Speech: A Privileged Modality

Luc E. JULIA
STAR Laboratory
SRI International
333, Ravenswood Ave.
Menlo Park, California 94025
julia@speech.sri.com

Adam J. CHEYER
Artificial Intelligence Center
SRI International
333, Ravenswood Ave.
Menlo Park, California 94025
cheyer@ai.sri.com

INTRODUCTION

Ever since the publication of Bolt's ground-breaking "Put-That There" paper [BOLT 80], providing multiple modalities as a means of easing the interaction between humans and computers has been a desirable attribute of user interface design. In Bolt's early approach, the style of modality combination required the user to conform to a rigid order when entering spoken and gestural commands. In the early 1990s, the idea of synergistic multimodal combination began to emerge [COHEN 89], although actual implemented systems (generally using keyboard and mouse) remained far from being synergistic. Next-generation approaches involved time-stamped events to reason about the fusion of multimodal input arriving in a given time window, but these systems were hindered by time-consuming matching algorithms. To overcome this limitation, we proposed [JULIA 93] a truly synergistic application and a distributed architecture for flexible interaction that reduces the need for explicit time stamping. Our slot-based approach is command directed, making it suitable for applications using speech as a primary modality. In this article, we use our interaction model to demonstrate that during multimodal fusion, speech should be a privileged modality, driving the interpretation of a query, and that in certain cases, speech has even more power to override and modify the combination of other modalities than previously believed.

MODEL OF INTERACTION

Numerous user studies [e.g., SIROUX 95, MELLOR 96] have shown that most subjects prefer combinations of spoken and gestural inputs. In such examples, whereas speech plays a strong role in the acquisition of commands, combining it with a pointing device provides significant (8%) improvement in performance (recognition and understanding) over the use of speech in isolation. Not surprisingly, gestures provide a fast and accurate means of locating specific objects, while voice commands are more appropriate for selecting describable sets of objects or for referring to objects not currently visible on the screen. Many of these studies also attempt to enumerate and classify the relationships between the modalities arriving for a single command (complementary, redundancy, transfer, equivalence, specialization, contradiction).

To model interactions where blended and unsorted modalities may be combined in a synergistic fashion with little need for time stamping, we first proposed a three-slot model known as VO^*V^*, such that

V or Verb is a word or a set of words expressing the action part of a command.
O^* or Object[s] is one or more objects to which the verb applies.
V^* or Variable[s] is one or more attributes or options necessary to complete the command.

Input modalities produced by the user (handwriting, speech, pen gestures) fill slots in the model, and interpretation occurs as soon as the triplets produce a complete command. A multimedia prompting mechanism is also provided to assist the user in fulfilling an incomplete command [MORAN 97]. In addition, multiple information sources may compete in parallel for the right to fill a slot, given scored modality interpretations. This model has been shown to be easily generalizable, and has been applied to various application domains, including Multimodal UNIX [LEFEBVRE 93].

SPEECH LEADS THE WAY

One might notice that whereas drawing an arrow or a circle does not change the meaning of the spoken utterance "Delete this house", drawing an arrow while simultaneously pronouncing the sentence "Show a photo of this hotel" or "Scroll in this direction" actually influences the gesture's meaning. Using this simple observation, we chose to privilege the speech modality in our model by allowing it to modify the content of an already-filled slot. This dictates that when speech is being produced, the interpretation agent must wait for the recognition result before giving a final interpretation for a command, even when

a command triplet is already completely filled. In addition, we have given speech the power to modify not only the V slot, but also the O^* and V^* slots. Let us consider the utterance "Remove all hotels in this area." During this interaction, if the user draws a circle, O^* is created containing all objects in the selected region, and the speech fills V and filters O^* to contain only hotels (complementary interactions). In a similar manner, the model can handle all classifications of multimodal interaction. If a crossout pen gesture is drawn instead of a circle (filling in the slots V="remove" and O^*={objects in gesture's boundary}), the recognized English command to reinforces the interpretation of the gesture (redundancy).

In our model, the general context of interaction also provides information to resolve ambiguities. However, even context, which normally has more weight than most input modalities, will be overridden by spoken input: what is said is what is meant. For instance, if the user draws an arrow from an object, a monomodal interpretation assigns "move" to V, and the object to O^*, instead of the other possible assignments such as scrolling (V="scroll", O^*="map", V^*=[arrow's length, arrow's direction]) or selection (O^*="object"). However, if the user is simultaneously vocalizing "scroll in this direction", the general context of the application is overwritten to give priority to the spoken request.

