

Multimodal Interaction : A new focal area for AI

Philip R. Cohen

Center for Human-Computer Communication
Oregon Graduate Institute of Science and Technology

<http://www.cse.ogi.edu/CHCC>

Abstract*

AI research has often been driven by popular visions – HAL 2001, Asimov’s Robot, Star Trek – and by critical application areas – medical expert systems, spoken dialogue systems, etc. These visions and applications serve to inspire and guide researchers, posing challenges, illustrating technical weaknesses, and generally channeling creative energy. Without doubt, the widely held vision of the autonomous robot, has exerted a substantial integrative force, such that numerous disciplines, ranging from mechanical engineering to cognitive science, can see how their intellectual endeavors can contribute to the overall endeavor. In this brief position paper, and in the accompanying talk, I would like to propose that the next generation of intelligent multimodal user interfaces can offer a similar intellectual focus for AI researchers. After providing a brief overview of our work in this area and two examples, I would like to suggest the potential impact that such interfaces could have in the relatively near-term.

Introduction

In this position paper, I would like to illustrate recent work in intelligent user interfaces, specifically the subfield called multimodal interfaces, and argue that multimodal interfaces should occupy a similar place in the AI research endeavor as autonomous robotics. Along the way, I will describe the kinds of skills that were needed to build two multimodal systems in our laboratory, the QuickSet and Rasa systems. As you will see, AI and Computer Science methodologies played a central role, but are informed by cognitive science and ethnographic

research and methodologies. This paper is not intended as a survey of the field (cf. [29]), but rather as an example. Numerous other groups, at CMU, DFKI, Illinois, Rutgers, SRI, and elsewhere are also engaged in such research, and the reader is urged to consult their work.

The paper first considers problems of building principled multimodal systems that fuse information at the semantic level. Issues to be addressed include deciding on relevant principles, observing users, and building an appropriate software architecture. Then, I provide two examples, QuickSet, a multimodal pen/voice system for interacting with map-based applications, and Rasa, a tangible multimodal system that enables users to employ paper-based interfaces. In each case, evaluations of the systems by the intended user population have taken place and are summarized. Finally, I describe the kinds of multidisciplinary methodologies that were used in building these systems.

Multimodal Interaction

Multimodal interfaces are those that allow the user to employ a substantial range of human sensory capabilities to obtain and interact with desired information, and to perform tasks. For instance, such interfaces typically enable a user to employ some subset of speech, gaze, body movements, pen strokes, haptics, etc. Multiple modes are advantageous for a variety of reasons, including: robustness, flexibility, ability to correct persistent errors in one mode by using another, ability to avoid expected errors by switching modes, expressiveness, and speed [3, 25, 27, 29]. Such systems are also appropriate to ubiquitous computing environments, small devices, and very large devices.

Principles

Because of the complexity of building multimodal systems, we choose to investigate a principled approach to their design. The principle we have adopted is to use each modality for its strengths and to overcome weaknesses of the other(s) [5]. By studying the strengths and weaknesses of various

* The work described here was supported in part by the National Science Foundation, the Information Systems and Information Technology Offices of DARPA under contract numbers DABT63-95-C-007 and N66001-99-D-8503, and also in part by ONR grants: N00014-95-1-1164, N00014-99-1-0377, and N00014-99-1-0380. The views presented here do not represent those of the US Government.

modalities, interfaces can be designed that can dramatically simplify human input, leading to more robust performance [5, 28]. In addition to inspiring new interface designs, such a principle can be applied at run time, whereby joint use of communication modes can compensate for errors in the individual modalities. For instance, joint use of spoken language and pen-based gesture, or speech and lip recognition, can produce better overall recognition rates than relying on the individual modes alone [23, 26, 29].

Proactive Empirical Research

Before building a complete multimodal system, it is important to understand how people would in fact interact multimodally. A series of proactive empirical studies was undertaken that investigated multimodal interaction in a variety of domains. Using high-fidelity Wizard of Oz simulations as well as actual system prototypes, it was discovered that by structuring an interface graphically, users' inputs could be channeled towards a linguistically simpler style, one that led to reduced parse ambiguity, bigram perplexity, and utterance disfluencies [2, 30], up to 2-8 fold. A number of these techniques have recently been implemented in Microsoft's MiPad prototype, and the predictions have been borne out [10].

Interface simulations investigating multimodal pen/voice interactions with map-based systems have also found that multimodal input is simpler than unimodal speech. In particular, multimodal speech and pen input to map-based applications is briefer, less syntactically complex, has fewer disfluencies, leads to fewer user errors, and is preferred over unimodal speech [25]. Furthermore, people adopt one of two styles of multimodal integration -- they gesture first, then speak (typically within 4 seconds), or they speak and gesture together, but do not speak first, gesture later [25]. Such results led directly to interface designs and to thresholds used in our QuickSet system [6]. Furthermore, the empirical simulation-based results were again borne out with QuickSet user testing.

