Multimodal Human-Computer Interaction

- Sharma, R., Pavlovic, V.I., Huang, T.S., “Toward multimodal human-computer interface”, *Proceedings of the IEEE*, 86-5, 853-869, 1998.
- Oviatt, S., “Ten myths of multimodal interaction,” *Communications of the ACM* vol:42 iss:11 pp:74-81, 1999.

Toward Multimodal Human–Computer Interface

RAJEEV SHARMA, MEMBER, IEEE, VLADIMIR I. PAVLOVIĆ, STUDENT MEMBER, IEEE, AND THOMAS S. HUANG, FELLOW, IEEE

Invited Paper

Recent advances in various signal-processing technologies, coupled with an explosion in the available computing power, have given rise to a number of novel human–computer interaction (HCI) modalities—speech, vision-based gesture recognition, eye tracking, electroencephalograph, etc. Successful embodiment of these modalities into an interface has the potential of easing the HCI bottleneck that has become noticeable with the advances in computing and communication. It has also become increasingly evident that the difficulties encountered in the analysis and interpretation of individual sensing modalities may be overcome by integrating them into a multimodal human–computer interface.

In this paper, we examine several promising directions toward achieving multimodal HCI. We consider some of the emerging novel input modalities for HCI and the fundamental issues in integrating them at various levels—from early "signal" level to intermediate "feature" level to late "decision" level. We discuss the different computational approaches that may be applied at the different levels of modality integration. We also briefly review several demonstrated multimodal HCI systems and applications. Despite all the recent developments, it is clear that further research is needed for interpreting and fusing multiple sensing modalities in the context of HCI. This research can benefit from many disparate fields of study that increase our understanding of the different human communication modalities and their potential role in HCI.

***Keywords*—** *Human–computer interface, multimodality, sensor fusion.*

NOMENCLATURE

AGR	Automatic gesture recognition.
ANN	Artificial neural network.
ASR	Automatic speech recognition.
BAC	Brain-activated control.
DP	Dynamic programming.
DTW	Dynamic time warping.
EBP	Error back propagation.
EEG	Electroencephalograph.
EM	Expectation maximization.
EMG	Electromyograph.
FIDO	Feature in, decision out.
FIFO	Feature in, feature out.
HCI	Human–computer interaction.
HMM	Hidden Markov model.
LRT	Likelihood ratio test.
MAP	Maximum *a posteriori* (estimator).
ML	Maximum likelihood (estimator).
MLP	Multilayer perceptron.
MOG	Mixture of Gaussians.
MS-TDNN	Multistate time-delay ANN.
OAA	Open agent architecture.
PDA	Personal digital assistant.
VR	Virtual reality.

Manuscript received July 15, 1997; revised November 30, 1997. The Guest Editor coordinating the review of this paper and approving it for publication was A. M. Tekalp. This work was supported in part by the National Science Foundation under Grant IRI-96-34618 and in part by the U.S. Army Research Laboratory under Cooperative Agreement DAAL01-96-2-0003.

R. Sharma is with the Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802 USA (e-mail: rsharma@cse.psu.edu).

V. I. Pavlović and T. S. Huang are with The Beckman Institute and Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801 USA (e-mail: vladimir@ifp.uiuc.edu; huang@ifp.uiuc.edu).

Publisher Item Identifier S 0018-9219(98)03281-2.

I. INTRODUCTION

With the ever increasing role of computers in society, HCI has become an increasingly important part of our daily lives. It is widely believed that as the computing, communication, and display technologies progress even further, the existing HCI techniques may become a bottleneck in the effective utilization of the available information flow. For example, the most popular mode of HCI still relies on the keyboard and mouse. These devices have grown to be familiar but tend to restrict the information and command flow between the user and the computer system. This limitation has become even more apparent with the emergence of novel display technology such as virtual reality [1]–[3] and wearable computers [4], [5]. Thus, in recent years, there has been a tremendous interest in introducing new modalities into HCI that will potentially resolve this interaction bottleneck.

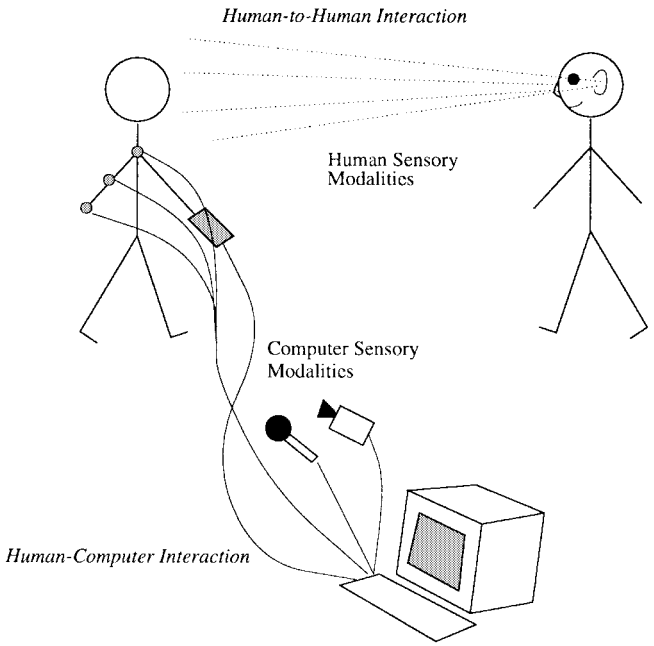


Fig. 1. Human-to-human interaction and human-to-computer interaction. Humans perceive their environment through five basic senses. HCI, on the other hand, need not be bounded to typical human senses.

One long-term goal in HCI has been to migrate the "natural" means that humans employ to communicate with each other into HCI (Fig. 1). With this motivation, ASR has been a topic of research for decades [6]. Some other techniques like automatic gesture recognition, analysis of facial expressions, eye tracking, force sensing, or EEG have only recently gained more interest as potential modalities for HCI. Though studies have been conducted to establish the feasibility of these novel modalities using appropriate sensing and interpretation techniques, their role in HCI is still being explored. A limiting feature of modern interfaces that has also become increasingly evident is their reliance on a single mode of interaction—a mouse movement, key press, speech input, or hand motion. Even though it may be adequate in many cases, the use of a single interaction mode proves to be inept in HCI. For example, in manipulating a [three-dimensional (3-D)] virtual object, a user may employ [two-dimensional (2-D)] mouse motion to select the object, then point with the mouse at a control panel to change the object's color. On the other hand, in a more natural setup, the same user would point at the object with his hand and say: "Make it green." Almost any natural communication among humans involves *multiple, concurrent modes of communication*. Surely, any HCI system that aspires to have the same naturalness should be multimodal. Indeed, studies have shown that people prefer to interact multimodally with computers, since among other things, such interaction eases the need for specialized training [3], [7]. The integration of multimodal input for HCI can also be seen from the perspective of *multisensor data fusion* [8]. Different sensors can, in that case, be related to different communication modalities. It is well known that multiple types of sensors may increase the accuracy with which a quantity can be measured by reducing the uncertainty in decision making [8], [9]. From a biological point of view, there is a clear evidence that the integration of multiple sensory modalities occurs in the human superior colliculus [9]. Yet, the use of multiple integrated interaction modalities in the HCI systems has not been adequately explored.

In this paper, we consider four basic questions relevant for multimodal HCI:

- *Why* integrate multiple modalities?
- *Which* modalities to integrate?
- *When* to integrate multiple modalities?
- *How* to integrate multiple modalities?

There are numerous potential benefits in integrating multiple modalities into HCI. The reasons range from the fact that natural human interaction itself has a multimodal character to the statistical advantages of combining multiple observations. *Why* integrate multiple modalities is discussed in Section II.

Natural human-to-human interaction is perceived through five basic senses and expressed through various actions such as voice, hand and body movements, facial expression, etc. However, an HCI system does not have to confine itself to these sensors and actions but may also take advantage of other computer-sensing modalities, like the EEG. We discuss some of the promising new HCI modalities in Section III.

Once the desired HCI modalities are selected, an important question to be addressed is *how* to combine them. To address this problem, it is helpful to know how the integrating modalities relate in a natural environment. Some modalities, like speech and lip movements, are more closely tied than others, such as speech and hand gestures. It is also plausible to assume that integration of such different combinations of modalities should be explored at *different* levels of integration. Depending on the chosen level of integration, the actual fusion can then be performed using numerous methods, ranging from simple feature concatenation to complex interaction of interface agents. The issues and techniques of *when* and *how* to combine multiple modalities are considered in Sections IV and V, respectively.

In Section VI, we briefly review some implemented multimodal HCI systems. These systems have incorporated multiple modalities at various integration levels into their interfaces. A particular speech/gesture system for controlling virtual-reality display is considered in greater detail as a case study of multimodal integration. This is followed by our discussion in Section VII and concluding remarks in Section VIII.

II. Why Multiple Modalities in HCI?

The interaction of humans with their environment (including other humans) is naturally multimodal. We speak about, point at, and look at objects all at the same time. We also listen to the tone of a person's voice and look at a person's face and arm movements to find clues about his feelings. To get a better idea about what is going on

around us, we look, listen, touch, and smell. When it comes to HCI, however, we usually use only one interface device at a time—typing, clicking the mouse button, speaking, or pointing with a magnetic wand. The "ease" with which this unimodal interaction allows us to convey our intent to the computer is far from satisfactory. An example of a situation when these limitations become evident is when we press the wrong key or when we have to navigate through a series of menus just to change an object's color. We next discuss the practical, biological, and mathematical rationales that may lead one to consider the use of *multimodal interaction* in HCI.

A. Practical Reasons

Practical reasons for multimodal HCI stem from some inherent drawbacks of modern HCI systems that undermine their effectiveness. HCI systems today are unnatural and cumbersome. They are based on "Stone Age" devices like the mouse, joystick, or keyboard, which limit the ease with which a user can interact in today's computing environments, including, for example, immersive virtual environments. Several studies have confirmed that people prefer to use multiple-action modalities for virtual object manipulation tasks [3], [7]. Such studies were based on the "Wizard of Oz" experiments, where the role of a multimodal computer is played by a human "behind the scenes" [10]. In [3], Hauptmann and McAvinney concluded that 71% of their subjects preferred to use both speech and hands to manipulate virtual objects rather than just one of the modalities alone. Oviatt has shown in [7] that 95% of the subjects in a map manipulation task tend to use gestures together with speech. Multiple modalities also *complement* each other. Cohen has shown [11], for example, that gestures are ideal for direct object manipulation, while natural language is more suited for descriptive tasks.

