Check for updates

Moving from traditional interfaces toward interfaces offering users greater expressive power, naturalness, and portability.

Ten Myths of Multimodal Interaction

Sharon Oviatt

Multimodal systems process combined natural input modes—such as speech, pen, touch, hand gestures, eye gaze, and head and body movements—in a coordinated manner with multimedia system output. These systems represent a new direction for computing that draws from novel input and output technologies currently becoming available. Since the appearance of Bolt's [1] "Put That There" demonstration system, which processed speech in parallel with manual pointing, a variety of multimodal systems has emerged. Some rudimentary ones process speech combined with mouse pointing, such as the early CUBRICON system [8]. Others recognize speech while determining the location of pointing from users' manual gestures or gaze [5].

Recent multimodal systems now recognize a broader range of signal integrations, which are no longer limited to the simple point-and-speak combinations handled by earlier systems. For example, the Quickset system integrates speech with pen input that includes drawn graphics, symbols, gestures, and pointing. It uses a semantic unification process to combine the meaningful multimodal information carried by two input signals, both of which are rich and multidimensional. Quickset also uses a multiagent architecture and runs on a handheld PC [3]. Figure 1 illustrates Quickset's response to the multimodal command "Airstrips... facing this way, facing this way, and facing this way," which was spoken while the user drew arrows placing three airstrips in correct orientation on a map.

Multimodal systems represent a research-level paradigm shift away from conventional windows-icons-menus-pointers (WIMP) interfaces toward providing users with greater expressive power, naturalness, flexibility, and portability. Well-designed multimodal systems integrate complementary modalities to yield a highly synergistic blend in which the strengths of each mode are capitalized upon and used to overcome weaknesses in the other. Such systems potentially can function more robustly than unimodal systems that involve a single recognition-based technology such as speech, pen, or vision.

Systems that process multimodal input also aim to give users better tools for controlling the sophisticated visualization and multimedia output capabilities that already are embedded in many systems. In contrast, keyboard and mouse input are relatively limited and impoverished, especially when interacting with virtual environments, animated characters, and the like. In the future, more balanced systems will be needed in which powerful input and output capabilities are better matched with one another.

As a new generation of multimodal systems begins to define itself, one dominant theme will be the integration and synchronization requirements for combining different modes strategically into whole systems. The computer science community is just beginning to understand how to design well integrated and robust multimodal systems. The development of such systems will not be achievable through intuition alone. Rather, it will depend on knowledge of the natural integration patterns that typify people's combined use of different input modes. This means that the successful design of multimodal systems will require guidance from cognitive science on the coordinated human perception and production of natural modalities. In this respect, multimodal systems can flourish only through multidisciplinary cooperation, as well as through teamwork among those with expertise in individual component technologies.

Multimodal Interaction: Separating Myth from Empirical Reality

In this article, 10 myths about multimodal interaction are identified as currently fashionable among computationalists and are discussed from the perspective of contrary empirical evidence. Current information about multimodal interaction is summarized from research on multimodal human-computer interaction, and from the linguistics literature on natural multimodal communication. In the process of uncovering misconceptions associated with each myth, information is highlighted on multimodal integration patterns and their temporal synchrony, the information carried by different input modes, the processibility of users' multimodal language, differences among users in multimodal integration patterns, and the reliability and other general advantages of multimodal system design. This state-of-the-art information is designed to replace popularized myths with a more

NORMAND COUSINEAU

accurate foundation for guiding the design of next-generation multimodal systems.

Myth #1: If you build a multimodal system, users will interact multimodally. Users have a strong preference to interact multimodally rather than unimodally, although this preference is most pronounced in spatial application domains [10]. For example, 95% to 100% of users preferred to interact multimodally when they were free to use either speech or pen input in a spatial domain [10]. However, just because users prefer to interact multimodally is no guarantee that they will issue every command to a system multimodally. Instead, they typically intermix unimodal and multimodal expressions. In a recent study, users' commands were expressed multimodally 20% of the time, with the rest just spoken or written [12].

