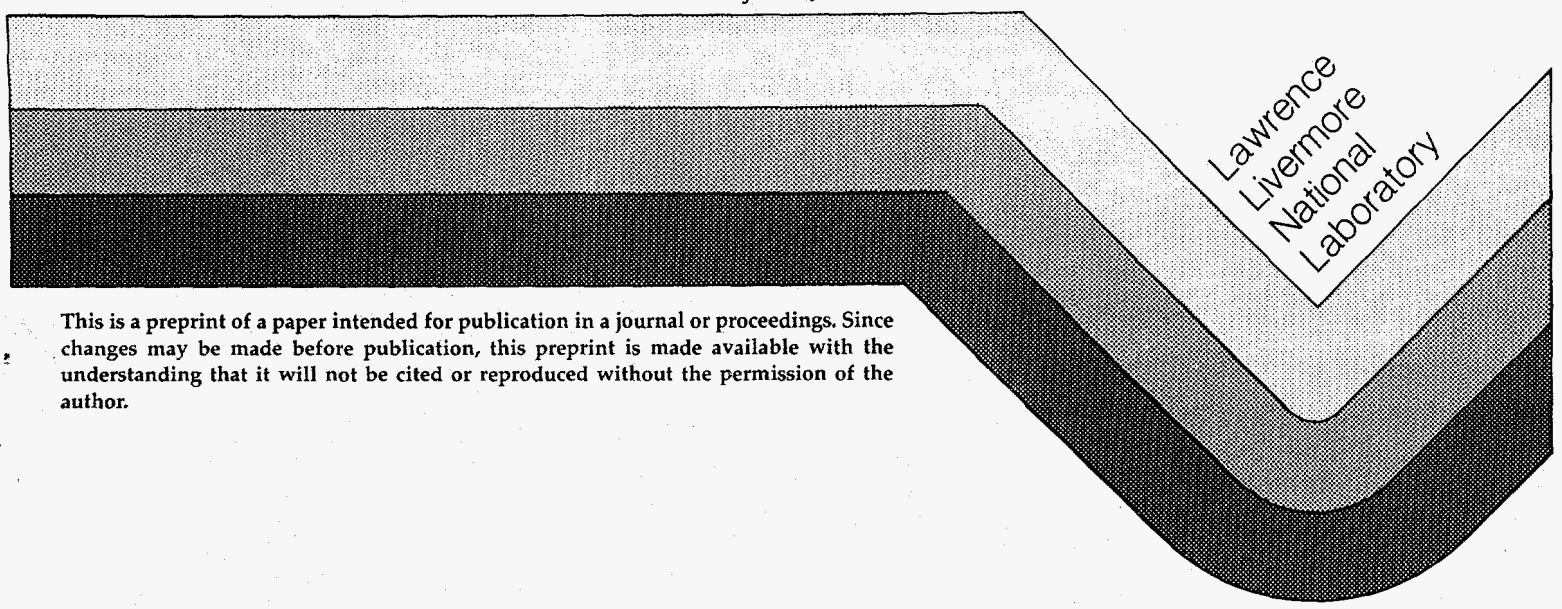
UCRL-JC-121501
PREPRINT

Multimodal Interfaces with Voice and Gesture Input

Andre D. MIlota
Meera M. Blattner

To be presented at
1995 IEEE International Conference
on Systems, Man and Cybernetics
Vancouver, BC, Canada
Oct. 23-25, 1995

July 20, 1995



DISCLAIMER

Multimodal Interfaces with Voice and Gesture Input

Andre D. Milota and Meera M. Blattner
University of California, Davis, and
Lawrence Livermore National Laboratory
Livermore, CA 94550, USA

ABSTRACT

The modalities of speech and gesture have different strengths and weaknesses, but combined they create a synergy where each modality corrects the weaknesses of the other. We believe that a multimodal system such as one interwining speech and gesture must start from a different foundation than ones which are based solely on pen input. In order to provide a basis for the design of a speech and gesture system, we have examined the research in other disciplines such as anthropology and linguistics. The result of this investigation was a taxonomy that gave us material for the incorporation of gestures whose meanings are largely transparent to the users. This study describes the taxonomy and gives examples of applications to pen input systems.

