HUMAN-COMPUTER INTERACTION, 1997, Volume 12, pp. 1-5

Introduction to This Special Issue on Multimodal Interfaces

Sharon Oviatt
Oregon Graduate Institute of Science and Technology

Wolfgang Wahlster
German Research Center for Artificial Intelligence (DFKI GmbH)
Saarbrücken

The growing emphasis on multimodal interface design is fundamentally inspired by the aim to support natural, flexible, efficient, and powerfully expressive means of human-computer interaction that are easy to learn and use. Multimodal interfaces represent a new direction for computing that draws from the myriad input and output technologies becoming available, and that potentially can integrate complementary modalities to yield more synergistic blends than has been possible previously. It is a research direction that seeks guidance from cognitive science expertise on the coordinated human perception and production of naturally co-occurring modalities (e.g., speech, gesture, gaze, and facial movements) during interaction with people and computers. In fact, the realization of successful multimodal systems is dependent on - and only can flourish through - extensive interdisciplinary cooperation, as well as teamwork among those representing expertise in the individual component technologies.

In this special issue, new work on next-generation multimodal interfaces includes articles based on PhD theses by Elizabeth Mynatt of Xerox Palo Alto Research Center, Robert Stevens of the University of York, and Yi Han of the University of Melbourne. It also includes articles based on extensive work by Steve Roth's group at Carnegie Mellon University and from Sharon Oviatt's laboratory at the Oregon Graduate Institute of Science and Technology. Reading these articles, it becomes clear that *multimodal interface* in no sense refers to a unified and specialized subarea of research. Rather, it is a collection of emerging research areas, like early evening stars gradually glowing more brightly. Some of the clusters already visible include multimodal interfaces that support computing for special populations (e.g., Mynatt, this issue; Stevens, Edwards, & Harling, this issue), research on coordinated multimodal input (e.g., Oviatt, this issue), and work on presentation planning and coordinated multimodal output (e.g., Han & Zukerman, this issue; Roth, Chuah, Kerpedjiev, Kolojejchick, & Lucas, this issue). In addition, new methods for generating research and system development are becoming available, including multimodal simulation techniques for collecting data (e.g., Oviatt, this issue) and agent architectures for implementing integrated multimodal systems (e.g., Han & Zukerman, this issue).

Multimodal interfaces have the potential to expand computing to encompass more challenging applications, for use by a broader spectrum of the population, and during more adverse usage conditions. It is not coincidental, then, that most of the articles in this issue present work in support of challenging applications; these include algebra instruction (Stevens et al., this issue), presentation of data summaries (Han & Zukerman, this issue; Roth et al., this issue), and interaction with complex spatial displays such as maps (Oviatt, this issue; Roth et al., this issue). In addition, two articles

We thank the many reviewers who provided insightful comments and valuable feedback to authors. Special thanks also to Tom Moran for his enthusiastic support of our developing the research content in this special issue, for helpful comments on the introduction, and for editorial advice throughout the project.

describe implemented systems that make computing accessible to the visually impaired (Mynatt, this issue; Stevens et al., this issue). With respect to expansion of usage contexts, one empirical article also addresses multimodal interface design for portable devices to be used in natural field settings (Oviatt, this issue). In these respects, multimodal systems are being designed to support computing for applications, user groups, and usage contexts that either have not been available or have been accommodated poorly in the past - which is expected to precipitate a major shift in our experience of the accessibility, utility, and quality of modern computing.

The Mynatt article describes the Mercator system, which was designed to transform salient components of today's graphical interface into a rich auditory form appropriate for the visually impaired. Mercator uses intuitive auditory icons to convey interface objects and their functionality. As a basis for approaching the difficult problems associated with navigation, the Mercator system adopts a hierarchical model, which blind users learn to traverse in place of a spatial one. To assist in alleviating the otherwise tedious traversals of a large hierarchical structure, techniques that permit interruption, navigation shortcuts, and previews are incorporated into Mercator's interface design. Mercator's goal is essentially to translate the visual mode of output into a parallel auditory one for supporting tasks such as text processing and mail. Two areas for further research and development are the extension of Mercator to accommodate spoken input and the incorporation of a tactile display as a complement to nonspatial and transitory auditory output.

In contrast, the Mathtalk system by Stevens et al. devises new ways to render spatially-oriented algebraic expressions so that visually impaired students can receive mathematics education. Like Mercator, Mathtalk also focuses on exploiting the auditory modality, and it utilizes both speech and nonspeech output. However, as part of its cognitive-educational focus, the thrust of Mathtalk is to provide blind students with an interface that supports active reading through control over the flow of information in algebra formulas. For example, it permits auditory glances so blind students can gain an overview of the size, balance, and structure of a formula using features like algebra earcons, while hiding the detail in complex expressions until the student is ready to analyze it. In addition, the synthetic spoken output of algebra terms is enhanced with natural prosodic cues that convey structure without taxing the short-term memory of listeners. This contrasts with commercial screen readers that speak lengthy strings of LaTex control characters mixed with algebra notation. The Mathtalk system currently is being extended by the Maths project to develop a suite of multimodal tools (including braille input, soft braille displays, speech and nonspeech output, etc.) for reading and manipulating mathematics information.