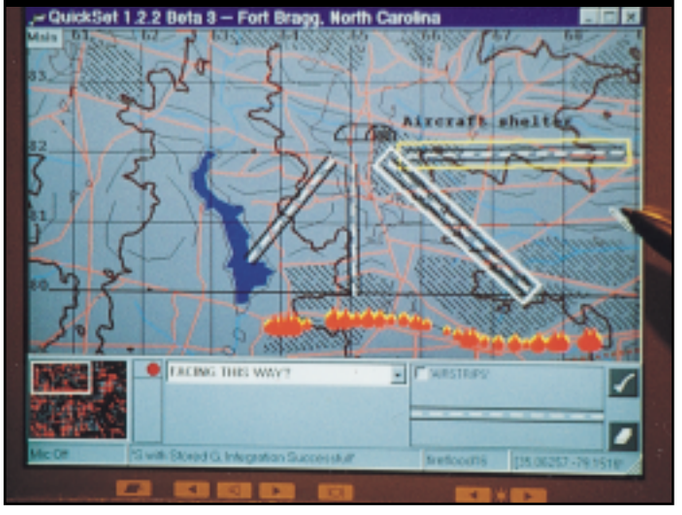
Predicting whether a user will express a command multimodally also depends on the type of action they are performing. In particular, they almost always express commands multimodally when describing spatial information about the location, number, size, orientation, or shape of an object. In the data summarized in Figure 2, users issued multimodal commands 86% of the time when they had to add, move, modify, or calculate the distance between objects on a map in a way that required specifying spatial locations [12]. They also were moderately likely to interact multimodally when selecting an object from a larger array—for example, when deleting a particular object from the map. However, when performing general actions without any spatial component, such as printing a map, users rarely expressed themselves multimodally—less than 1% of the time [12].

To summarize, users like being able to interact multimodally, but they don't always do so. Their natural communication patterns involve mixing unimodal and multimodal expressions, with the multimodal ones being predictable based on the type of action being performed. These empirical results emphasize that future multimodal systems will need to distinguish between instances when users are and are not communicating multimodally, so that accurate decisions can be made about when parallel input streams should be interpreted jointly versus individually. This data also suggests that knowledge of the type of actions to be included in an application should influence the basic decision of whether to build a multimodal system at all.

Myth #2: Speech and pointing is the dominant multimodal integration pattern. Since the development of Bolt's [1] "Put That There" system, computationalists have viewed speak-and-point as the prototypical form of a multimodal integration. In Bolt's original system, semantic processing was based on spoken input, but the meaning of a deictic term such as "that" was resolved by processing the x,y coordinate indicated by pointing at an object. Other multimodal systems also have attempted to resolve deictic expressions by tracking the direction of the human gaze [5].