Fusion

The core of any multimodal system is its method for fusing information derived from each mode. Depending on the characteristics of the data, information can be fused "early," at the level of signal features, and/or "late" at the level of meaning. An example of the former is audio-visual fusion, in which information about lip movement (in terms of so-called "visemes") and about the spoken words (in terms of phonemes) can be combined, resulting

in better overall speech recognition, especially in noisy environments [23, 29].

In order to fuse information at the level of meaning, many groups have adopted unification of typed feature structures as the main symbolic information fusion process [8, 11, 24]. Typed feature structures are directed-acyclic attribute-value graphs, whose attributes are arranged in a type hierarchy. Such data structures are commonly used in computational linguistics to represent lexical, grammatical, and semantic information [1, 13, 22]. In the case of multimodal interaction, the meanings of the signals in each mode would be represented in such structures. Modality fusion occurs when the structures are unified [1, 11], subject to various constraints [25]. Feature structure unification, a generalization of term unification in logic programming coupled with type reasoning, is appropriate as a fusion operation because it combines complementary and redundant information, but rules out inconsistent information. In tests of QuickSet's mutual disambiguation of modalities based on feature structure unification, it was found that multimodal interaction led to a 20-40% error rate reduction over unimodal spoken language processing [26].

Fusion Architecture

Because machine perception is errorful, the software architecture needs to be designed from the ground up to handle errors. Given that recognizers produce a large number of hypotheses, with even a modest number of interpretations per mode, examining their cross-product quickly can become expensive. Thus, there needs to be rapid filtering of those combinations that could not possibly unify [14, 33], a statistical assessment of the joint probabilities of the remaining cross-modal recognition hypotheses, and finally a unification process to combine the fine structure of the interpretations. A hybrid symbolic/statistical process that can handle an arbitrary number of modes, with a variety of spatial and temporal relationships (e.g, some precede others, some co-occur, etc.) needs to be developed to arrive at the best overall interpretation of the inputs [31, 34]. Here, statistical speech and natural language processing, machine learning, pattern recognition, and sensor fusion techniques play a crucial role.

Distributed Software Architecture

The system's ability to handle parallel asynchronous input is critical to its usability. This characteristic is a significant departure from the design of current graphical user interface software, which



Figure 1: Left: QuickSet operating on a handheld PC; Right: Collaborating QuickSet operating on a 50" plasma display with touch overlay. User is speaking through a wireless microphone while drawing.

assumes that its input is certain and sequential.¹ To address these needs, we have employed a multi-agent architecture [4, 15, 18] that offers fault-tolerant, distributed, asynchronous operations, with a facilitated or direct communication model.

Example 1: QuickSet

QuickSet is a collaborative handheld multimodal system based on a multiagent architecture, which controls numerous applications, including community fire and flood control, military simulators (ModSAF), exercise initialization (ExInit), and virtual terrain environments (the Naval Research Laboratory's Dragon II and SPAWAR's CommandVu). The system enables users to create point, line, and area entities on a PC screen, using a variety of form factors ranging from handheld or wearable devices to wall-sized displays, simply by speaking and sketching. For example, the user can create and position an M1A1 company at a given location and with a given orientation and posture by saying: "M1A1 company facing one two zero degrees in defensive posture," while touching the desired location. In contrast, a user of a graphical user interface (GUI) would have to locate the desired unit in a browser or palette, drag the icon onto the screen, and fill in various parameters in a dialogue box. Likewise, by speaking and sketching, the user can create linear and area features, such as unit boundaries, objectives, routes, fortifications, air corridors, no go/slow go areas, drop zones, supply routes, cultural features, etc.

QuickSet Architecture

QuickSet consists of a set of software agents, including speech recognition, natural language processing, text-to-speech, gesture recognition, multi-

modal integration, a map-based user interface, a database system, and an application bridge. Continuous speech and continuous gesture are processed in parallel, with n-best recognition results from each mode represented as typed feature structures. After parsing, the resulting interpretations are collected by the multimodal integrator. The integrator operates as a multimodal chart parser [12], storing partial feature structure interpretations in a multimodal chart, which is operated upon by rules, and subjected to constraints. The basic fusion operation is unification of feature structures [11].