Also working in a spatial domain, Oviatt presents empirical research demonstrating the performance advantages of coordinated multimodal interaction with maps. To support portability as well as more expressive and flexible user input, Oviatt analyzes spoken, pen-based, and multimodal pen-voice input as ways to interact with dynamic map systems. The results of this research identify performance advantages when interacting with maps multimodally rather than unimodally - including faster task completion, fewer errors, fewer input disfluencies, briefer and less complex language input, and greater user satisfaction. Oviatt's article also presents data on how the structure of a visual display influences the simplicity and processability of users' language input. It outlines several interface techniques for designing map interfaces to guide user input to match processing capabilities, so that robust system processing can be optimized. From a methodological viewpoint, this work is based on a rapid semiautomatic simulation method that offers a tool for investigating a range of new multimodal systems involving different input and output capabilities.

A comprehensive information visualization workspace is presented by Roth et al., in which users are able to explore data presented in diverse but coordinated formats using a collection of related systems called Visage, Sagebrush, Sagebook, and selective dynamic manipulation (SDM). With these tools, users also can structure and dynamically manipulate their own data summaries displayed as tables, charts, maps, and the like (see color plates in this issue), and import these displays into briefings to be communicated to others. From a long-term perspective, this work on automated presentation planning aims to build graphically articulate visualization systems. Visage, Sage, and SDM are designed to offer an information-centric means of coordinating displays so that users can focus on tracking critical information as they attempt to navigate and structure large volumes of complex data. To further enhance the usability of these visualization tools, Roth et al. discuss the addition of spoken input as a complement to direct manipulation. Spoken input permits users to work

more quickly and efficiently to achieve certain goals, for example when issuing iterative commands, locating out-of-view objects, and so forth. Like Oviatt's work, Roth et al.'s interest in combined multimodal input is motivated by the need to accommodate a challenging visual-spatial application domain, and by an increasing awareness of the need to integrate multimodal input and output capabilities in a strategic manner.

Finally, the Han and Zukerman article represents work toward automating multimodal presentation planning, based on a multiagent architecture using a blackboard. The system they describe, called Magpie, implements a set of agents that communicate among one another to satisfy the constraints needed for layout of text and graphics in data summaries such as tables. Magpie enables the dynamic creation of modality-specific agents needed to select and integrate basic components of the data presentation. In particular, Magpie's constraint propagation mechanism enables agents to cooperate in planning constraints that restrict the modality and screen space available for laying out each display component. In light of time and space limitations, the system's multiagent architecture resolves resource competition among agents representing alternative modalities. It is clear that agent architectures of this kind will play a substantial role in the implementation of future integrated multimodal systems.

To realize successful multimodal systems of the future, many key research challenges and scientific issues remain to be addressed. Among these challenges are the continued development of new component technologies, including those needed to render dynamic visual displays, to present information in tactile form, and to process human speech, gaze, and manual gestures. In addition, strategies will be needed for coordinating input and output modalities, for resolving integration and synchronization issues among modes, and for using information in one input mode to disambiguate noisy or error-prone input in another. A general theory of communicative interaction also will be needed to provide a foundation for handling interactive dialogue in a manner independent of the specific input and output modes used in any given system. Other key challenges include consideration of the entire interactive input-output cycle between human and computer, including the impact of multimedia system displays and feedback on subsequent multimodal user input - rather than viewing input and output as separate research domains. Finally, future research will be needed to evaluate the adequacy of multimodal systems for supporting human performance under the expanded conditions of usage that are anticipated.

NOTES

Authors' Present Addresses. Sharon Oviatt, Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, P.O. Box 91000, Portland, OR 97291-1000. E-mail: oviatt@cse.ogi.edu. Wolfgang Wahlster, German Research Center for Artificial Intelligence (DFKI GmbH), Stuhlsatzenhausweg 3, D-66123, Saarbrücken, Germany. E-mail: wahlster@dfki.uni-sb.de.

ARTICLES IN THIS SPECIAL ISSUE

Han, Y., & Zukerman, I. A. (1997). A mechanism for multimodal presentation planning based on agent cooperation and negotiation. *Human-Computer Interaction, 12*, 187-226.

Mynatt, E. D. (1997). Transforming graphical interfaces into auditory interfaces for blind users. *Human-Computer Interaction, 12*, 7-45.

Oviatt, S. L. (1997). Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction, 12*, 93-129.

Roth, S. F., Chuah, M. C., Kerpedjiev, S., Kolojejchick, J., & Lucas, P. (1997). Toward an information visualization workspace: Combining multiple means of expression. *Human-Computer Interaction, 12*, 131-185.

Stevens, R. D., Edwards, A. D. N., & Harling, P. A. (1997). Access to mathematics for visually disabled students through multimodal interaction. *Human-Computer Interaction, 12*, 47-92.