Another drawback of current advanced single-modality HCI is that it lacks robustness and accuracy. For example, modern ASR systems have advanced tremendously in recent years. However, they are still error-prone in the presence of noise and require directed microphones or microphone arrays. Automatic gesture-recognition systems have just recently gained popularity. Whether they use a stylus or a glove or are vision based, they are still constrained to the recognition of few predefined hand movements and are burdened by cables or strict requirements on background and camera placement [12]. However, concurrent use of two or more interaction modalities may loosen the strict restrictions needed for accurate and robust interaction with the individual modes. For instance, spoken words can *affirm* gestural commands, and gestures can *disambiguate* noisy speech. Gestures that complement speech, on the other hand, carry a complete communicational message only if they are interpreted together with speech and, possibly, gaze. The use of such multimodal messages can help reduce the complexity and increase the naturalness of the interface for HCI. For example, in computer-vision-based gesture recognition, in addition to the input from the images, the gesture recognition could be influenced by the speech, gaze direction, and content of the virtual display. To exploit this multimodality, for example, instead of designing a complicated gestural command for the object selection, which may consist of a deictic gesture followed by a symbolic one (to symbolize that the object that was pointed at by the hand is supposed to be selected), a simple concurrent deictic gesture and verbal command "this" can be used (as will be discussed in Section VI).

Another pragmatic reason for using multiple modalities in HCI, particularly with redundant input, is to enable physically or cognitively handicapped people access to computers (or computer-controlled devices). For example, the use of hand gestures and American Sign Language with the help of computer vision, the use of eye tracking combined with speech recognition, and the use of EEG-based control would help the physically challenged. With multimodality built into the HCI, the need for building special-purpose interfaces for individual disability will be greatly eased.

B. Biological Reasons

A rationale for integration of multiple sensory modalities can be found in nature. Human beings as well as other animals integrate multiple senses.

Studies of superior colliculus have shown that different senses are initially segregated at the neural level. When they reach the brain, sensory signals converge to the same target area in the superior colliculus, which also receives signals from the cerebral cortex, which, in turn, modulates resultant behavior. A majority (about 75%) of neurons leaving the superior colliculus are *multisensory*. This strongly suggests that the use of multimodality in HCI would be desirable, especially if the goal is to incorporate the naturalness of human communication into HCI. We further expose these issues in a related section on biological foundations for multimodal integration (see Section V-A). Another thorough discussion pertaining to this topic, including additional references, can be found in [9].

C. Mathematical Reasons

More insight on *why* but also *how* and *when* to integrate multiple modalities comes from the field of *sensory data fusion*. Data fusion as a field of study has existed for many decades. However, its main thrust has been in the area of target detection. The goal of data fusion for target detection is to find optimal ways of integrating different sensory data (radar, infrared, etc.), which produce "best" detection rates. The reason for combining different sensors has its origins in statistical data analysis. The disadvantage of using a single sensor system is that it may not be able adequately to reduce the uncertainty for decision making. Uncertainty arises when features are missing, when the sensor cannot measure all relevant attributes, or when observations are ambiguous [9]. On the other hand, it is well known that it is statistically advantageous to combine multiple observations from the same source because improved estimates are obtained using redundant observations [8]. It is also known

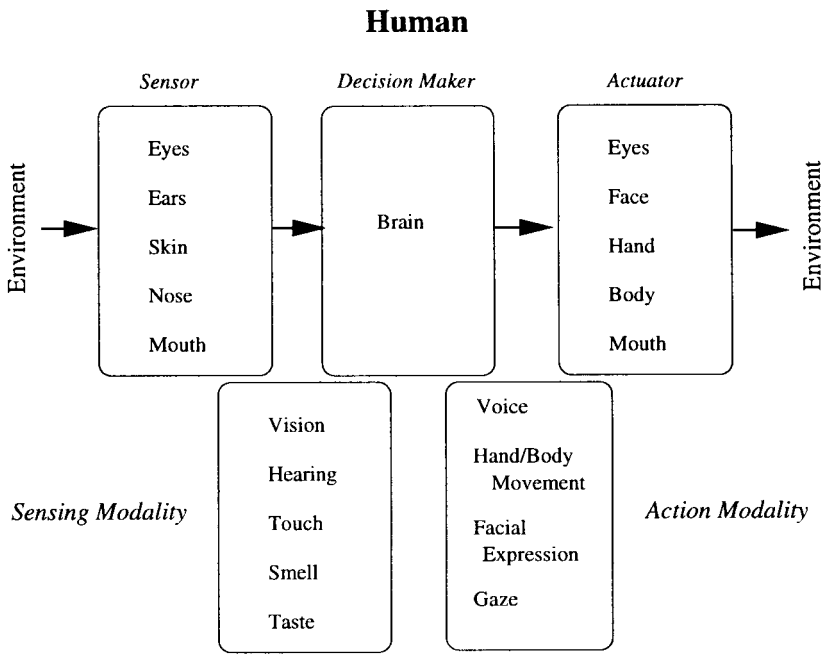


Fig. 2. Modalities for human sensing and action. Human beings sense the environment in which they live through their senses. They act on the environment using numerous actuators.

that multiple types of sensors may increase the accuracy with which a quantity can be observed. Formally, if x_i and x_j are two (statistically independent) estimates of an observed quantity, the minimum mean square error (MSE) combination of the two estimates results in

$$x_{ij} = \left(\Sigma_i^{-1} + \Sigma_j^{-1}\right)^{-1}\Sigma_i^{-1}x_i + \left(\Sigma_i^{-1} + \Sigma_j^{-1}\right)^{-1}\Sigma_j^{-1}x_j \tag{1}$$

where Σ_i and Σ_j are the variances of x_i and x_j, respectively. Moreover, the variance of the fused estimate Σ_{ij} is given by

$$\Sigma_{ij}^{-1} = \Sigma_i^{-1} + \Sigma_j^{-1}. \tag{2}$$

Thus, the variance of the fused estimate Σ_{ij} is "smaller" than the variances of either of the two original estimates. This can be easily generalized to more than two redundant observations. Clearly, the advantage of having multimodal HCI is substantiated by the purely statistical point of view.

III. Modalities for HCI

Humans perceive the environment in which they live through their *senses*—vision, hearing, touch, smell, and taste. They act on and in it using their *actuators* such as body, hands, face, and voice. Human-to-human interaction is based on sensory perception of actuator actions of one human by another, often in the context of an environment (Fig. 2). In the case of HCI, computers perceive actions of humans. To have the human–computer interaction be as natural as possible, it is desirable that computers be able to interpret all natural human actions. Hence, computers should interpret human hand, body, and facial gestures, human speech, eye gaze, etc. Some computer-sensory modalities are analogous to human ones. Computer vision and ASR mimic the equivalent human sensing modalities. However, computers also possess sensory modalities that humans lack. They can accurately estimate the position of the human hand through magnetic sensors and measure subtle changes of the electric activity in the human brain, for instance. Thus, there is a vast repertoire of human-action modalities that can potentially be perceived by a computer.

In the rest of this section, we review the individual modalities for HCI. The modalities are discussed under the two categories of *human-action modalities* and *computer-sensing modalities*. Fig. 3 shows how the two categories relate to each other. A particular human-action modality (e.g., speaking) may be interpreted using more than one computer-sensing modality (e.g., audio and video). We discuss a sampling of issues related to each of the individual modalities, some of which may be resolved by using multimodal integration in HCI.

A. Human-Action Modalities for HCI

A large repertoire of human actions could possibly be incorporated into HCI by designing suitable sensing mechanisms. Historically, the action modalities most exploited for HCI are based on hand movements. This is largely due to the dexterity of the human hand which allows accurate selection and positioning of mechanical devices with the help of visual feedback. Appropriate force and acceleration can also be applied easily using the human hand. Thus, the hand movement is exploited in the design of numerous interface devices—keyboard, mouse, stylus, pen, wand, joystick, trackball, etc. The keyboard provides a direct way of providing text input to the computer, but the speed is obviously limited and can only be improved to a certain rate. Similarly, hand movements cause a cursor to move on the computer screen (or a 3-D display). The next level of action modalities involves the use of *hand gestures*,

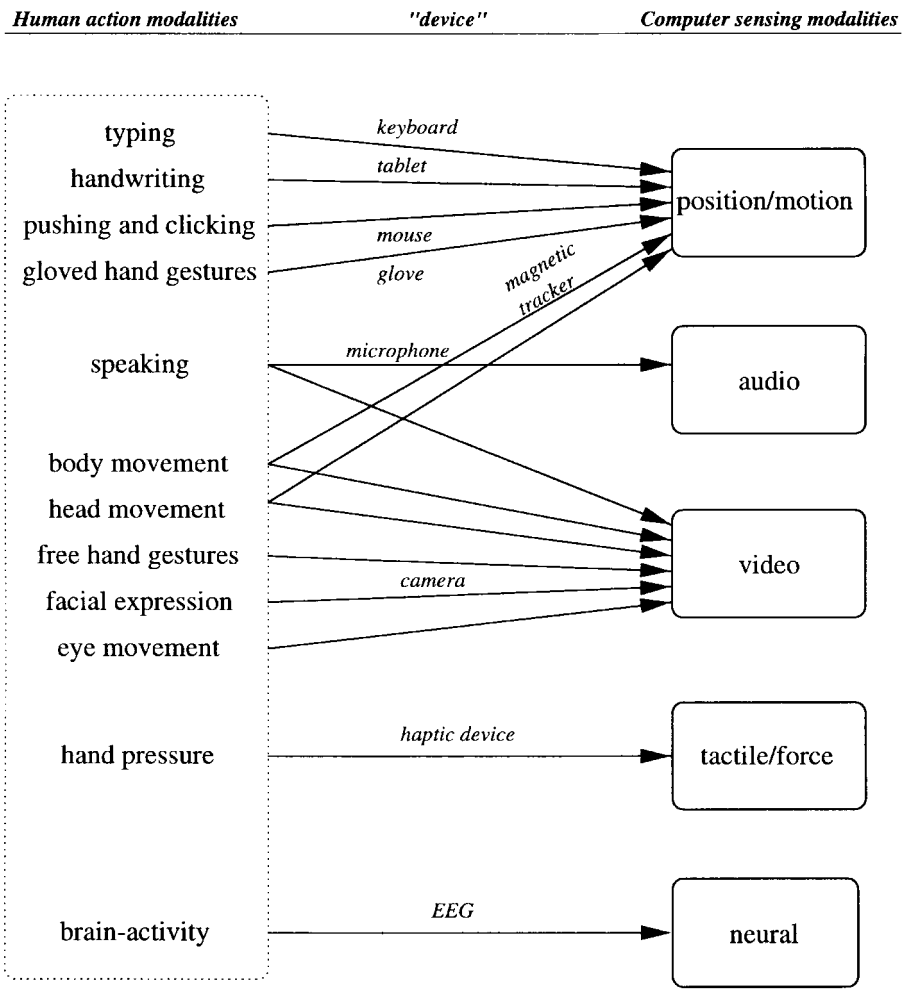


Fig. 3. Mapping of different human-action modalities to computer-sensing modalities for HCI. Multiple human actions, such as facial expressions and hand or eye movement, can be sensed through the same "devices" and used to infer different information.

ranging from simple pointing through manipulative gestures to more complex symbolic gestures such as those based on American Sign Language. With a glove-based device, the ease of hand gestures may be limited, but with noncontact video cameras, free-hand gestures would be easier to use for HCI. The role of free-hand gestures in HCI could be further improved (requiring lesser training, etc.) by studying the role of gestures in human communication. A multimodal framework is particularly well suited for embodiment of hand gestures into HCI.