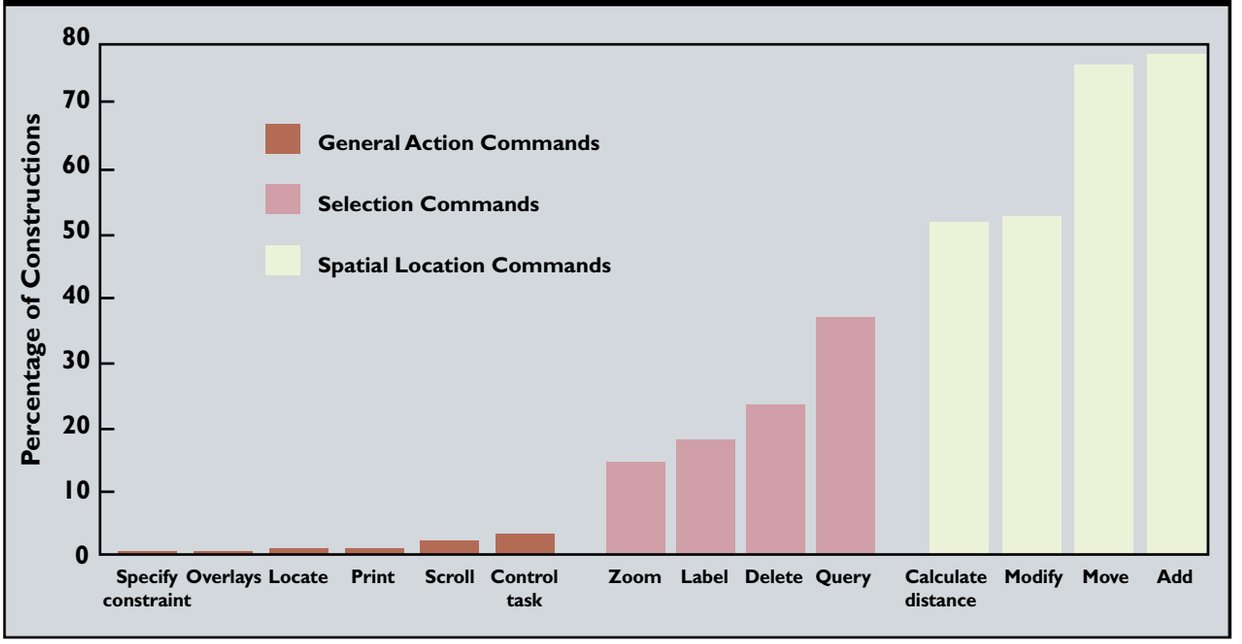
However, this concept of multimodal interaction as point-and-speak makes only limited use of new input modes for *selection* of objects—just as the mouse does. In this respect, it represents the persistence of an old mouse-oriented metaphor. In contrast, modes that transmit written input, manual gesturing, and facial expressions are capable of generating symbolic information that is more richly expressive than simple object selection. For example, studies of users' integrated pen/voice input indicate that a speak-and-point pattern only comprises 14% of all spontaneous multimodal utterances [12]. Instead, pen input is used more often to create graphics, symbols and signs, gestural marks, digits and lexical content. During interpersonal multimodal communication, linguistic analysis of spontaneous manual gesturing also confirms that deictic gestures (pointing) account for less than 20% of all gestures [6]. This data highlights the fact that any multimodal system designed exclusively to process speak-and-point will fail to provide users with much useful functionality. For this reason, specialized algorithms for processing deictic-point relations will have only limited practical use in the design of future multimodal systems.

Figure 1. Quickset multimodal pen/voice system on a handheld PC.



Myth #3: Multimodal input involves simultaneous signals. Another common assumption is that signals involved in any multimodal construction will

Figure 2. Percentage of commands that users expressed multimodally as a function of type of task command—with high levels during spatial location commands (right), moderate levels during selection commands (middle), and negligible levels during general action commands (left).

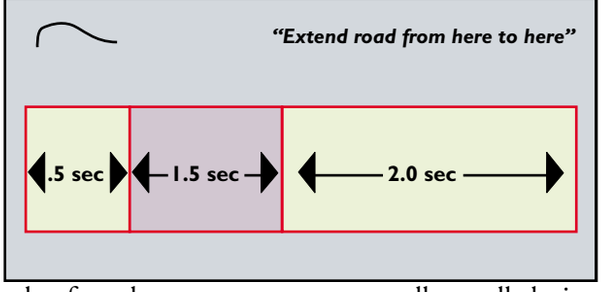


co-occur temporally. This temporal overlap then determines which signals to combine during system processing. For example, successful processing of the deictic term "that square" would rely on interpretation of pointing when the word "that" is spoken in order to extract the intended referent. However, one empirical study indicated that users often do not speak deictic terms at all and, when they do, the deictic frequently is not overlapped in time with their pointing. In fact, it has been estimated that as few as 25% of users' commands actually contain a spoken deictic that overlaps with the pointing needed to disambiguate its meaning [12].

Beyond the issue of deixis, users' spoken and pen-based input frequently do not overlap at all during multimodal commands to a computer. As illustrated in Figure 3, they are sequentially integrated about half the time, with pen input preceding speech and a brief lag between input signals of one or two seconds [12]. This finding is consistent with linguistics data revealing that both spontaneous gesturing and signed language often precede their spoken lexical analogues during human communication [4, 7]. The degree to which gesturing precedes speech is greater in topic-prominent languages such as Chinese than it is in subject-prominent ones like English [6].

In short, although speech and gesture are highly interdependent and synchronized during multimodal interaction, synchrony does not imply simultaneity. The empirical evidence reveals that multimodal signals often do not co-occur temporally at all during human-computer or natural human communication. Therefore, computationalists should not count on conveniently overlapped signals in order to achieve successful processing in the multimodal architectures they build.

Figure 3. A sequentially integrated multimodal command, with pen input preceding speech and a brief lag between signals.