These modality agents communicate through a central facilitator via a common language, currently the Interagent Communication Language for the Open Agent Architecture [4, 18]. The presence of the facilitator enables agents to connect, disconnect, and reconnect without restarting the system. However, as a central component, the facilitator can be a bottleneck to high bandwidth information transfer, and potentially a single point of failure. To overcome these problems, we have developed a successor multiagent system, the Adaptive Agent Architecture, which offers direct as well as facilitated communication, and supports fault-tolerant operation based on the theory of teamwork [7, 15, 16]. Importantly, the agents can reside on a variety of different types of machines, located anywhere on the Internet. If appropriately time-synchronized, distributed agents can participate in analyzing users' multimodal inputs. In particular, handheld systems, such as PDAs, can run the interface, but off-load computationally intensive multimodal processing to servers operating elsewhere.

Evaluation

Multimodal interaction with QuickSet was recently compared with interaction via a standard graphical user interface (GUI) for the task of placing entities

¹ To see this, try moving the mouse on your keyboard while typing.

on a map [3]. It was found that multimodal interaction led to a 4-9 fold increase in the speed with which military users could create entities, lines and areas of various types. Although there were no more errors with multimodal interaction than with the GUI, the time to correct multimodal errors was again 4-fold faster than the time needed to correct GUI errors. Furthermore, all the users preferred interacting multimodally to using the GUI. This is just one study, comparing one GUI with multimodal interaction, but it is indicative of the potential this style of interface can offer.

Paper too?

One virtue of employing a distributed multiagent architecture as the core software architecture is that it supports a variety of platforms and hardware configurations. In particular, it is appropriate for paper-based environments.

To understand why paper is important, consider the scene below (Figure 2) from a US Army Division command post taken at a frenetic time during an exercise. On the computer screens shown here, and on the 17 other screens arrayed around the room, are military command and control systems. Here is a quiz: What's missing from this picture?



Figure 2: Scene from a US Army Division command post during an exercise (photo courtesy of William Scherlis).

Indeed, *no one* is using these or any of the other 17 systems. What the officers are in fact doing can be found in the next photo (Figure 3). They are standing on chairs, plotting the positions of military units using an 8-foot by 6-foot paper map and Post-it notes. You will notice that *they have turned their backs on computer science and technology*.

In general, it is fair to say that despite the best efforts of researchers and numerous well-funded development efforts, many military users persist in

employing paper rather than computers.² However, the officers are not simply trying to be difficult or overly conservative in rejecting digital systems. Rather, the systems they have received are missing qualities that they value highly. The users tell us that they continue to use paper maps because they have extremely high resolution, are malleable, cheap, lightweight, and can be rolled up and taken anywhere. Importantly, *paper does not fail*, and it supports face-to-face collaboration among the staff members. For example, Figure 4 illustrates the kinds of collaboration officers engage in with paper maps, which are simply unsupported by present day computer systems.



Figure 3: Officers tracking units with paper maps and Post-it notes in preference to computer systems (Photo courtesy of William Scherlis)

²They are not alone in this preference. It has been observed that medical personnel in emergency and intensive care facilities, as well as air traffic controllers, prefer to interact with physical objects, such as paper and pencil, rather than use a computer [9, 17]. What is common among all these environments is their life-and-death nature, and the users' absolute requirement for safety and robustness.



Figure 4: Multiuser collaboration around a command post map.

Thus, for a variety of factors, officers (as well as medical and air traffic control personnel) prefer paper to digital systems. *We believe there is no reason they cannot have the benefits of both.*

To provide such advantages, we have adapted the QuickSet system to use paper-based interfaces, thereby enabling officers to employ their highly practiced mode of operation when using digital systems [19-21]. The new Rasa system, developed for David McGee's Ph.D. thesis, provides both sets of benefits without substantial task overhead in virtue of its understanding the symbology drawn on the paper maps and Post-it notes, as well as its understanding of the spoken language used in creating and naming units. Essentially, both paper maps and Post-it notes rest on touch or pen-sensitive digitizers. Digital ink used in drawing symbols on a Post-it note is routed to the QuickSet symbology recognizer, which creates an n-best list of hypothesized units that the user drew. Accompanying speech is recognized and parsed, and the fused interpretation then waits at the integrator for a location feature to arrive. The act of placing the note on the map, which is overlaid on a touch-sensitive digitizer, then supplies the desired location. It is most important to note that in this tangible multimodal system, the physical artifacts, i.e., the paper map and Post-Its, *become* the computational interface. For example, moving a Post-it note on the map moves it in the digital system. In response, Rasa projects system updates onto the paper map. Rasa can also provide data to other displays and visualizations.