INTRODUCTION

The objective of this study is to examine the interaction between speech and gesture with the intent of designing a system that uses both modalities to their best advantage. These modalities have different strengths and weaknesses. Some of their qualities are inherent to the media through which they flow, some to the state of the art in recognition technology, some to human factors. Others are attributable to the emergent properties of the complex application interfaces in which the technologies will be employed. It is our goal to explore and evaluate the synergistic opportunities presented by a variety of modalities. As many of the technologies used for these modalities are still awkward, expensive, and lack robustness we shall confine our studies to areas where the environment is friendly, the users are knowledgeable, and the potential gains are large. We are implementing the gestures used in our examples below in a language called "Scribble."

The notion of gesture encompasses a large variety of movements. The Oxford English Dictionary says that gesture is a movement of the body, or any part of it, that is considered expressive of thought or feeling. This definition of gesture must be modified in our context. An *utterance* is a speech sequence preceded and followed by silence; it may be co-extensive with a sentence. All gesturing that occurs in association with speech and which is bound up with the total utterance is referred to as *gesticulation* [1]. For our purposes, an utterance is composed of speech sequences with gesticulation. In our current research, we are concerned with gestures that function independently of speech or sign languages. Many gestures placed in the context of an utterance use head motion, for example, nodding the head to indicate agreement or comprehension. The role of facial expressions and other head movements, such as gaze, are important for the subject of multimodal communication, but we do not consider them as part of this study. The area of gestures required for writing is not included because it is based on different approach than the ones we discuss. There are also cases where people are using gesture to accomplish some entirely different task, such as driving, while they are talking. This is not considered gesture by most researchers in the field, and it is certainly not gesticulation.

Specifically, Scribble has voice input and a flat surface with pen input, so gestures we describe are those that are applicable to that context. Gesticulation, under ordinary circumstances, is not recorded. When used for input to a computer display gesture should be considered another modality than pure gesture, because it is performing an operation. In our case, a gesture is always visible on the screen (although this doesn't have to be the case). With respect to the definition of the use of gesture as input on a flat surface, there are several different notions of what gesture is:

(a) Gesture is time-dependent and requires recognition of a sequence of actions by the user. For example, a line drawn left to right is not considered have the same referent as the same line drawn right to left.

(b) Gesture is input that is not exact, such as handwritten versus typed characters, but is not time-dependent. In this case, the line given as an example in (a) is considered independently of the direction of motion of the pen at the time it was drawn.

Anthropologists study only time-dependent gestures which are not put into a display medium. We will use both of these notions of gesture, but differentiate between them when necessary.

INTERLACING GESTICULATION AND SPEECH

Gesticulation, as the speech it accompanies, is organized into phrases [1]. A *tone unit* is a phonologically defined unit of speech, closely matching units of content or ideas. There are also *gesture phrases* where a limb moves from a rest position to engage in a movements and returns to rest. Gesture phrases and tone units are produced from the same underlying unit of meaning [1]. In other words, studies indicate that gesticulation is related to the speech it accompanies in that it is organized separately, but brought into coordination with speech because it is being employed in the same overall aim. Gesture is visual medium and utilizes both space and time, while speech is auditory and utilizes time.