Myth #4: Speech is the primary input mode in any multimodal system that includes it. Historically, linguists and computationalists alike have viewed speech as a primary input mode—with gestures, head and body movements, direction of gaze, and other input secondary. Speech is viewed as self-sufficient, with other modes being redundant accompaniments that carry little new or significant information. This perspective has biased early multimodal systems toward mainly processing speech input, and also toward the primitive speak-and-point integrations in

The flexibility of a multimodal interface can accommodate a wide range of users, tasks, and environments for which any given single mode may not suffice.

which a secondary mode is used only for simple selection. Sometimes secondary modes also have been viewed as useful when the primary speech signal is degraded (for example, in a noisy environment), in which case they might supply needed information when confidence in speech recognition is low. However, such views fail to acknowledge that other modes can convey information that is not present in the speech signal at all—for example, spatial information specified by pen input [10], and manner of action information specified by gesturing [6]. Multimodal systems that ignore the sources of such information will systematically fail to recognize many types of spontaneous multimodal construction.

Table 1. Percentage of multimodal commands per user involving simultaneous (SIM) vs. sequential (SEQ) integration of spoken and written signals.

User	SIM	SEQ
SIM integrators:		
U1	**86**	14
U2	**92**	8
U3	**94**	6
U4	**100**	0
SEQ integrators:		
U5	31	**69**
U6	25	**75**
U7	17	**83**
U8	11	**89**
U9	0	**100**
U10	0	**100**
U11	0	**100**

Speech also is not primary in terms of being the first input signal during multimodal constructions. Pen input precedes speech in 99% of sequentially-integrated multimodal commands, and in the majority of simultaneously-integrated ones as well [12]. This earlier production of manually-oriented input (writing or gestures) is believed to provide context, and also to assist users in planning their speech.

In short, speech is neither the exclusive carrier of important content, nor does it have temporal precedence over other input modes. As a result, the belief that speech is primary risks underexploiting the valuable roles to be played by other modes in next-generation multimodal architectures.

Myth #5: Multimodal language does not differ linguistically from unimodal language. It frequently is assumed that "language is language is language," so why should multimodal language differ in its basic form from other unimodal types of language—such as speech, writing, or keyboard? In fact, it recently has been demonstrated that multimodal pen/voice language is briefer, syntactically simpler, and less disfluent than users' unimodal speech [10]. In one study, a user added a boat dock to an interactive map system by speaking: *"Place a boat dock on the east, no, west end of Reward Lake."* However, when interacting multimodally using pen/voice input the same user completed the same action by indicating: **[draws rectangle]** *"Add dock."*

When free to interact multimodally, users selectively eliminate many linguistic complexities. As illustrated here, they prefer not to speak error-prone spatial location descriptions ("on the east, no, west end of Reward Lake") if a more compact and accurate alternative is available, such as pen input. They also use far less linguistic indirection and fewer co-referring expressions, which reduces the need for anaphoric tracking and resolution during natural language processing [11]. In other significant ways, multimodal language is simply different than spoken or textual language. For example, during pen/voice commands users' language departs from the subject-verb-object word order typical of English [12]—a difference that also has important implications for successful natural language processing.

In short, multimodal language is different than traditional unimodal forms of natural language, and in many respects it is substantially simplified. One implication for computationalists is that multimodal language may be easier to process, which could support more robust systems in the future.

Myth #6: Multimodal integration involves redundancy of content between modes. It often is claimed that the propositional content conveyed by different modes during multimodal communication contains a high degree of redundancy. However, the dominant theme in users' natural organization of multimodal input actually is complementarity of content, not redundancy—see Figure 4. For example,

speech and pen input consistently contribute different and complementary semantic information—with the subject, verb, and object of a sentence typically spoken, and locative information written [12]. Even during multimodal correction of system errors, when users are highly motivated to clarify and reinforce their information delivery, speech and pen input rarely express redundant information—less than 1% of the time. During human communication, linguists also have documented that spontaneous speech and gesturing do not involve duplicate information [2, 6].

In short, actual data highlights the importance of complementarity as a major organizational theme during multimodal communication. The designers of next-generation multimodal systems therefore should not expect to rely on duplicated information when processing multimodal language.