In addition to hand movements, a dominant action modality in human communication is the production of sound, particularly spoken words. The production of speech is usually accompanied by other visible actions, such as lip movement, which can be exploited in HCI as well. Where the human is looking can provide a clue to the intended meaning of a particular action or even serve as a way of controlling a display. Thus, eye movements can be considered a potential action modality for HCI. The facial expression and body motion, if interpreted appropriately, can help in HCI. Even a subtle "action" like a controlled thought has been investigated as a potential candidate for HCI.

B. Computer-Sensing Modalities for HCI

What action modality to use for HCI is largely dependent on the available computer-sensing technology. We next discuss the broad categories of the computer-sensing modalities and consider how the above human-action modalities might by measured and interpreted.

1) Position and Motion Sensing: A large number of interface devices have been built to sense the position and motion of the human hand and other body parts for use in HCI. The simplest such interface is the keyboard, where the touch of a particular key indicates that one of a set of possible inputs was selected. More accurate position and motion sensing in a 2-D plane is used in interface devices such as a mouse, light pen, stylus, and tablet [7], [13]. Three-dimensional position/motion sensing is commonly done through a joystick or a trackball. For a brief history of HCI technology covering these familiar computer-sensing modalities, we refer the reader to [14]. In position/motion sensing, both relative and absolute measurements are made, dictated by the type of position/motion transducer used. With the advent of 3-D displays and virtual reality, there was a need to track the position of head, fingers, and

other main body parts for controlling the display. For tracking the head (to display the graphics with the correct perspective), various forms of sensors have been employed. Electromagnetic fields [15] are the most popular method but are expensive and restricted to a small radius, typically about 5–20 ft. Ultrasonic tracking requires line of sight and is inaccurate, especially at long ranges, because of variation in the ambient temperature [16]. Other methods might include tracking of infrared light-emitting diodes or inertial trackers using accelerometers. Attempts to solve hand tracking resulted in mechanical devices that directly measure hand and/or arm joint angles and spatial position. This group is best represented by *glove-based devices* [17]–[21]. Glove-based gestural interfaces require the user to wear a cumbersome device and generally carry a load of cables that connect the device to a computer. This may hinder the ease and naturalness with which the user interacts with the computer-controlled environment.

2) Audio Sensing: The direct motivation for sensing the sound waves using a (set of) microphone(s) and processing the information using techniques known as ASR is to be able to interpret speech, the most natural human-action modality for HCI. Significant advances have been made toward the use of ASR for HCI [6]. The current ASR technology is still not robust, however, especially outside controlled environments, under noisy conditions and with multiple speakers [22]. Methods have been devised, for example, by using microphone arrays and noise cancellation techniques to improve the speech recognition. However, these tend to work only for the environments for which they are designed. An active research area is concerned with making ASR sufficiently robust for use in HCI. For instance, it has been demonstrated conclusively that the recognition rate for speech can be improved by using visual sensing to analyze the lip motion simultaneously [23]. Other visual sensing modalities such as gesture analysis may also help in improving speech interpretation [24].

3) Visual Sensing: A video camera, together with a set of techniques for processing and interpreting the image sequence, can make it possible to incorporate a variety of human-action modalities into HCI. These actions include hand gestures [12], lip movement [23], gaze [25]–[27], facial expressions [28], and head and other body movements [29], [30]. For example, with the help of specially designed cameras and lighting, eye movements can be tracked at greater than 250 Hz and can be potentially used for controlling a display, either directly or indirectly, by designing multiresolution displays [31], [32]. Similarly, visually interpreted gestures can allow a tetherless manipulation of virtual-reality [33] or augmented-reality displays [34]. Use of visual sensing for HCI suffers difficulties from both a theoretical and practical standpoint. The problem, such as visual interpretation of hand gestures, is still not well understood, particularly when it is desirable not to put restrictions on the hand movements for more natural HCI [12]. From a practical standpoint, visual sensing involves the processing of huge amounts of information in real time, which could put undue demands on the processing power of the system being controlled. Furthermore, visual sensing requires an unoccluded view of the human, putting restrictions on the motion of the user and the physical setting for HCI. Nonetheless, the use of computer vision for improving HCI continues to be a topic of very active research [35]. Visual sensing can be especially useful in conjunction with other sensing modalities [36], such as lip reading with audio [23], lip reading with eye tracking [32], visual gesture recognition with speech [24], etc.

4) Tactile and Force Sensing: The dexterity of the human hand for accurately positioning a mechanical device can be coupled with application of "force," which can be sensed by using appropriate *haptic* devices. The computer sensing of touch and force is especially important for building a proper feel of "realism" in virtual reality. The key idea is that by exerting force or touch on virtual objects (with the corresponding haptic display for feedback), the user will be able to manipulate the virtual environment in a natural manner. Situations where such realism is especially important include, for example, simulation of surgery for training [37], [38]. Force sensing is a topic of very active research since is it difficult to design suitable devices with the desired accuracy without constraining the user [39], [40]. A better force sensing for HCI may also be obtained by simultaneously considering the sensing of position and motion.

5) Neural Sensing: One computer-sensing modality that has been explored with increasing interest is based on the monitoring of brain electrical (EEG) activity [41]–[44]. The brain activity can be monitored noninvasively from the surface of the scalp and used for directly controlling the computer display (Fig. 4). This form of interaction is also termed BAC. The "hands-free" nature of the resulting HCI makes it attractive for head-mounted displays and situations (such as aircraft piloting) where hands are being used in other tasks. Another very big impetus for pursuing this sensing modality is as a means of HCI for the physically disabled [45]. However, it requires training (using biofeedback and self-regulation) so that specific brain responses may be modulated [46]. There are many theoretical and applied open problems that need to be addressed for BAC, for example, how user distractions and/or increased workload affect such an interface, etc. An alternative approach includes sensing surface EMG signals [47]. Approaches have also been suggested for using multimodal sources that include eye tracking and monitoring of muscle tension in conjunction with EEG [48], [49].

IV. When to Integrate the HCI Modalities

The previous section introduced different types of modalities that may be "integrated" for multimodal HCI. Different sensing modalities yield disparate signal forms and rates. That makes successful integration of such signals a difficult and challenging task. In this section, we consider the problem of *when* to integrate the multiple modalities, which, in turn, determines the abstraction level at which the modalities are fused. Should they be fused at the "raw" sensory data level or at the higher "decision" level? How are

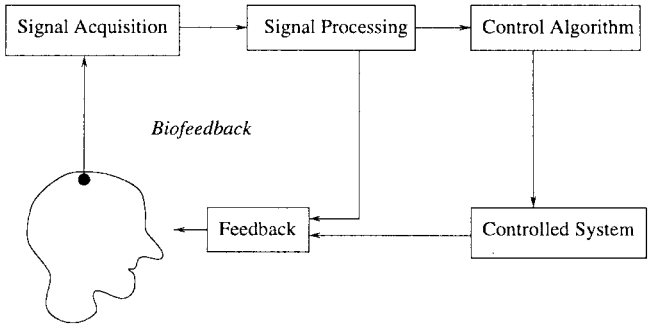


Fig. 4. The main components of an EEG-based system for controlling a display using brain activity.

multiple modalities synchronized? In solving this problem, one should perhaps consider the following questions.

- How closely coupled are the modalities in natural human interaction?
- Does this coupling remain in HCI?
- What are the possible levels of multimodal integration in HCI?

The first question is not easy to answer. The answers mostly originate in psychobehavioral studies concerned with the interaction of modalities. For instance, it is known that gestures and speech are intimately connected and are claimed to arise from a single mental concept [50]. Gestures occur synchronously with their semantically parallel speech units or just before them [50]. It is also claimed that the gaze follows the hand during gestural actions [51]. Speech and lip movements are also closely coupled [23], even more so than gestures and speech.

A question remains, however, as to whether such coupling persists when the modalities are used for HCI. Several "Wizard of Oz"-type studies have confirmed that it does. Oviatt [7], for example, has extensively studied the interaction of drawing gestures and speech. She has concluded that the integration occurs on a semantic level where gestures are used to convey information on location, while speech conveys the information on subject and action (verb) in a sentence.

To study the levels of multimodal integration in HCI, one can use the theoretical and computational apparatus developed in the field of *sensory data fusion*. For the most part, three distinct levels of integration can be distinguished [8]:

1) data fusion;
2) feature fusion;
3) decision fusion.

Integration of sensory data according to the above levels of fusion is depicted in Fig. 5. Another, more refined version of this classification can be found in [52].