In the well-designed utterance, speech and gesture can be used in complementary ways. Gesture can be used together with speech to provide additional or parallel meanings [1]. In these cases the gesture is an integral part of the utterance and the speech and gesture are formed together. Gesture may be used for economy or to disambiguate speech; gesture can be used to indicate the spatial or other visual properties that may be difficult to describe through words. Hand gestures lend themselves to specifying analog quantities and selecting among graphical elements. They are used more quickly, conveniently, and naturally than other media at certain times. Gesture may also be used in alternation with speech. In these cases it serves the role of a spoken element. The term "mixed syntax" is used to describe this type of utterance [2]. Alternation may also occur with a listener fails to understand the spoken elements of the utterance. It was shown by Efron [3] that there are cultural differences in gesticulatory styles, for example, Southern Italians "draw pictures" with their hands, while East European Jews use gestures that are abstract and bring out relationships. Cultures differ not only in the extent to which they use gesture but also the sort of information that is conveyed.

Anthropologists believe that gesture can be elaborated into a flexible and functionally general communicative code to a degree comparable to spoken language [1]. Since gestural expressions are fully integrated with spoken aspects, they are both planned at the outset, so ideas must be encoded in both gestural and verbal forms. There are linguists that believe that ideas are encoded in an abstract propositional form that is the same form used to encode verbal information [4]. Others believe that representations are modality specific [5]. One of our objectives is to design a language that can be used either with all speech or all gesture as well as interlaced speech and gesture. This study only pertains to interlaced speech and gesture.

EXISTING VOICE AND GESTURE SYSTEMS

One of the first multi-modal interfaces was created by Carbonell [6] in 1970. Bolt's "Put That There" system [7] constituted a significant milestone in the development of multi-modal user interfaces. It successfully fused two imperfect modalities to create one of greater robustness and efficiency. Since then he has published almost a dozen or more research papers in the area. He generally employs complex devices like Polhemus cubes, data gloves and eyetrackers to capture a much wider scope of gesture in three dimensions.

Salisbury built a multi-modal interface for an AWACS system [8]. It employed speech and mouse input to view radar data and issue commands to aircraft. In addition to utilizing deictic (pointing) gestures it also employed what they called a range-bearing selection gesture which did not actually set a range or bearing but allowed the user to make these measurements. Researchers at SRI have been developing a number of multimodal applications. The Shoptalk system [9] employed direct manipulation to control the scope of anaphoric references, select continuous quantities, and to point out deictic references. It also had some interesting features that addressed the problems of natural language coverage and made the process of modality fusion visible to the user. While this program employed the keyboard for its natural language input other SRI projects have looked at speech and and writing as well [10]. The SRI researchers have also done extensive work in the area of system integration and have explored the use of multi-agent architecture.

The issue of vague pointing or correction of dectic inaccuracy or what is referred to as pars-pro-toto is dealt with by the XTRA system[11]. This interface uses the utterence to resolve ambiguous pointing gestures which may encompass multiple referents to select the correct ones. The ESPRIT project also spawned a multi-modal interface, called ACORD[12], which used spoken input to resolve pointing gestures. Currently the most popular sort of multi-modal interface which employs speech are the window navigators. Here the spoken commands are generally used to navigate between windows as well as to select menu items. This technique, while it does not utilize the advantages of different media, does take advantage of the multiple channels and reduces some of the switching of the users hands between keyboard and mouse [13].

A rather compelling study of the advantages of multi-modal input was done by simply augmenting an existing graphics editer, MacDraw, with a speech input device[14]. Pausch reported a 21.23% speed increase even with this

simple architecture. In a later study he found that accelerator keys also increased the interaction efficiency by 9.92% to 14.51% depending upon the users [15]. He however concluded, "...that as the number of accelerator keys grow and the key-strokes become less obvious the savings that voice input provides will grow." Kullberg developed a sophisticated system[16] which allowed the user to dynamically remap gestures by examples provided thru the speech channel.