Myth #7: Individual error-prone recognition technologies combine multimodally to produce even greater unreliability. Another common misconception is that any multimodal system incorporating two error-prone recognition technologies, such as speech and handwriting recognition, will result in compounded errors and even greater performance unreliability. However, multimodal systems actually can support *more* robust recognition, not less—such that the error-handling problems typical of recognition technologies become more manageable. In part, this increased robustness is due to leveraging from users' natural intelligence about when and how to deploy input modes effectively. In a flexible multimodal interface, people will avoid using an input mode that they believe is error-prone for certain content. Their language also is simpler, as discussed previously, which further minimizes errors. When a recognition error does occur, users alternate input modes in a way that tends to resolve it effectively. This error resolution occurs because the confusion matrices differ for any given lexical content for the different technologies involved in the mode alternation.

The increased robustness of multimodal systems also depends on designing an architecture that integrates modes synergistically. In a well-designed and optimized multimodal architecture, there can be *mutual disambiguation* of two input signals [9]. For example, if a user says "ditches" but the speech recognizer confirms the singular "ditch" as its best guess, then parallel recognition of several graphic marks could result in recovery of the correct plural interpretation. This recovery can occur in a multimodal architecture even though the speech recognizer initially ranked the plural interpretation "ditches" as a less preferred choice on its n-best list.

Figure 5 illustrates another example of mutual disambiguation from a Quickset user's log. In this case, the user said "pan" and drew an arrow. Although neither the speech nor gesture were first on their n-best lists, the correct interpretation was recovered successfully on the final multimodal n-best list. This recovery was achievable because inappropriate signal pieces are discarded during the unification process, which imposes semantic, temporal, and other constraints on legal multimodal commands.

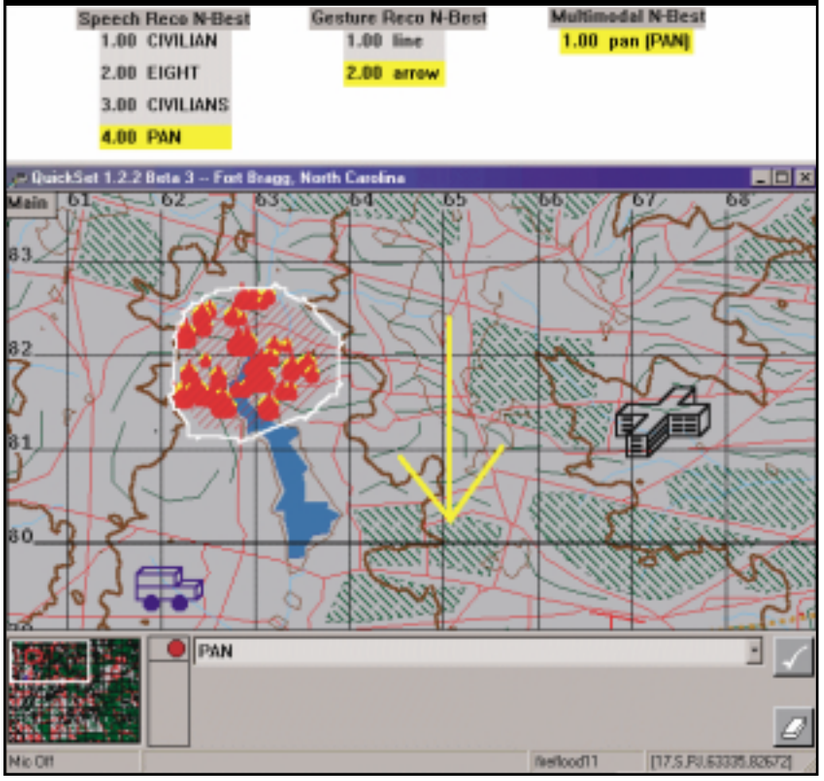
Due to mutual disambiguation, the parallel recognition and semantic interpretation that occurs in a multimodal architecture can yield a higher likelihood of correct interpretation than recognition based on either single input mode. This improvement is a direct result of the disambiguation between signals that can occur in a well-designed multimodal system, which exhibits greater performance stability and overall robustness as a result. The superior error-handling characteristics of multimodal systems represent a major performance advantage. During the next decade, we are increasingly likely to see new media for which recognition is error-prone being embedded within multimodal architectures in order to harness and stabilize them more effectively.