Data fusion is the lowest level of fusion [Fig. 5(a)]. It involves integration of raw observations and can occur only in the case when the observations are of the same type. This type of fusion does not typically occur in multimodal integration for HCI since the modes of interaction are of a

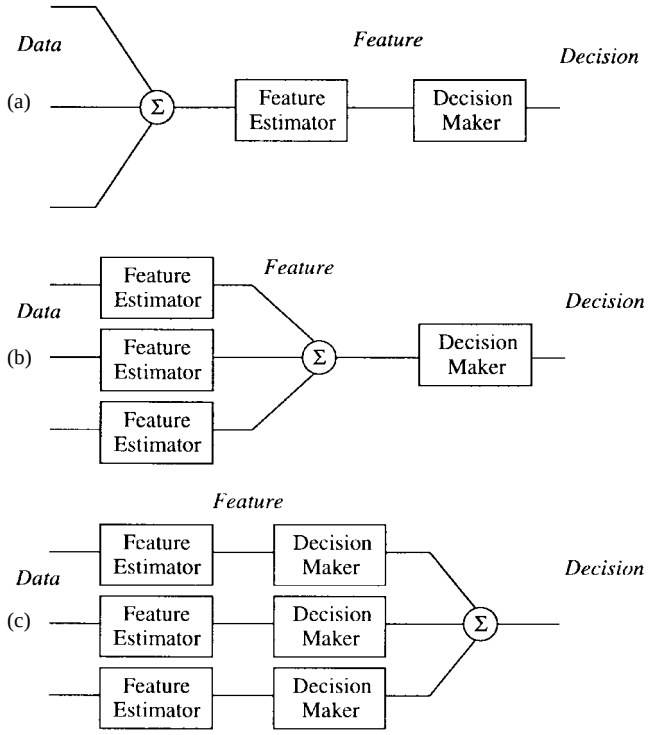


Fig. 5. The three different levels for fusing the multiple sensing modalities. (a) Data fusion fuses individual modes' data. (b) Feature fusion combines features of individual modalities. (c) Decision fusion integrates individual modes' decisions.

different nature (gestures and speech, for instance) and are observed using different types of sensors (video camera and microphone, for instance). It can occur, for example, when one or more cameras are used to capture visual information on one object. Data-level fusion is characterized by the highest level of information detail out of the three fusion types. It also assumes a high level of synchronization of the multimodal observations. Unfortunately, data fusion is also highly susceptible to noise, specific nature of individual sensors, sensor failures, and sensor alignment [52].

Feature fusion is more commonly found in integration of modalities for HCI. It assumes that each stream of sensory data is first analyzed for features, after which the features themselves are fused [Fig. 5(b)]. This type of fusion is appropriate for closely coupled and synchronized modalities, possibly speech and lip movement. Feature-level fusion retains less detailed information than data fusion but is also less sensitive to noise. However, feature sets can be quite large. This high cardinality can result in soaring computational cost for this fusion approach [52].

The type of fusion most commonly found in HCI is the so-called decision-level fusion. Decision-level fusion is based on the fusion of individual mode decisions or interpretations [Fig. 5(c)]. For example, once an arm movement is interpreted as a deictic (pointing) gesture and a spoken sentence is recognized as "Make *this* box white," the two can be fused to interpret that a particular object (box) needs to be painted white. Synchronization of modalities in this case pertains to synchronization of decisions on a semantic level. Decision fusion is the most robust and resistant to

individual sensor failure. It has a low data bandwidth and is generally less computationally expensive than feature fusion. One disadvantage of decision-level fusion is that it potentially cannot recover from loss of information that occurs at lower levels of data analysis and thus does not exploit the correlation between the modality streams at the lower integration levels.

Finding an optimal fusion level for a particular combination of modalities is not straightforward. A good initial guess can be based on the knowledge of the interaction and synchronization of those modes in a natural environment. However, it still remains necessary to explore multiple levels of fusion in order to determine the optimal combination of the desired modalities.

V. How to Integrate the HCI Modalities

As mentioned in the previous section, the level (data *versus* feature *versus* decision) at which the integration is done strongly influences the actual computational mechanism used for the fusion. In this section, we discuss the different mechanisms that may be used for the integration of multiple modalities in the context of HCI. First, we discuss the plausible biological basis for integration. This is followed by a general model of fusion that is used to discuss "how" to carry out the integration at the feature and decision levels.

A. Biological Foundations

An insight into *how* to combine multiple modalities can be gained from neurological models of sensor fusion. One such model was proposed by Stein and Meredith in [53]. The model suggests that the fusion of sensory neurons coming from individual sensors occurs in the brain structure know as *superior colliculus*. Superior colliculus is thought of as being responsible for orienting and attentive behavior. Two facts relevant to multimodal fusion can be gathered from their model.

1) *Evidence accruement:* Sensory evidence in superior colliculus seems to be *accrued* rather than averaged over different sensors inputs. In other words, the response of multisensory neurons leaving the colliculus is stronger when multiple weak-input sensory signals are present than when a single strong signal exists.
2) *Contextual dependency:* Besides receiving input from sensory neurons, superior colliculus also receives signals from the cerebral cortex. These signals modulate the fusion of sensory neurons, thus inducing *contextual feedback*. They are also responsible for different combinations of sensory signals based on the context.

Another important issue for multimodal fusion is addressed in the studies of perceptual sensory fusion. This issue tackles the problem of how *discordances* between individual sensors are dealt with. Discordances usually arise as a consequence of sensor malfunctioning. Dealing with them is clearly of utmost importance for the proper functioning of a system. According to Bower [54], there are four possible ways of dealing with sensory discordances:

1) *blind fusion:* sensor observations are fused without any regard for individual discordances;
2) *fusion with sensor recalibration:* an attempt is made to recalibrate discordant sensors;
3) *fusion with sensor suppression:* discordant sensors are suppressed;
4) *no fusion:* discordant sensors are not fused.

The last three fusion categories indicate the existence of feedback in the signal fusion processes in biological entities. Obviously, this provides one with a good rationale to consider those issues when tackling the problem of multimodal fusion for HCI.

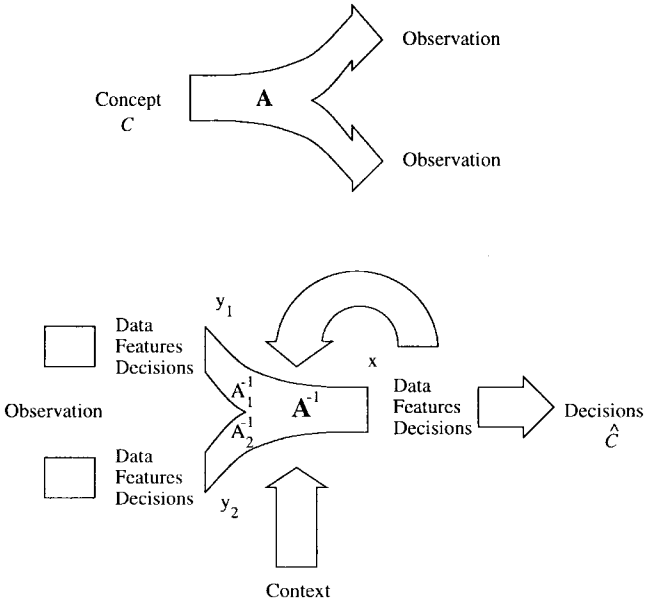


Fig. 6. A general model for multimodal production and fusion in HCI. Multiple observations are produced through different action modalities from the same concept "C." To infer the driving concept, multiple observations need to be reintegrated in the multimodal fusion process.

B. A General Fusion Model

We consider a general model of multimodal fusion for HCI built on the foundations of sensory data fusion theory while also taking into account the biological evidence about integration of multiple senses. The model is depicted in Fig. 6. The fusion model is built under the assumption that each concept behind any human action is expressed through multiple action modalities and is perceived through multiple sensory modalities. The goal of multimodal fusion is then to integrate different abstractions of the observed actions (data, features, or decisions) such that it can best infer a decision about the driving concept. As part of such a process, contextual knowledge can be used to constrain the search space of the problem.

Let us denote by y_i, $i = 1, 2, \cdots, N$ observation abstractions of the concept $\mathcal{C}$ perceived through one of N sensory modalities. As mentioned before, such abstractions

can be raw observations (data), features estimated from raw observations, or even individual mode decisions based on unimodal observations. Let Y_i be the set of all observations of modality i, $Y_i = \{y_{i1}, y_{i2}, \cdots\}$. Let x be an abstraction of concept $\mathcal{C}$. Again, this could be the concept $\mathcal{C}$ itself or a feature in the concept space $\mathcal{X}$. The *production model* in additive noise environment can then be defined as

$$y_i = A_i(x) + \omega_i,\ i = 1, 2, \cdots, N \tag{3}$$

where $A_i(\cdot)$ is the mapping that transduces a mental concept to an observable action and ω_i is an additive noise. The task of inferring the best concept $\hat{x}$ given some cost function J, sets of observations $Y_1, \cdots, Y_N$, and context $\kappa(x) = 0$ can then be formulated as the following constrained optimization problem:

$$\hat{x} = \arg,\ \min_{x \in \mathcal{X}} J(x|Y_1, \cdots, Y_N),\ \text{such that}\ \kappa(x) = 0. \tag{4}$$

A common way of defining the inference problem employs the *Bayesian framework* (see, for instance, [55] for an overview of Bayesian inference). In this framework, the cost function is given as the probability of making a wrong decision $J(x|Y_1, \cdots, Y_N) = 1 - P(x|Y_1, \cdots, Y_N)$. Assuming, for the sake of simplicity, that context plays no role, this leads to

$$\hat{x}_{\text{MAP}} = \arg,\ \max_x P(x|Y_1, \cdots, Y_N). \tag{5}$$

Hence, (4) now assumes the form of the MAP estimator. Another equivalent form of this criterion is often found in data fusion and detection literature and follows from the *Bayes* inference rule:

$$\hat{x}_{\text{MAP}}\colon \frac{P(Y_1, \cdots, Y_N|\hat{x})}{P(Y_1, \cdots, Y_N|x)} > \frac{P(x)}{P(\hat{x})}, \qquad \forall x \in \mathcal{X}. \tag{6}$$

Test (6) is commonly referred to as the *Bayes LRT* [56]. The quantity on the left side of (6) is known as the likelihood ratio $\Lambda(Y)$, and the quantity on the right side is the threshold η.

A frequently used assumption in the integration of multiple modalities is that of *conditionally independent observations*. In other words

$$P(y_1, \cdots, y_N|x) = P(y_1|x) \cdots P(y_N|x). \tag{7}$$

Under this assumption, the MAP LRT reduces to

$$\hat{x}_{MAP}\colon \quad \frac{P(Y_1|\hat{x})}{P(Y_1|x)} \cdots \frac{P(Y_N|\hat{x})}{P(Y_N|x)} > \frac{P(x)}{P(\hat{x})} \tag{8}$$

$$: \quad \Lambda(Y_1) \cdots \Lambda(Y_N) > \eta, \forall x \in \mathcal{X}. \tag{9}$$

This equation can be viewed as the distributed version of the Bayes (fusion) rule.