A TAXONOMY OF GESTURE WITH ACCOMPANYING VOICE

In order to create a gestural system for flat surfaces with pen input, we wish to make the gestural language as transparent as possible. Various kinds of problems may occur in learning a new language. In this research we hope to anticipate many of these problems by examining the way speech and gesture expressions are currently in use. The goal of creating this taxonomy has been to create a language based on speech and gesture that uses these modalities in ways with which we are already acquainted. We believe that the models for existing pen-based interfaces should not be used for multimodal systems of pen and voice. For this reason we have gone back to studies in disciplines that have examined these modalities. Any input device has certain *affordances* [17], that is, it lends itself to certain types of use, while making other types awkward or more difficult. The affordances of pens are an important consideration in the underlying design of Scribble.

PART A. GENERAL CATEGORIES

1.0 COMBINING SPEECH AND GESTURE

Speech and gesture systems can take three forms: All speech, all gesture, and intertwined speech and gesture. With intertwined speech and gesture, either speech can be parallel or gesture can replace spoken elements alternating with speech. In actual practice, both of these may be used within one utterance and they are a matter of degree. For the taxonomy, we will differentiate them.

- 1.1. Uses of speech and gesture in parallel
  - 1.1.1 for economy of utterance
  - 1.1.2 to illustrate of spatial or visual attributes
  - 1.1.3 to disambiguate an utterance
  - 1.1.4 to provide redundancy
  - 1.1.5 for emphasis
  - 1.1.6 for organization
  - 1.1.7 for confirmation
- 1.2. Uses of alternating speech and gesture
  - 1.2.1 to supply nonverbal equivalents
  - 1.2.2. a needed word may be unknown
  - 1.2.3. a time delay may have occurred

2. 0 GESTURAL TYPOLOGIES

Nespoulous and Lecours [18] classify gestures into arbitrary, mimetic, and deictic. The classifications will be important to the material in this paper, so we elaborate on each of these types.

2.1. Arbitrary Gestures

These are gestures that cannot be interpreted without being learned, hence are *opaque*. A *transparent* gesture can be deduced. Transparency and opaqueness are dependent upon culture to some degree.

- 2.1.1 Referential: refering to actions, object, circumstances, etc.
- 2.1.2 Modalizing: refering more specifically to the individual's opinion, for example, shrugging the shoulders to indicate "I don't know."

2.2. Mimetic (imitative) Gestures

These gestures are iconic in nature and transparent to the observer. These gestures often require both hands.

- 2.2.1. Analogical mimetic: relationship between the gesture and its referent, as when the gesturer draws an object in space.
- 2.2.2. Connotative mimetic: represent one of the secondary features of a referent to represent the whole.

2. 3 Deictic (pointing) Gestures

A deictic gesture cannot be used without the referent being present in the situation in which the gesture occurs. The gesture should be transparent within the context.

- 2.3.1 Specific deictic: pointing to a specific object with the purpose of referring to that object, its function, or an attribute.
- 2.3.2. Generic deictic: pointing to a whole class of objects, their functions or attributes by pointing to an object in that class.
- 2.3.3. Mimetic deictic: pointing involves an additional motion that selects among objects on a screen. For example, following a line with a pen to distinguish it from other objects.

3.0 ILLUSTRATIVE GESTURES

Some other categories of gesture that are important for our exposition is what Nespoulous and Lecours call *illustrative gestures*. These subtypes are:

3.1. Deictic illustrative: deictic is as above.

3.2. Spatiographic illustrative: an outline of the spatial configuration of the referent of one of the lexical items.

3.3. Kinemimic illustrative: outlines the action of one of the lexical items.

3.4 Pictomimic illustrative: outlines of the properties of referent, for example, big or small.

PART B. GESTURE FOR INTERACTION WITH A COMPUTER INTERFACE

4.0 SEMANTIC CATEGORIES
- 4.1 Manipulate (re-orient)
- 4.2 Change (correct, modify, replace by, undo)
- 4.3. Create or destroy
- 4.4. Establish relationship
- 4.5. Retrieve/store
- 4.6. Name
- 4.7. Confirm