Figure 4. In the Quickset architecture, the semantic unification process capitalizes on the complementarity of information supplied by different modes, as well as exerting linguistic and temporal constraints on what satisfies an acceptable semantic blend.

Myth #8: All users' multimodal commands are integrated in a uniform way. When users interact multimodally, there actually can be large individual differences in integration patterns. In a recent study, users adopted either a simultaneous or sequential integration pattern when combining speech and pen input. For example, Table 1 shows that users 1–4 spoke and wrote so their signals were overlapped temporally, whereas users 5–11 combined signals sequentially.

Each user's dominant integration pattern was iden-

Figure 5. An example of mutual disambiguation in which the fourth choice "pan" on the speech n-best list and the second choice "arrow" on the gesture list both were pulled up during unification to produce a correct multimodal interpretation.



tified when they first began interacting with the system, and then persisted throughout their session. That is, each user's integration pattern was established early and remained consistent, although two distinct integration patterns were observed among different users. These findings imply that multimodal systems that can detect and adapt to a user's dominant integration pattern could lead to considerably improved recognition rates.

Myth #9: Different input modes are capable of transmitting comparable content. As an alternative extreme to the view that speech is primary, the concept of "alt-modes" also has emerged recently. This myth characterizes different input modes as fully able to transmit comparable propositional content. According to this technology-oriented perspective, simple translation is possible among different modes, which basically are interchangeable. Those who assert this myth believe that it is possible to design an idealized "everyperson information kiosk"—with tailorable input and output modalities to suit any user's physical, perceptual, or cognitive limitations. In the everyperson information kiosk, diverse communication modalities would be coordinated in a mechanistic plug-and-play manner to create the ultimate multimodal translation device.

Although the everyperson information kiosk may be an admirable goal, its presumptions fail to acknowledge that different modes represented by the emerging technologies that recognize speech, handwriting, manual gesturing, head movements, and gaze each are strikingly unique. They differ in the type of information they transmit, their functionality during communication, the way they are integrated with other modes, and in their basic suitability to be incorporated into different interface styles. None of these modes is a simple analogue of another in the sense that would be required to support simple one-to-one translation.

Different modes basically vary in the degree to which they are capable of transmitting similar information, with some modes relatively more comparable (speech and writing) and others less so (speech and gaze). Although speech and writing may convey many similar concepts, they still differ in the range and precision of their expressivity. For example, it often is infeasible to speak complex spatial shapes, relations among graphic objects, or precise location information—although such information is trivial to sketch using a pen. And whereas speech delivers information to a listener in a direct and intentional way, a modality like gaze reflects the speaker's focus of interest more passively and unintentionally, and may not convey useful information at all during periods of blank staring. Such extreme differences between input modes make them suitable candidates for qualitatively different interface styles. For example, speech input may function well within a command or conversational interface, whereas gaze may be more compatible as part of a noncommand interface concept.

Myth #10: Enhanced efficiency is the main advantage of multimodal systems. It often is assumed that the enhanced speed and efficiency enabled by parallel input is the primary performance advantage of a multimodal system, compared with a unimodal or graphical interface. For example, during multimodal pen/voice interaction in a spatial domain, a speed-up of 10% has been documented in compar-

ison with a speech-only interface [10]. However, this efficiency advantage may be limited to spatial domains, since it has not been demonstrated when task content is verbal or quantitative in nature [10].

There are other advantages of multimodal systems that are more noteworthy in importance than modest speed enhancement. For example, task-critical errors and disfluent language can drop by 36–50% during multimodal interaction [10]. Users' strong and nearly universal preference to interact multimodally likely constitutes another more consequential advantage. A third more significant advantage is the flexibility that multimodal systems permit users in selecting and alternating between input modes. Such flexibility makes it possible for users to alternate modes so that physical overexertion is avoided for any individual modality. It also permits substantial error avoidance and easier error recovery, as discussed previously. Finally, the flexibility of a multimodal interface can accommodate a wide range of users, tasks, and environments—including users who are temporarily or permanently handicapped, usage in adverse settings (noisy environments, for example) or while mobile, and other cases for which any given single mode may not suffice. In many of these real-world instances, integrated multimodal systems have the potential to support entirely new capabilities that have not been supported at all by previous traditional systems.