Sometimes, however, the prior probabilities of individual multimodal concepts $P(x)$ are not known. In the absence of this knowledge, it is best to assume that all concepts are equally likely. This leads us to another criterion for multimodal fusion, known as the ML criterion

$$\hat{x}_{\text{ML}} = \arg,\ \max_x P(Y_1, \cdots, Y_N|x). \tag{10}$$

Again, using the independence-of-observations assumption from (7), the ML criterion becomes

$$\begin{aligned} \hat{x}_{\text{ML}} &= \arg,\ \max_x \prod_{i=1}^{N} P(Y_i|x) \\ &= \arg,\ \max_x \sum_{i=1}^{N} \log P(Y_i|x). \end{aligned} \tag{11}$$

Often, a weighted form of the ML is used to emphasize the importance of individual modes

$$\hat{x}_{\text{ML}_\text{w}} = \arg,\ \max_x \sum_{i=1}^{N} w_i, \log P(Y_i|x) \tag{12}$$

where w_i are the modal weights such that $\sum_{i=1}^{N} w_i = 1$. For example, in many bimodal speech-recognition applications, weights w_i are employed to indicate the dominant communication mode in different noise environments.

Equipped with these tools, one can now approach the task of multimodal fusion from the perspective of fusion on different integration levels, as discussed in Section IV. In the following discussion, we address two categories of multimodal fusion techniques often found in HCI—feature-level and decision-level fusion techniques.

C. Feature-Level Fusion

Feature-level fusion techniques are concerned with integration of features from individual modalities into more complex multimodal features and decisions. Using the terminology from data-fusion literature [52], feature-fusion techniques can be classified as twofold:

- FIFO techniques;
- FIDO techniques.

FIFO techniques yield fused multimodal features. This implies the need for an additional feature classifier to infer the multimodal decision. Kalman filters are an often accounted type of FIFO fusers. Unlike FIFO integrators, FIDO integrators do not require a separate classification unit. Feature fusion and classification are inherently connected in this architecture. FIDO fusion frequently employs structures known as *probabilistic networks*, such as artificial neural networks and HMM's.

1) Kalman Fusion: Multimodal fusion on a feature level can easily be formulated using the *Kalman filter* approach [57], [58]. Instead of performing fusion of a time series of feature vectors, however, the fusion is in this case performed over a sequence of features belonging to different modalities. The Kalman fusion approach is based on the assumption that the production model (3) is known. For ease of discussion, we assume that this model is linear and that there are only two modalities to be fused. Assume also that the additive noise in (3) is independent across different modalities, $\omega_1 \perp \omega_2$. We can then combine observations from the two modalities into a single vector $y^T = [y_1^T y_2^T]$ and write

$$y = Ax + \omega \tag{13}$$

where

$$A^T = [A_1^T A_2^T]$$
$$\omega^T = [\omega_1^T \omega_2^T] \tag{14}$$
$$\Omega = E[\omega\omega^T].$$

Let $\hat{x}_-$ and $\hat{x}_+$ be the estimates of x before and after new observations from two modalities are fused, respectively, and let Σ_- and Σ_+ be their corresponding variances. Then

$$\hat{x}_j = \Sigma_j[A_j^T \Omega_j^{-1} y_j + \Sigma_-^{-1}\hat{x}_-] \tag{15}$$
$$\Sigma_j^{-1} = \Sigma_-^{-1} + A_j^T \Omega_j^{-1} A_j,\ j = 1, 2. \tag{16}$$

Similarly, from (13), we can obtain the following Kalman filter equation:

$$\hat{x}_+ = \Sigma_+[A^T \Omega^{-1} y + \Sigma_-^{-1}\hat{x}_-] \tag{17}$$
$$\Sigma_+^{-1} = \Sigma_-^{-1} + A^T \Omega^{-1} A. \tag{18}$$

Last, substituting (14) into (17) and using (15), we have

$$\hat{x}_+ = \Sigma_+[\Sigma_1^{-1}\hat{x}_1 + \Sigma_2^{-1}\hat{x}_2 - \Sigma_-^{-1}\hat{x}_-] \tag{19}$$
$$\Sigma_+^{-1} = \Sigma_1^{-1} + \Sigma_2^{-1} - \Sigma_-^{-1}. \tag{20}$$

Equation (19) clearly has the same form as (8).

From (19), it is obvious that in the Bayesian case with independent observations, the fused evidence is formed by averaging the evidence from individual modalities. However, a general Bayesian inference framework can result in a highly nonlinear evidence fusion.

The above feature fusion equations can easily be generalized for the fusion case of N conditionally independent modalities. Similarly, an *extended Kalman filter* [57], [58] approach can be used if (3) is nonlinear. Extensions to multiscale Kalman fusion have also been explored [59]. In general, the Kalman filter approach is the most commonly found approach in feature-to-feature fusion.

2) ANN's: An ANN is a type of statistical classifier often used in pattern recognition. More precisely, it is a graph where variables are associated with nodes of the graph and variable transformations are based on propagation of numerical messages along the links of the graph [60]. The network is used for classification of inputs y into N classes $X \in \mathcal{X} = \{X_1, X_1, \cdots, X_N\}$. In other words, given an observation y, the network estimates the density $P(x|y)$. To show that a structure like an ANN can indeed be used as a density estimator, consider that according to the Bayes rule

$$\begin{aligned} P(x = X_1|y) &= \frac{P(y|x = X_1)P(x = X_1)}{P(y)} \\ &= \frac{P(y|x = X_1)P(x = X_1)}{\sum_i P(y|x = X_i)P(x = X_i)} \\ &= \frac{1}{1 + \exp\left\{\sum_i -\ln\left[\frac{P(y|x=X_1)}{P(y|x=X_i)}\right] - \ln\left[\frac{P(x=X_1)}{P(x=X_i)}\right]\right\}} \\ &= \frac{1}{1 + \exp(z_1)} \\ &= g(z_1), \end{aligned} \tag{21}$$

$g(\cdot)$ is called the logistic sigmoid function. Assuming that conditional probabilities are Gaussian (or of general exponential family) with identical covariance matrices, it readily follows that

$$z_1 = w_1^T y + w_{11}. \tag{22}$$

Thus, the conditional density computation can be achieved using a simple one-layer linear network with the sigmoid node transfer functions. In fact, it can be shown (see [61], for example) that a network of the form

$$u_k(y) = v_k^T g[z(y)] + v_{k1} \tag{23}$$

with "properly" selected weights v and w yields $P(x = X_k|y) = u_k(y)$. The network of the above type is also known as the MLP and is frequently used in pattern-recognition tasks. Many other ANN architectures have been proposed in a variety of different contexts [62].

Selection of network weights is done through network training. Based on a corpus of training data with known classification, the network weights are modified until the network "learns" the classification. Training procedures differ widely depending on the network architecture. For MLP, for instance, a gradient-based optimization technique, also know as EBP, is often used [63]. If optimized to a global minimum of the MSE cost function, such networks indeed behave as MAP density estimators. In practice, however, the weights usually correspond to a local minimum of the cost function, and the networks only approximate MAP estimators.

One drawback of "classical" ANN architectures is their inability efficiently to handle temporal sequences of features. Mainly, they are unable to compensate for changes in temporal shifts and scales. Several modified architectures have emerged that can handle such tasks. One of the most often used architectures is the so-called MS-TDNN [64]. This architecture handles changes in pattern temporal scale using the dynamic programming approach.

The adaptation of ANN's to multimodal fusion has often been inspired by biological origins. However, fusion architectures have mostly followed the line of regular ANN architectures with straightforward concatenation of features from multiple modalities into a joint feature vector. Some attempts have been made to design specific fusion architectures and fusion nodes. For example, Meier *et al.* in [65] designed an architecture that combines two TDNN's using a layer of "combination" nodes. The combination nodes' activation is determined as the weighted sum of the individual mode networks' output scores.

3) HMM's: HMM's are, like the ANN's, a special case of probabilistic Bayes networks [60]. They have been used successfully for more than a decade in the field of ASR [6]. Unlike ANN's, HMM's are designed to model the posterior densities of observations $P(y|x = X)$ over time and can, therefore, be used as ML estimators (10). An HMM is a doubly stochastic process, a network with *hidden* and *observable* states. The hidden states "drive" the model dynamics—at each time instance, the model is in

one of its hidden states. Transitions between the hidden states are governed by probabilistic rules. The observable states produce outcomes during hidden-state transitions or while the model is in one of its hidden states. Such outcomes are measurable by an outside observer. The outcomes are governed by a set of probabilistic rules. Thus, an HMM can be represented as a triplet (A, b, pi), where A is called the (hidden) state transition matrix, b describes the probabilities of the observation states, and π is the initial hidden-state distribution. It is common to assume that the hidden-state space is discrete and that the observables are allowed to assume a continuum of values. In such cases, b is usually represented as a MOG probability density functions (pdf's). The process of association of different HMM's with different concepts is denoted as training. In this process, the parameters of the HMM (A, b, pi) are modified so that the chosen model "best" describes the spatio/temporal dynamics of the desired concept. The training is, again, achieved by optimizing the likelihood measure $\log[P(y|x = X)]$ over the model parameters. Such optimization involves the use of computationally expensive EM procedures, like the Baum–Welch algorithm [66]. However, any such training procedure involves a step based on DP, which in turn has a DTW property. This means that the variability in duration of training samples is accounted for in the model. The same is true for the recognition or model-evaluation process. A probability of the observation's being produced by each HMM is evaluated using a DP *forward/backward* or *Viterbi* algorithm. Obviously, the larger the number of trained HMM's, the more computationally demanding the recognition procedure. To help address this problem successfully, an external set of rules or *grammar* is imposed, which describes the language sentence structure or how the trained units can be "connected" in time [33], [67]. Several problems are related to the typical use of the HMM as a recognition tool. For example, in its original formulation, an HMM is a first-order stochastic process. This implies that the (hidden) state of the model at time instance i depends only on the state at time $i-1$. While this model may be adequate for some processes, it often results in lower recognition rates for the processes that do not follow the first-order Markov property. As in speech, such problems can be somewhat reduced by extending the parameter vectors with the time derivatives of the original parameters [68]. It is also possible to derive higher order HMM's; however, such models do not share the computational efficiency of the first-order models [6]. Another possible drawback of classical HMM's is the assumption that pdf's of the observables can be modeled as MOG's. The main reason for modeling the observables as MOG's is to ease the training. In such cases, the HMM parameters can be efficiently computed using the Baum–Welch algorithm. Extensions in this direction have been achieved in speech recognition by using neural networks to model the observation pdf's [61]. Unfortunately, the training procedure in that case is computationally overwhelming. Also, in the original formulation, an HMM is assumed to be *stationary*. Nonstationary versions of HMM's have been recently formulated for speech recognition [69].