The system has been evaluated with military users, and has been found to be as fast as paper, and is preferred to paper [21]. Moreover, users' work

continues when the system or computer communications fail, and the effort to synchronize them when the system is brought on line is well within users' tolerance. Clearly, given the situation shown in Figure 3, we would hypothesize that it would offer superior usability to the existing command-and-control systems. This hypothesis will be tested in field experiments.

Concluding Remarks

With this very brief paper, I hope to have called attention through the QuickSet and Rasa examples to the potential for a variety of subfields within AI and computer science (e.g., speech, natural language, vision, distributed systems, machine learning, knowledge-representation and reasoning, human-computer interaction), in collaboration with numerous other disciplines (cognitive science, ethnography, linguistics, pattern recognition, sensor fusion, etc.), to contribute to radically changing the human-computer interface. In the development and evaluation of QuickSet, we see a direct progression from proactive empirical research, to system development, and finally to formal laboratory and field user testing. For Rasa, our taking an ethnographic perspective and observing the actual "work practice," enabled us to identify both problems and opportunities for technology. In a very real sense, neither of these systems could have been developed without multidisciplinary collaboration. No one methodology was employed, nor could it have been. The research within each "core" discipline was focused by the multidisciplinary goals to produce a synergistic whole.

Other domains ripe for multimodal interaction include in-home access to digital information, mobile computing, geographic information systems (GIS), computer-aided design (CAD), games. Overall, multimodal interaction can benefit society by enlarging the base of users to children, users with disabilities, or users whose physical situations of computer usage are changing. A concentrated effort to research and develop such systems can have enormous scientific and technological payoffs, and would be a worthy complement to the other focal areas of AI research.

Acknowledgements

Many thanks to my colleagues at CHCC/OGI, who made this research possible, including (but not limited to): Josh Clow, Marcus Huber, Michael Johnston, Sanjeev Kumar, David McGee, James Pittman, Sharon Oviatt, Ira Smith, Matt Wesson, and Lizhong Wu.

References

1. Carpenter, R., *The logic of typed feature structures*. 1992, Cambridge University Press: Cambridge, England.
2. Cohen, P.R., *The role of natural language in a multimodal interface*, in *Proceedings of the Conference on User Interface Software Technology (UIST'92)*. 1992, ACM Press: New York.
3. Cohen, P.R., McGee, D., and Clow, J., The efficiency of multimodal interaction for a map-based task, in the *Proceedings of the Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, Washington, May, 2000, Association for Computational Linguistics, 331-338.
4. Cohen, P.R., Cheyer, A., Wang, M.Q., et al., *An Open Agent Architecture*, in *Working notes of the AAAI Spring Symposium Series on Software Agents*. 1994, AAAI Press, reprinted in *Readings in Agents*, Huhns and Singh (eds.), Morgan Kaufmann Publishers, Inc., San Mateo, California, 1997: Stanford, Calif. 1-8.
5. Cohen, P.R., Dalrymple, M., Moran, D.B., et al., *Synergistic use of natural language and direct manipulation*, in *Proceedings of the Human-Factors in Computing Systems Conference (CHI'89)*. 1989, ACM Press: New York.
6. Cohen, P.R., Johnston, M., McGee, D., et al., *QuickSet: Multimodal interaction for distributed applications*, in *Proceedings of the Fifth ACM International Multimedia Conference*, E. Glinert, Editor. 1997, ACM Press: New York. 31-40.
7. Cohen, P.R. and Levesque, H.J., *Teamwork*. Nous, 1991. **25**(4): 487-512.
8. Denecke, M., and Yang, J., Partial information in multimodal dialogue, in the *Proceedings of the International Conference on Multimodal Interaction*, Beijing, China, 2000, 624-633.
9. Gorman, P., Ash, J., Lavelle, M., Lyman, J., Delcambre, L., and Maier, D., *Bundles in the Wild: Managing information to solve problems and maintain situation awareness*. Library Trends, 2000. **49**(2).
10. Huang, X., Acero, A., Chelba, C., Deng, L., Duchene, D., Goodman, J., Hon, H., Jacoby, D., Jiang, L., Loynd, R., Mahajan, M., Mau, P., Meredith, S., Mughal, S., Neto, S., Plumpe, M., Wang, K., Wang, Y., MIPAD: A next generation PDA prototype, in the *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000, Chinese Friendship Publishers.
11. Johnston, M., P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, I. Smith., *Unification-based multimodal integration.*, in the *Proceedings of the Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, 1997, 281-288.
12. Johnston, M., *Unification-based Multimodal Parsing*, in the *Proceedings of the Proceedings of COLING-ACL 98: The 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, 1998, Association for Computational Linguistics.
13. Kay, M., *Functional Grammar*, in *Proceedings of the Fifth Annual Meeting of the Berkeley Linguistics Society*. 1979: Berkeley, California. 142-158.
14. Kiefer, B.K., H.-U., Nederhof, M.-J., *Efficient and robust parsing of word hypotheses graphs*, in *VerbMobil: Foundations of speech-to-speech translation*, W. Wahlster, Editor. 2000, Springer: Berlin, Germany. 280-295.
15. Kumar, S., Cohen, P. R., *Towards a Fault-Tolerant Multi-Agent System Architecture.*, in the *Proceedings of the Proceedings of The Fourth International Conference on Autonomous Agents (Agents 2000)*, Barcelona, Spain, 2000, AAAI Press, 459-466.
16. Levesque, H.J., Cohen, P.R., and Nunes, J., *On Acting Together*, in *Proceedings of American Association for Artificial Intelligence (AAAI-90)*. 1990: San Mateo, California.
17. Mackay, W.E., *Is paper safer? The role of flight strips in air traffic control*. ACM Transactions on Human-Computer Interaction, 1999. **6**(4): 311-340.
18. Martin, D.L., Cheyer, A., J., Moran, D. B., *The Open Agent Architecture: A framework for building distributed software systems*. Applied Artificial Intelligence, 1999. **13**(January-March): 91-128.
19. McGee, D., Cohen, P. R., and Wu, L., *Something from nothing: Augmenting a paper-based work practice via multimodal interaction*, in the *Proceedings of the Proceedings of the Conference on Design of Augmented Reality Environments*, Denmark, 2000, Association for Computing Machinery.
20. McGee, D., Cohen, P. R., *Creating Tangible Interfaces by Augmenting Physical Objects with*