4.0 GESTURES FOR ATTRIBUTES

Referring to the attributes of objects is important to pen-based systems. For this reason we have added a category specifically for attributes of objects. We differentiate these attributes from the spatiographic, kinomimic, or pictomimic because these are more specific to attributes of displayed referents.
- 4.1. Intensity
- 4.2. Direction
- 4.3. Velocity
- 4.4. Accuracy
- 4.5. Size
- 4.6. Orientation
- 4.7. Location

5.0 RELATIONSHIPS

Relationships tend to be more abstract than properties of individual objects. Relationships are often indicated by speech with examples given by gesture.
- 5.1 Relation of attributes (example: bigger or smaller)
- 5.2 Order (including hierarchies)
- 5.3 Conditionals
- 5.4 Categorize or abstract (example: vehicle and car)
- 5.5 Selection of a set of objects
- 5.6 Aggregate operations (find the max)

Studies in collaborative work have looked at gestures governing social interaction and topics related to human-to-human interaction. An example of this might be a motion to indicate a speaker wishes to interrupt another speaker. Because our application is to convey information to a computer, we will not make substantial use of the studies made in social interaction.

A computer language requires elements that have no mimetic or deictic counterparts. This has been a problem in the design of icons. Space does not permit a discussion of the design of a set of arbitrary gestures for an interface language. There is an intermediate set of gestures where existing arbitrary gestures are used. For example, a line or an "X" over a displayed element may mean deletion by conventions used in visual symbols displayed in public places (as well as for other things),

EXAMPLES OF THE TAXONOMY IN USE

Our challenge is to use the taxonomy to discover new yet obvious gestures (with speech) to convey information transparently and rapidly. Below we identified three categories: deictic, attributive, and spaciographic-pictomimic, where we believe much improvement can be made over existing multimodal systems.

Deictic Gestures

Our definition of *selection* is a deictic gesture where the objective is the identification of a referent displayed on a screen. (Another notion of selection is used in database retrieval.) A tap to change state is not deictic. A phrase such as "move this here" can have two deictic references. The accompanying speech disambiguates a single referent from a set of references, i.e., "this" versus "these."

1. Selection of a single referent: The identification of a single referent could be from a menu or a graphical or textual object displayed. An ambiguity could arise between pointing at an x-y coordinate (pixel) or an object, i.e., "Insert a word here." Is the insertion with regard to the object (text) or at the particular x-y-location on the screen? The language and context must be used to disambiguate this problem.

2. Deictic selection of a referent within a set: Identification of an object intertwined with other objects can be difficult. The object may have no identifiable characteristics that can be easily verbalized. Simple pointing may be ambiguous. Pointing may be combined with a motion such as following a contour, if the object is long and thin, such as a line or a street on a map. *Deictic-pictomimic* selection is used in Scribble. If the object is a car moving on a road among many other cars, a motion of the pen that follows the car can select it from the other cars together with speech input of "this car."

3. Deictic selection of a set of objects: In graphical programs, selection of a set is often done by encircling the desired objects. Selection of sets of objects has barely been explored in the literature. In the process of selection a user often forgets to circle one of the objects displayed. Selection of a set of objects allows the user to circle

additional objects after selection of the first has been made and connect the two sets with a line (arc). Scratching gestures can convey similar information content to circling. We used scratching or scribbling (a back and forth freehand motion) for a number of different purposes in Scribble. Scribbling over an area is one of our selection methods while scribbling in some pen-based systems is used for erasing. The voice input disambiguates gestures used for multiple purposes.

Attributive Gestures

Attributive gestures can be used with or without deictic reference.

1. Moving the displayed object: To move the entire display, a *flick* operation can be used, as with some other pen systems. The flick designates the direction of motion. The flick can also be interpreted to show speed and orientation. For example, a curved flick indicates a turn or change in orientation of the display. Circling motions with the appropriate words indicates re-orientation, while spiraling means reorientation and enlarging or shrinking depending upon the gesture.