Feature fusion context can be introduced into HMM's by modeling the observations as concatenated multimodal feature vectors. Such integration architectures have been considered for the fusion of speech and lip movements [70] and speech and hand gestures [71]. However, possibly due to the differences in the time scale of the features from the two modalities, such architectures do not perform well. One attempt to alleviate this problem was introduced in [72] through an HMM-like architecture called the Boltzmann zipper. In the Boltzmann zipper, each hidden state can "belong" to only one of the multiple modalities (audio or video, for example) but not to both, as is the case in classical multimodal HMM's. This architecture has been applied to bimodal speech recognition and has shown improvement in fusion characteristics [23] over the concatenation approach. HMM's have been utilized with much more success as individual feature classifiers at the decision-level multimodal fusion (see Section V-D).

D. Decision-Level Fusion

Fusion on the decision level is the most frequently followed approach to multimodal integration. As depicted in Fig. 5, it involves fusion of concepts (decisions) from the individual modes to form a unique multimodal concept. An underlying assumption of this type of fusion is that the basic features of the individual modes are not sufficiently correlated to be fused at the feature level. Therefore, feature fusers/classifiers mentioned in Section V-C cannot be directly used for fusion. However, they are often used here too, but in a slightly different role, as decision makers for the individual modes. Several types of decision-level fusion mechanisms are commonly found in HCI systems. In the following discussion, we consider two such mechanisms: *frames* and *software agents*.

1) Frames: The concept of frames is commonly found in artificial intelligence literature. A frame is a unit of a knowledge source describing an object [73]. Each frame has a number of *slots* associated with it. The slots represent possible properties of the object, actions, or an object's relationship with other frames. This last property facilitates a mechanism for designing networks of frames for a particular context with links describing contextual semantics. Such networks are also known as semantic networks [74]. In the multimodal HCI context, different modalities can be associated with individual frame slots. Different modalities can describe particular properties of a virtual object. Speech can, for instance, designate the object's color, while gestures can imply the object's location. This is a case of the complementary role of modalities. It is also possible that multiple modalities indicate the same property of an object. In such cases, fusion can be achieved by selecting the property with the lowest joint cost. In the Bayesian framework, this is equivalent to choosing the highest prior or posterior joint probability. An alternative may be to consider the Dempster–Shafer combination of evidence

[75]. In that case, the evidence is actually accrued, which is consistent with the observed properties of biological data-fusion systems.

Frame-based multimodal HCI systems have been utilized ever since Bolt's early "Put-That-There" system [76]. This system used speech, gaze, and hand gestures to manipulate virtual objects. Many recent systems still use the same mechanism. For example, [77] used speech and pen-gesture frame fusion to design an interface for a calendar program. Many simple frame-based approaches have also been implemented for bimodal (audio and video) speech recognition [65], [70], [78]. Such approaches basically assume one-frame/one-slot networks for each of the two modalities. The slots describe phonemes observed through speech and lip movements. Two frames are fused by selecting the phoneme with the highest joint probability, $P(\text{video}|\text{phoneme})P(\text{audio}|\text{phoneme})$. The classifiers for the individual modes are commonly of the HMM type.

2) Software Agents: Software agents have recently emerged as a valuable tool for HCI [79]. A software agent is a software entity that functions continuously and autonomously in a particular environment, often inhabited by other agents and processes [79], [80]. Agents should be able to perform their activities without human intervention over long periods of time. They should learn from their experience and communicate and cooperate with other agents. Groups of agents can play roles of "digital butlers," "personal filters," or "digital sisters-in-law" [81]. In that respect, software agents are particularly useful in overcoming some problems of present-generation user interfaces. For example, current direct manipulation interfaces are inappropriate for large search spaces, are not adaptive and easy to learn, cannot learn from examples themselves, and, most important, cannot easily integrate multiple interaction modalities. On the other hand, software agents can be task oriented, flexible, adaptive, and can integrate modalities by delegating modal interaction to different communicating subagents.

OAA [82] is especially suitable for multimodal fusion tasks. In this architecture, one agent can autonomously handle speech recognition while another handles gestures and a third processes eye movements, for instance. The modal agents, in turn, communicate with a central agent known as a *facilitator*, which handles their interactions with other agents in the system who wish to receive multimodal information, such as a multimodal interpretation agent. This multimodal integration architecture provides an opportunity for implementation of sensor discordances detection, evidence accruement, and contextual feedback. However, the complexity of the architecture is greater compared to some of the other integration techniques. To handle this burden, the OAA facilitates distributed computing, in which different agents can exist on different computer platforms, ranging from workstations to hand-held personal assistants. One implementation of this architecture has been used in QuickSet, a multimodal interface for military simulation [83], which uses speech, handwriting, and pen gestures.

Fig. 7. A visual computing environment for structural biology (MDScope) used as a testbed for building a speech/gesture interface with decision-level fusion.

VI. Multimodal HCI Systems and Applications

Although there has been substantial research interest toward developing multimodal HCI systems, relatively few implemented HCI systems exhibit such multimodality. We first present in Section VI-A a description of an implemented speech/gesture system that can be considered as a case study of multimodal integration with fusion at the decision level. This is followed by a brief review of other reported multimodal systems and applications in Section VI-B.

A. A Gesture/Speech Interface for Controlling a 3-D Display

We summarize a case study in building a speech/gesture interface for a virtual-reality application (further details are given in [24]). The particular virtual environment that we considered is used by structural biologists in the Theoretical Biophysics Group at the University of Illinois at Urbana-Champaign. The system, called MDScope [84], is a set of integrated software components that provides an environment for simulation and visualization of biomolecular systems in structural biology (Fig. 7). To keep the interaction natural in a complex environment like MDScope, it is desirable to have as few devices attached to the user as possible. Motivated by this, we developed techniques that enable spoken words and simple free-hand gestures to be used while interacting with 3-D graphical objects in this virtual environment. The hand gestures are detected through a pair of strategically positioned cameras and interpreted using a set of computer-vision techniques that we term AGR. These computer-vision algorithms are able to extract the user hand from the background, extract positions of the fingers, and distinguish a meaningful gesture from unintentional hand movements using the context. The context of a particular virtual environment is used to place the necessary constraints to make the analysis robust and to develop a command language that attempts optimally to combine speech and gesture inputs.

The key goal of our work was to simplify model manipulation and rendering to such a degree that biomolecular modeling assumes a playful character; this will allow the researcher to explore variations of the model and concentrate on biomolecular aspects of the task without undue distraction by computational aspects. This helps in the process of developing an understanding of important properties of the molecules, in viewing simulations of molecular dynamics, and in "playing" with different combinations of molecular structures. One potential benefit of the system would be in reducing the time to discover new compounds—in research toward new drugs, for example.

The general AGR problem is difficult because it involves analyzing the human hand, which has a very high degree of freedom, and because the use of the hand gesture is not so well understood. (See [12] for a recent survey of vision-based AGR.) However, we use the context of the particular virtual environment to develop an appropriate set of gestural "commands." The gesture recognition is done by analyzing the sequence of images from a pair of cameras positioned such that they facilitate robust analysis of the hand images. The background is set to be uniformly black to further help with the real-time analysis without using any specialized image-processing hardware. In addition to recognizing a pointing finger, we have developed an HMM-based AGR system for recognizing a basic set of manipulative hand gestures. The gesture commands are categorized as being either dynamic (e.g., move back, move forward) or static (e.g., grab, release, stop, up, down). We have also developed a gesture command language for MDScope that is mainly concerned with manipulating and controlling the display of the molecular structures.

For integration of speech and gesture within the MDScope environment, a real-time decoding of the user's commands was required in order to keep pace with the hand gestures. Thus we used "word spotting," the task of detecting a given vocabulary of words embedded in unconstrained continuous speech. The recognition output stream consisted of a sequence of keywords and fillers constrained by a simple syntactical network. The recognizer that followed was developed by modifying the HMM implementation of HTK Toolkit by Entropic Research.

To utilize the information input from the user in the form of spoken words and simple hand gestures effectively, we designed a command language for MDScope that combines speech with gesture using a frame-based architecture. This command language employs the basic syntax of $<$ `action` $<$ `object` $<$ `modifier` $>$ and emphasizes both *complementary* and *reenforcing* roles of spoken and gestural modes. The $<$ `action` $>$ component is spoken (e.g., "rotate"), while the $<$ `object` $>$ and $<$ `modifier` $>$ are specified by a combination of speech and gesture. An example is speaking "this" while pointing, followed by a modifier to clarify what is being pointed to, such as "molecule," "helix," "atom," etc., followed by speaking "done" after moving the hand according to the desired motion. Another example of the desired speech/gesture capability is the voice command "engage" to query MDScope for the molecule that is nearest to the tip of the pointer and to make the molecule blink to indicate that it was selected and to save a reference to that molecule for future use. Once engaged, the voice command "rotate" converts the gesture commands into rotations of the chosen molecule and the command "translate" converts them into translations. When finished, the command "release" deselects the molecule and allows the user to manipulate another molecule.

This application shows a case where computer-vision and speech-recognition techniques are used for building a natural human–computer interface for a virtual-reality environment using spoken words and free hand gestures. The previous interface of MDScope was a keyboard and a magnetically tracked pointer. This is particularly inconvenient since the system is typically used by multiple (six to eight) users, and the interface hinders the interactive nature of the visualization system. Hence, incorporating voice command control in MDScope enabled the users to be free of keyboards and to interact with the environment in a natural manner. The hand gestures permitted the users easily to manipulate the displayed model and "play" with different spatial combinations of the molecular structures. The integration of speech and hand gestures as a multimodal interaction mechanism was more powerful than using either mode alone.

B. Other Multimodal HCI Systems

One of the first multimodal HCI systems can be accredited to Bolt [76]. His "Put-That-There" system fused spoken input and magnetically tracked 3-D hand gestures using a frame-based integration architecture. The system was used for simple management of a limited set of virtual objects such as selection of objects, modification of object properties, and object relocation. Even though the naturalness of the interaction was hindered by the limitations of the technology at that time, "Put-That-There" has remained the inspiration of all modern multimodal interfaces. The rest of this section focuses on some of its descendants.