- Multimodal Language, in the *Proceedings of the Proceedings of International Conference on Intelligent User Interfaces*, Santa Fe, New Mexico, 2001.
21. McGee, D.R., *Augmenting environments with multimodal interaction*, in *Dept. of Computer Science and Engineering*. forthcoming, Oregon Graduate Institute of Science and Technology: Beaverton, OR.
 22. McKeown, K.R., *Generating natural language text in response to questions about database structure*. 1982.
 23. Neti, C., Iyengar, G., Potamianos, G., Senior, A., Perceptual interface for information interaction: Joint processing of audio and visual information for human-computer interaction, in the *Proceedings of the Proceedings of the International Conference on Spoken Language Processing (ICSLP'2000)*, Beijing, China, 2000, Chinese Friendship Publishers, 11-14.
 24. Nigay, L., and Coutaz, J., A generic platform for addressing the multimodal challenge, in the *Proceedings of the Proceedings of the Conference on Human Factors in Computing Systems (CHI'95)*, 1995, ACM Press, 98-105.
 25. Oviatt, S.L., *Multimodal interactive maps: Designing for human performance*. Human Computer Interaction, 1997. **12**: 93-129.
 26. Oviatt, S.L., Mutual disambiguation of recognition errors in a multimodal architecture, in the *Proceedings of the Human Factors in Computing Systems (CHI'99)*, New York, 1999, ACM Press, 576-583.
 27. Oviatt, S.L., Cohen, P. R., *Multimodal interfaces that process what comes naturally*. Communications of the ACM, 2000. **Volume 43**(3): 45-49.
 28. Oviatt, S.L., *Taming recognition errors with a multimodal interface*. Communications of the ACM, 2000. **43**(9): 45-51.
 29. Oviatt, S.L., Cohen, P. R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., Ferro, D., *Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions for 2000 and beyond*. Human Computer Interaction, 2001. **15**(4): 263-322.
 30. Oviatt, S.L., Cohen, P.R., and Wang, M.Q., *Toward Interface Design for Human Language Technology: Modality and Structure as Determinants of Linguistic Complexity*. Speech Communication, 1994. **15**(3-4).
 32. Stork, D.G., Wolff, G., and Levine, E., *Neural network lipreading system for improved speech recognition*, in *Proceedings of the International Joint Conference on Neural Networks Vol. II*. 1992. 286-295.
 33. Suhm, B., Myers, B., and Waibel, A., Model-based and empirical evaluation of multimodal interactive error correction, in the *Proceedings of the Human Factors in Computing Systems (CHI'99)*, New York, 1999, ACM Press, 584-591.
 34. Wu, L., , Oviatt, S., L. and Cohen, P. R., *Statistical multimodal integration for intelligent HCI*, in *Neural Networks for Signal Processing*, Y.H. Hu, Larsen, J., Wilson, E., and Douglas, S., Editor. 1999, IEEE Press: New York. 487-496.