2. Scratching: Scratching is the motion of rapidly moving a pen back and forth over a displayed object. This motion made over an object with a finger is a common gesture observed in collaborative work as pointed out in 1.3. This motion can also indicate intensity together with selection. For example, "make this circle blue," would change the filled color of an object to blue. Repeated scratching can intensify the color (as in "make it bluer") and specify a subset of the area to which this is applied, similar to drawing motions.

3. Kinomimic motion: Indicating how fast an object is to move along a path is an example of kinomimic motion. A car is "dragged" along a road to indicate the speed. Similarly, the dragging can be accompanied by "this is 50 miles per hour."

Pictomimic and Spaciographic Gestures

The spaciographic gesture is commonly used to characterize an object by its shape. Spaciographic can be used with fuzzy or inexact gestures to indicate approximate shape. Whenever possible verbal input would disambiguate the gesture.

1. Find operations (identify and retrieve objects, functions, or attributes): Drawing a triangle for the purpose of finding one or all triangles in a database.

2. Error correction: Gesture may be used to correct the shape of a curve by drawing another curve over it [16].

DISCUSSION

Speech and gesture were our forms of communication for thousands of years before writing was introduced. After the introduction of writing, pens (quills), pencils, and brushes were used because of the incredible precision and deftness they afford [19]. Punched cards, and later keyboards with a cathode ray tube display were used as input devices to computers because speech recognition and other more natural input devices were technically beyond our ability to produce. Times have changed and a wide variety of input devices are now available that take advantage of human senses. The success of the mouse was partly responsible for an examination of the use of gesture in the computer interface.

The pen is considered the most natural and ergonomic computer input device; it is small, unobtrusive, flexible, and can be manipulated easily by casual users. However, current pen interfaces are awkward and difficult to use. Oviatt [20] makes the case that pen alone may never be the universal interface that replaces the keyboard and mouse as the primary input device. Handwriting is slower than typing and recognition of all pen-written symbols is error prone and ambiguous [21]. Speech understanding has recognition problems that make speech a difficult modality for input. Also, speech is not an effective technology for the input and manipulation of graphical objects. When speech and gesture are combined through the use of pen and voice, something surprising happens--they complement each other overcoming the problems experienced by each modality. Oviatt recommends that the following strategy be adopted in connection with pen and voice interfaces: "Combine naturally complementary modalities in a manner that optimizes the individual strengths of each, while simultaneously overcoming each of their weakness."

In this investigation we are examining intertwined speech and gesture to create a foundation for Scribble, a speech and gesture system using pen and a flat surface for input. We have to other disciplines to examine what the use of speech and gesture studies conducted on was in the field. The results are given in a taxonomy. We are using gestures based on these but translated for use as computer input. Our ultimate goal is to design a multimodal system using all of our senses.

ACKNOWLEDGMENTS

Meera Blattner is also with the Department of Biomathematics, M.D. Anderson Cancer Research Hospital, University of Texas Medical Center, Houston.
This work was performed with partial support of NSF Grant #IRI-9213823 and under the auspices of the U.S. Depart. of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

REFERENCES

[1] Kendon, Adam, Current issues in The Study of Gesture, in *The Biological Foundations of Gestures,* In Nespoulous, Perron, and Lecours (Eds.), pp 23-47, Lawrence Erlbaum Assoc., Publishers, Hillsdale, New Jersey, 1986.

[2] Slama-Cazacu, T., Nonverbal components in message sequence: "Mixed syntax. In W.C. McCormack & S.A. Wurm (eds.), pp 217-227, *Language and Man: Anthropological Issues* The Hague: Mouton, 1976.

[3] Efron, D., *Gesture and Environment.* New York: Kings Crown Press. Republished as Gesture, Race and Culture. The Hague, Mouton, 1972.

[4] Pylyshyn, Z.W., What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin,* 80, pp 1-8.

[5] Shepard, R.N., The mental image, American Psychologist, 33, 1978, pp 125-137.