EXCEPTION

There are a few specialized exceptions to our assertion: in one example, if a user draws an arrow pointing to the left while saying "Scroll right" (contradiction), studies have shown that the gestural interpretation should win, as users tend to make more errors for directional commands when they speak than when they gesture with a pen.

IMPLEMENTED SYSTEMS

During the past few years, we have implemented multimodal pen and voice applications in several map-based domains (Figure 1). These systems provide an environment in which to explore the role of the speech modality during complex multimodal interactions [CHEYER 95], and hence have been designed to allow the user the greatest possible freedom of interaction during modality production. Based on observations with actual subjects, we have generated sets of rules allowing the system to make predictions when interpreting potentially ambiguous input.

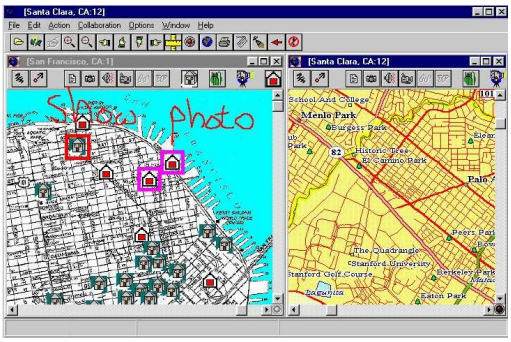

Figure 1: Multimodal Map Applications

As shown in Figure 1, the pen can also be used to enter handwritten queries. Since handwritten phrases are treated similarly to spoken input, perhaps we should change the title of this paper to "Natural Language: A Privileged Modality". Handwritten input can be appropriate for certain situations (e.g., in a noisy environment or a quiet one where speech is inappropriate); however, we have found that most users prefer spoken input when possible, especially in synergistic combination with gestures.

COMPONENT AGENTS

Our demonstration systems are built on top of the Open Agent Architecture™, a framework for constructing distributed applications by assembling and organizing communities of multiple software agents [OAA, COHEN 94]. Agent processes control and combine the input modalities, they access databases, the Internet, and natural language systems. For speech recognition, we use a continuous speech, speaker-independent, noise-robust system based on SRI's DECIPHER™ technology and commercialized by Nuance Communications [NUANCE]. For gesture recognition, we use proprietary algorithms [JULIA 95]; handwriting recognition is provided by CIC [CIC].

REFERENCES

[BOLT 80] Bolt R.A., “Put-That There: Voice and Gesture at the Graphics Interface”, ACM Computer Graphics, 1980, vol. 14-3, pp 262-270.

[CHEYER 95] Cheyer A. and Julia L., “Multimodal Maps: An Agent-based Approach”, CMC'95, Eindhoven (the Netherlands), 1995, pp 103-113.

[CIC] http://www.cic.com.

[COHEN 89] Cohen P.R. et al., “Synergistic Use of Direct Manipulation and Natural Language”, CHI'89, New York (USA), 1989, pp 227-233.

[COHEN 94] Cohen P., Cheyer A., Wang M. and Baeg S., “An Open Agent Architecture”, AAAI'94, Stanford (USA), 1994, pp 227-233.

[JULIA 93] Julia L. and Faure C., “A multimodal interface for incremental graphic document design”, HCII'93, Orlando (USA), 1993, poster sessions, p 186.

[JULIA 95] Julia L. and Faure C., “Pattern Recognition and Beautification for a Pen Based Interface”, ICDAR'95, Montreal (Canada), 1995, pp 58-63.

[LEFEBVRE 93] Lefebvre P., Duncan G. and Poirier F., “Speaking with computers: a multimodal approach”, EUROSPEECH’93, München (Germany), 1993, pp. 1665-1668.

[MELLOR 96] Mellor B.A., Baber C. and Tunley C., “In goal-oriented multimodal dialogue systems”, ICSLP’96, Philadelphia (USA), 1996, pp. 1668-1671.

[MORAN 97] Moran D., Cheyer A., Julia L., Martin D. and Park S., “The Open Agent Architecture and Its Multimodal User Interface”, IUI’97, Orlando (USA), 1997, pp. 61-68.

[NUANCE] http://www.nuancecom.com.

[OAA] http://www.ai.sri.com/~oaa.

[SIROUX 95] Siroux J., Guyomard M., Jolly Y., Multon F. and Remondeau C., “Speech and Tactile-Based Georal System”, EUROSPEECH’95, Madrid (Spain), 1995, pp. 1943-1946.