QuickSet [83], [85] is a multimodal interface for control of military simulations using hand-held PDA's. It incorporates voice and pen gestures as the modes of interaction. This interface belongs to the class of decision-level fusers. It follows the OAA [82] with ten primary agents connected through a central facilitator. Recognition of pen gestures sensed through the PDA is conducted by the gesture agent. The agent utilizes ANN and HMM classifiers for concurrent gesture recognition. Multiple modalities in QuickSet play reenforcing roles. The modalities can automatically disambiguate each other using joint ML estimation. Alternatively, unimodal interaction can be enabled when one of the modes becomes unreliable. Situations like that may occur in high noise environments such as field posts during military exercises.

Another multimodal interface system was built for interacting with a calendar program called Jeanie [77]. The interface consists of autonomous speech, pen gesture, and handwriting-recognition modules. The speech-recognition

module is built upon the JANUS speech translation system [86]. It includes a semantic parser that can efficiently deal with unknown words and unrecognized fragments. The pen-gesture recognition module uses TDNN's to classify a small number of pen strokes sensed through a PDA, while the handwriting recognizer employs MS-TDNN's as classifiers. The fusion of the three modalities is performed in a frame-based fashion (see Section V-D1) followed by the dialogue manager interpretation of the fused information according to the current context. In the absence of recognition errors, the multimodal interpreter performs with 80% accuracy, while in its worst case, the accuracy drops to a low 35%.

VisualMan [87] is an application-independent multimodal user interface that combines eye gaze, voice, and 3-D motion. The interface is used for a 3-D virtual object manipulation in a Windows-based environment. Three-dimensional motion and eye gaze are integrated to provide positional information about the manipulated object, whereas spoken commands independently determine manipulative actions.

Finger-Pointer [88] integrates deictic free-hand gestures and simple static hand postures with voice commands for video presentation system control. Multimodal integration of the two modalities assumes that their roles are complementary—gestures specify positional information and identify simple object attributes (number of slides, for example), while spoken commands identify actions.

Virtual-World [89] is a combination of a flexible-object simulator and a multisensory user interface. It integrates hand gestures and hand motion detected with a glove-based device with spoken input. A dialogue manager is employed to map sensing device outputs to application parameters and results using a set of rules.

The integration of speech and lip reading has been extensively explored in recent years. The main goal of this fusion task has been to improve the recognition of speech in high noise environments. Approaches ranging from feature- to decision-level fusion have been tested for that purpose. For instance, [70] experimented with HMM-based feature fusion on feature and joint ML fusion on decision level (see Section V-C3). Results from their 40 word-recognition tasks indicate that decision-level bimodal fusion yields better accuracy compared to the feature-level approach. Similar results have been observed in [65], [72], and [78]. In [78], HMM fusion on feature level was tested against the decision level using the machine-learning C4.5 algorithm and showed again the superiority of the decision-level fusion. Fusion on the two levels tested in an ANN/TDNN setup by [65] confirmed the above observations.

An interesting application of multimodal interfaces lies in the domain of virtual autonomous agents. Autonomous agents are autonomous behaving entities in a dual virtual/real world who perform some actions in response to their perceived environment. Natural interaction and communication between the agents on one side and a human user on the other is crucial for effective system performance. We briefly survey a few such systems from the multimodal HCI standpoint.

The Artificial Life Interactive Video Environment (ALIVE) [90] is a system that allows wireless, full-body interaction between a human and a world of autonomous agents. The system uses a real-time, vision-based interface to detect and interpret the user's body motion in the context of a current virtual world and its artificial inhabitants. ALIVE has been used in numerous applications ranging from entertainment agents and personal teachers and trainers to interface agents and PDA's.

A similar approach is used in Smart Rooms [91]. Smart Rooms play the role of an invisible butler in trying to interpret what the user is doing and help him accomplish his tasks easily. Based on agent architecture, Smart Rooms use visual and auditory sensors to interpret hand gestures and speech and identify the user, for example. Extensions of this approach have been implemented as Smart Desks, Smart Clothes, and smart car interiors [92].

Another multimodal autonomous agent system called Neuro Baby (NB) [93], [94] was designed as a form of human companion. NB employs recognition of the user's voice intonation and mood as well as eye tracking and hand shaking via robotic hand. A feedback to the user is provided through the autonomous character's facial expressions and speech. The system has also evolved into a network NB, the intent of which is to provide two users, separated by a cultural gap, a way to communicate feelings in a nonverbal fashion.

VII. Discussion

With the massive influx of computers in society, the limitations of current human–computer interfaces have become increasingly evident. HCI systems today restrict the information and command flow between the user and the computer system. They are, for the most part, designed to be used by a small group of experts who have *adapted themselves* to the available HCI. For a casual user, however, the HCI systems are cumbersome and obtrusive and lack the "intelligence" expected by the user. Further, the HCI systems tend to confine the user to a less natural, *unimodal* means of communication. The ease with which such unimodal interfaces allow one to communicate with computers is far from satisfactory.

Integration of more than one modality into an interface would potentially overcome the current limitations and ease the information-flow bottleneck between the user and the computer. Ideally, in a multimodal HCI setup, the computers would adapt to the needs of the user. Cumbersome and obtrusive devices in one interaction modality would be replaced by more natural interface devices in other modalities. Such modalities do not necessarily have to be the ones employed in an analogous human-to-human communication. Computers possess sensory modalities that could prove to be superior to those that humans possess. Inaccurate interaction in one modality can be complemented by a more accurate interpretation in another modality or can be improved by combined interpretation through multiple modalities.

Some of the reported experimental HCI systems support the evidence of the potential benefits of multimodality in HCI. However, the limitations of these current systems are quite evident. A major hindrance is perhaps still in the inadequacies of the individual modalities that are used in the multimodal interface. For example, the performance of ASR is still highly context dependent, often below the desired robustness. Visual sensing involves real-time processing of huge amounts of data and thus suffers difficulties from both a theoretical and practical standpoint. Force sensing lacks suitable devices with desired accuracy without constraining the user. Sensing of neural information requires extensive training. These problems have restricted today's multimodal interfaces to a very narrow class of domains where the problems can be minimized. Further progress is needed in developing an understanding of the limitations of the individual modalities—to help both in making them better for HCI and for making them more suitable for integration with other modalities.

Strategies and techniques for multimodal integration are only beginning to emerge. The questions of *when* and *how* to merge multiple modalities have not yet been addressed in a satisfactory manner. More tightly coupled modalities such as speech and lip reading may call for integration at lower, feature levels, whereas modalities such as hand gestures and spoken language possibly require a semantic-level integration. Current multimodal HCI systems are often built using ad hoc approaches based more on intuition rather than systematic techniques and studies on human subjects. This can result in awkward solutions where the interaction requires unnatural mode couplings or where mode transitions need to be induced by the press of a button.

To alleviate the current problems, a massive effort focusing on several fundamental questions is necessary. How to increase robustness and accuracy of individual interface devices while minimizing their obtrusiveness? How to effectively couple multiple modalities in a natural manner? How to provide flexibility and adaptivity of the interaction? Better methodologies, perhaps based on systematic human studies, may need to be developed for evaluation of the multimodal interfaces. These evaluation techniques would ultimately determine the effectiveness of proposed techniques and architectures for multimodal HCI.

VIII. Concluding Remarks

Motivated by the tremendous need to explore better HCI paradigms, there has been a growing interest in developing novel sensing modalities for HCI. To achieve the desired naturalness and robustness of the HCI, multimodality would perhaps be an essential element of such interaction. Clearly, human studies in the context of HCI should play a larger role in addressing issues of multimodal integration. Even though a number of developed multimodal interfaces seem to be domain specific, there should be more systematic means of evaluating them. Modeling and computational techniques from more established areas such as sensor fusion may shed some light on how systematically to integrate the multiple modalities. However, the integration of modalities in the context of HCI is quite specific and needs to be more closely tied with subjective elements of "context." There have been many successful demonstrations of HCI systems exhibiting multimodality. Despite the current progress, with many problems still open, multimodal HCI remains in its infancy. A massive effort is perhaps needed before one can build practical multimodal HCI systems approaching the naturalness of human–human communication.

References

[1] J. A. Adam, "Virtual reality," *IEEE Spectrum*, vol. 30, no. 10, pp. 22–29, 1993.

[2] H. Rheingold, *Virtual Reality.* New York: Summit Books, 1991.

[3] A. G. Hauptmann and P. McAvinney, "Gesture with speech for graphics manipulation," *Int. J. Man-Machine Studies*, vol. 38, pp. 231–249, Feb. 1993.

[4] S. Mann, "Wearable computing: A first step toward personal imaging," *IEEE Computer Mag.*, vol. 30, pp. 25–32, Feb. 1997.

[5] R. W. Pickard and J. Healey, "Affective wearables," in *Proc. Int. Symp. Wearable Computing*, Cambridge, MA, Oct. 1997.

[6] L. R. Rabiner and B. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall, 1993.

[7] S. Oviatt, A. DeAngeli, and K. Kuhn, "Integration and synchronization of input modes during multimodal human–computer interaction," in *Proc. Conf. Human Factors in Computing Systems (CHI'97)*, Atlanta, GA, pp. 415–422.

[8] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proc. IEEE*, vol. 85, pp. 6–23, Jan. 1997.

[9] R. R. Murphy, "Biological and cognitive foundations of intelligent data fusion," *IEEE Trans. Syst., Man, Cybern.*, vol. 26, pp. 42–51, Jan. 1996.

[10] D. Salber and J. Coutaz, "Applying the Wizard of Oz technique to the study of multimodal systems," in *Proc. EWHCI'93*, Moscow, Russia, 1993.

[11] P. R. Cohen, M. Darlymple, F. C. N. Pereira, J. W. Sullivan, R. A. Gargan, Jr., J. L. Schlossberg, and S. W. Tyler, "Synergic use of direct manipulation and natural language," in *Proc. Conf. Human Factors in Computing Systems (CHI'89)*, Austin, TX, pp. 227–233.

[12] V. I. Pavlović, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human–computer interaction: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 677–695, 1997.

[13] A. Waibel, M. T. Vo, P. Duchnowski, and S. Manke, "Multimodal interfaces," *Artificial Intell. Rev.*, vol. 10, pp. 299–319, Aug. 1995.

[14] B. A. Myers, "A brief history of human computer interaction technology," Human Computer Interaction Institute, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, Tech. Rep. CMU-CS-TR-96-163, 1996.