Conclusion

The ability to develop future multimodal systems depends on knowledge of the natural integration patterns that typify people's combined use of different input modes. Given the complex nature of users' multimodal interaction, cognitive science will play an essential role in guiding the design of robust multimodal systems. In this respect, a multidisciplinary perspective will be more central to successful system design than it has been in traditional domains previously tackled by computer science.

The design of multimodal systems that blend input modes synergistically depends on intimate knowledge of the properties of different modes and the information content they carry, what characteristics are unique to multimodal language and its processibility, and how multimodal input is integrated and synchronized. It also relies on predicting when users are likely to interact multimodally, and how alike different users are in their integration patterns. Finally, optimizing the robustness of multimodal architectures depends on a clear understanding of the advantages of this type of system, compared with unimodal ones. In the future, specific design challenges will include developing multimodal architectures that can handle the time-critical nature of parallel interdependent input signals, as well as ones optimized for error avoidance and robustness.

Ten myths regarding multimodal interaction have been identified and discussed from the viewpoint of contrary empirical evidence. In separating myth from reality, the goal has been to reveal the nature of multimodal interaction more clearly, which in turn provides a better foundation for guiding the design of future multimodal systems. C

References

1. Bolt, R.A. Put that there: Voice and gesture at the graphics interface. *ACM Computer Graphics 14*, 3 (1980), 262–270.
2. Cassell, J., Pelachaud, C., Badler, N., et al. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Computer Graphics, Annual Conference Series*, ACM Press, NY, 1994, 413–420.
3. Cohen, P., Johnston, M., McGee, D., et al. Quickset: Multimodal interaction for distributed applications. In *Proceedings of the Fifth ACM International Multimedia Conference* (New York, NY) ACM Press, NY, 1997, 31–40.
4. Kendon, A. Gesticulation and speech: Two aspects of the process of utterance. In M. Key, Ed. *The Relationship of Verbal and Nonverbal Communication*. The Hague, Mouton, 1980, 207–227.
5. Koons, D.B., Sparrell, C.J. and Thorisson, K.R. Integrating simultaneous input from speech, gaze, and hand gestures. In *Intelligent Multimedia Interfaces.* M. Maybury, Ed. MIT Press, Menlo Park, CA, 1993, 257–276.
6. McNeill, D. *Hand and Mind: What Gestures Reveal about Thought.* University of Chicago Press, Chicago, IL, 1992.
7. Naughton, K. Spontaneous gesture and sign: A study of ASL signs co-occurring with speech. In *Proceedings of the Workshop on the Integration of Gesture in Language & Speech* (Oct. 7–8, Newark and Wilmington, DE). L. Messing, Ed., University of Delaware, 1996, 125–134.
8. Neal, J.G. and Shapiro, S.C. Intelligent multi-media interface technology. In *Intelligent User Interfaces*. J.W. Sullivan and S.W. Tyler, Eds. ACM, NY, 1991, 11–43.
9. Oviatt, S.L. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the Conference on Human Factors in Computing Systems CHI'99* (May 18–20, Pittsburgh, PA). ACM Press, NY, 1999, 576–583.
10. Oviatt, S.L. Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction 12*, (1997), 93–129.
11. Oviatt, S.L. and Kuhn, K. Referential features and linguistic indirection in multimodal language. In *Proceedings of the International Conference on Spoken Language Processing.* Sydney, ASSTA Inc., 2339–2342.
12. Oviatt, S.L., DeAngeli, A. and Kuhn, K. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of Conference on Human Factors in Computing Systems CHI'97* (March 22–27, Atlanta, GA). ACM Press, NY, 1997, 415–422.

Sharon Oviatt (oviatt@cse.ogi.edu) is a professor in the Department of Computer Science and Engineering at the Oregon Graduate Institute of Science and Technology (OGI), and Co-Director of the Center for Human-Computer Communication.

This research was supported by Grant No. IRI-9530666 from the National Science Foundation, Grant No. DABT63-95-C-007 from DARPA, and by grants, contracts, and equipment donations from Boeing, Intel, Microsoft, NTT Data, and Southwestern Bell.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

© 1999 ACM 0002-0782/99/1100 $5.00