[6] Carbonell, J.R., Mixed-initiative man-computer instructional dialogues, Technical Report, Bolt Beranek & Newman Inc., Report No. 171, Cambridge, MA, May 1970.

[7] Bolt, R. A., Put-That-There: Voice and Gesture at the Graphics Interface, *ACM Computer Graphics, 14(3),* 1980, pp 262-270.

[8] Salisbury, M.W., Hendrickson, J.H., T. L. Lammers, and C. Fu, C., Talk and draw: bundling speech and graphics, Computer, Vol. 23, No 8, 1990, IEEE Press, pp 59-65.

[9] Cohen, P.R., Dalrymple, M., Moran, D. B., Pericra, F. C. N., Sullivan, J. W., Gargan, R. A., Jr., Schlossberg, J. L. , Tyler, S. W. (1989) Synergistic Use of Direct Manipulation and Natural Language, *Proceedings of the ACM Conference on Human Computer Interaction (SIGCHI),* pp 227-233.

[10] Oviatt, Sharon L., and Cohen, Philip R. The Contributing Influence of Speech and Interaction on Human Discourse Patterns, in *Intelligent User Interfaces* (J. W. Sullivan and S. W. Tyler, Eds), 1991, ACM Press/Addison-Wesley, Frontier Series, Reading, MA, pp 69-83.

[11] Wahlster, Wolfgang, User Discourse Models for Multimodal Communication, in *Intelligent User Interfaces* (J. W. Sullivan and S. W. Tyler, Eds), 1991, ACM Press/Addison-Wesley, Frontier Series, Reading, MA, pp 45-68.

[12] Lee, J., and Zeevat, H., Integrating Naturla Language and Graphics in Dialogue, INTERACT'90 Conference Proceedings short papers, Elsvier Science Publishers, B. V. North-Holland, 1990, pp.479-484.

[13] C. Schmandt,D. Hindus,M. S. Ackerman,S. Manandhar, Observations on using speech input for window navigation. Human-Computer Interaction. INTERACT '90. Proceedings of the IFIP TC 13 Third International Conference. Editors: D. Diaper, D. Gilmore, G. Cockton, B. Shackel, Elsvier Science Publishers[B.V. (North-Holland)] pp.787-793, Cambridge, UK, Aug. 1990,

[14] Pausch, R., and Leatherby, J.H. A Study Comparing Mouse-Only Input vs. Mouse-Plus-Voice Input fro a Graphical Editor, Proc. of the AVIOS '90 VBoice I/O systems Applications Conf., September 1990, pp 227-231.

[15] R. Pausch and J. H. Leatherby, Voice Input vs. Keyboard Accelerators: A User Study, Report No. TR-91-22. Department of Computer Science, University of Virginia,Charlottesville, VA, October 1991.

[16] R. L. Kullberg, Mark Your Calender! Learning Personalized Annotation from Intergrated Sketch and Speech, ACM Conference on Human Factors in Computing Systems, CHI '95, Denver, Colorado, May 1995, pp. 302-303.

[17] Gaver, William W., Technology Affordances, Proceedings of ACM CHI '91, Human Factors in Computing Sys., pp 79-84 .

[18] Nespoulous, Jean-Luc, and Lecours, Andre Roch, Gestures: Nature and Function, In Nespoulous, Perron, and Lecours (eds), *The Biological Foundations of Gestures,* Lawrence Erlbaum Assoc., Publishers, Hillsdale, new Jersey, 1986.

[19] Carr, Robert, and Shafer, Dan (1991) *The Power of PenPoint,* Addison-Wesley, Reading, MA.

[20] Oviatt, Sharon L. (1994) PEN/VOICE: Complementary Multimodal Communication *Speech Technology,* pp 22-25.

[21] Cohen, Philip R. (1992) The Role of Natural Language in a Multimodal Interface, *Proceedings of the ACM Conference on User Interface Systems and Technology (UIST),* November 15-18, pp 143-149.