[15] F. H. Raab, E. B. Blood, T. O. Steiner, and H. R. Jones, "Magnetic position and orientation tracking system," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 15, pp. 709–718, 1979.

[16] R. Azuma, "Tracking requirements for augmented reality," *Commun. ACM*, vol. 36, no. 7, pp. 50–52, 1993.

[17] T. Baudel and M. Baudouin-Lafon, "Charade: Remote control of objects using free-hand gestures," *Commun. ACM*, vol. 36, no. 7, pp. 28–35, 1993.

[18] S. S. Fels and G. E. Hinton, "Glove-Talk: A neural network interface between a Data-Glove and a speech synthesizer," *IEEE Trans. Neural Networks*, vol. 4, pp. 2–8, Jan. 1993.

[19] D. J. Sturman and D. Zeltzer, "A survey of glove-based input," *IEEE Comput. Graph. Applicat. Mag.*, vol. 14, pp. 30–39, Jan. 1994.

[20] D. L. Quam, "Gesture recognition with a DataGlove," in *Proc. 1990 IEEE National Aerospace and Electronics Conf.*, 1990, vol. 2.

[21] C. Wang and D. J. Cannon, "A virtual end-effector pointing system in point-and-direct robotics for inspection of surface flaws using a neural network based skeleton transform," in *Proc. IEEE Int. Conf. Robotics and Automation*, May 1993, vol. 3, pp. 784–789.
[22] B. H. Juang, "Speech recognition in adverse environments," *Comput. Speech Language*, vol. 5, pp. 275–294, 1991.
[23] D. Stork and H.-L. Lu, "Speechreading by Boltzmann zippers," in *Machines that Learn*. Snowbird, UT: 1996.
[24] R. Sharma, T. S. Huang, V. I. Pavlović, Y. Zhao, Z. Lo, S. Chu, K. Schulten, A. Dalke, J. Phillips, M. Zeller, and W. Humphrey, "Speech/gesture interface to a visual computing environment for molecular biologists," in *Proc. Int. Conf. Pattern Recognition*, Aug. 1996, pp. 964–968.
[25] F. Hatfield, E. A. Jenkins, M. W. Jennings, and G. Calhoun, "Principles and guidelines for the design of eye/voice interaction dialogs," in *Proc. 3rd Ann. Symp. Human Interaction with Complex Systems*, Dayton, OH, 1996, pp. 10–19.
[26] T. E. Hutchinson, "Computers that sense eye position on the display," *Computer*, vol. 26, pp. 65–67, July 1993.
[27] R. J. K. Jacob, "What you look at is what you get," *Computer*, vol. 26, pp. 65–67, July 1993.
[28] I. A. Essa and A. P. Pentland, "Coding analysis, interpretation, and recognition of facial expressions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 757–763, 1997.
[29] D. M. Gavrila and L. S. Davis, "Toward 3-D model-based tracking and recognition of human movement: A multi-view approach," in *Proc. IWAFGR'95*, Zurich, Switzerland, June 1995, pp. 272–277.
[30] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 780–785, 1997.
[31] C. Nodine, H. Kundel, L. Toto, and E. Krupinski, "Recording and analyzing eye-position data using a microcomputer workstation," *Behavior Res. Methods, Instruments Comput.*, vol. 24, no. 3, pp. 475–584, 1992.
[32] C. Lansing and G. W. McConkie, "A new method for speechreading research: Tracking observers' eye movements," *J. Acad. Rehabilitative Audiology*, vol. 28, pp. 25–43, 1994.
[33] V. I. Pavlović, R. Sharma, and T. S. Huang, "Gestural interface to a visual computing environment for molecular biologists," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Killington, VT, Oct. 1996, pp. 30–35.
[34] R. Sharma and J. Molineros, "Computer vision-based augmented reality for guiding manual assembly," *Presence: Teleoperators Virtual Environ.*, vol. 6, pp. 292–317, June 1997.
[35] F. K. H. Quek, "Eyes in the interface," *Image Vision Comput.*, vol. 13, Aug. 1995.
[36] R. Sharma, T. S. Huang, and V. I. Pavlović, "A multimodal framework for interacting with virtual environments," in *Human Interaction with Complex Systems*, C. A. Ntuen, E. H. Park, and J. H. Kim, Eds. Norwell, MA: Kluwer, 1996, pp. 53–71.
[37] R. M. Satava and S. B. Jones, "Virtual environments for medical training and education," *Presence: Teleoperators Virtual Environ.*, vol. 6, no. 2, pp. 139–146, 1997.
[38] S. L. Delp, P. Loan, C. Basdogan, and J. M. Rosen, "Surgical simulation: An emerging technology for training in emergency medicine," *Presence: Teleoperators Virtual Environ.*, vol. 6, no. 2, pp. 147–159, 1997.
[39] M. Bergamasco, "Haptic interfaces: The study of force and tactile feedback systems," in *Proc. IEEE Int. Workshop Robot on Robot and Human Communication*, 1995, pp. 15–20.
[40] T. R. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*. Cambridge, MA: MIT Press, 1992.
[41] Z. A. Keirn and J. I. Aunon, "Man-machine communications through brain-wave processing," *IEEE Eng. Med. Biology Mag.*, vol. 9, pp. 55–57, 1990.
[42] D. J. McFarland, G. W. Neat, R. F. Read, and J. R. Wolpaw, "An EEG-based method for graded cursor control," *Psychobiology*, vol. 21, pp. 77–81, 1993.
[43] V. T. Nasman, G. L. Calhoun, and G. R. McMillan, "Brain-actuated control and HMDS," in *Head Mounted Displays*, New York: McGraw-Hill, 1997, pp. 285–312.
[44] W. Putnam and R. B. Knapp, "Real-time computer control using pattern recognition of the electromyogram," in *Proceedings of the Fifteenth Annual International Conference on Engineering in Medicine and Biology Society*. New York: IEEE Press, 1993, vol. 15, pp. 1236–1237.
[45] H. S. Lusted and R. B. Knapp, "Controlling computers with neural signals," *Sci. Amer.*, pp. 82–87, Oct. 1996.
[46] T. Elbert, B. Rockstroh, W. Lutzenberger, and W. Birbaumer, *Self-Regulation of the Brain and Behavior*. New York: Springer-Verlag, 1984.
[47] S. Suryanarayanan and N. R. Reddy, "EMG-based interface for position tracking and control in VR environments and teleoperation," *Presence: Teleoperators Virtual Environ.*, vol. 6, no. 3, pp. 282–291, 1997.
[48] A. M. Junker, J. H. Schnurer, D. F. Ingle, and C. W. Downey, "Loop-closure of the visual cortex response," Armstrong Aerospace Medical Research Laboratory, Wright–Patterson Air Force Base, Dayton, OH, Tech. Rep. AAMRL-TR-88-014, 1988.
[49] D. W. Patmore and R. B. Knapp, "A cursor controller using evoked potentials and EOG," in *Proc. RESNA 18th Ann. Conf.*, 1995, pp. 702–704.
[50] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought*. Chicago, IL: Univ. of Chicago Press, 1992.
[51] J. Streeck, "Gesture as communication I: Its coordination with gaze and speech," *Commun. Monographs*, vol. 60, pp. 275–299, Dec. 1993.
[52] B. V. Dasarathy, "Sensor fusion potential exploitation—Innovative architectures and illustrative approaches," *Proc. IEEE*, vol. 85, pp. 24–38, Jan. 1997.
[53] B. Stein and M. A. Meredith, *The Merging of Senses*. Cambridge, MA: MIT Press, 1993.
[54] T. G. R. Bower, "The evolution of sensory system," in *Perception: Essays in Honor of James J. Gibson*, R. B. MacLeod and H. L. Pick, Jr., Eds. Ithaca, NY: Cornell Univ. Press, 1974, pp. 141–153.
[55] D. Heckerman, "A tutorial on learning with Bayesian networks," Microsoft Corp., Seattle, WA, Tech. Rep. MSR-TR-95-06, Mar. 1995.
[56] P. K. Varshney, *Distributed Detection and Data Fusion*. New York: Springer-Verlag, 1996.
[57] R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Kalman Filtering*. New York: Wiley, 1992.
[58] C. K. Chui and G. Chen, *Kalman Filtering with Real-Time Applications*. Berlin/Heidelberg, Germany: Springer-Verlag, 1991.
[59] K. C. Chou, A. S. Willsky, and A. Benveniste, "Multiscale recursive estimation, data fusion, and regularization," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 464–478, Mar. 1994.
[60] M. I. Jordan and C. M. Bishop, "Neural networks," in *CRC Handbook of Computer Science* A. Tucker, Ed. Boca Raton, FL: CRC Press, 1996.
[61] H. A. Boulard and N. Morgan, *Connectionist Speech Recognition. A Hybrid Approach*. Norwell, MA: Kluwer, 1994.
[62] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *IEEE Computer Mag.*, pp. 31–44, Mar. 1996.
[63] P. J. Werbos, "Generalization of back propagation with applications to a recurrent gas market model," *Neural Networks*, vol. 1, pp. 339–356, 1988.
[64] P. Haffner, M. Franzini, and A. Weibel, "Integrating time alignment and neural networks for high performance continuous speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Toronto, Ont., Canada, Apr. 1991, pp. 105–108.
[65] U. Meier, W. Hürst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1996.
[66] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.
[67] T. E. Starner and A. Pentland, "Visual recognition of American sign language using hidden Markov models," in *Proc. Int. Workshop Automatic Face and Gesture Recognition*, Zurich, Switzerland, June 1995, pp. 189–194.
[68] L. W. Campbell, D. A. Becker, A. Azarbayejani, A. F. Bobick, and A. Pentland, "Invariant features for 3-D gesture recognition," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Killington, VT, Oct. 1996, pp. 157–162.
[69] D. X. Sun and L. Deng, "Non-stationary hidden Markov models for speech recognition," in *Image Models (and Their Speech Model Cousins)*, S. E. Levinson and L. Shepp, Eds. New York: Springer-Verlag, 1996, pp. 161–182.

[70] A. Adjoudani and C. Benoit, "Audio-visual speech recognition compared across two architectures," in *Proc. Eurospeech'95 Conf.*, Madrid, Spain, 1995, vol. 2, pp. 1563–1566.
[71] V. Pavlović, G. Berry, and T. S. Huang, "Integration of audio/visual information for use in human–computer intelligent interaction," in *Proc. IEEE Int. Conf. Image Processing*, Santa Barbara, CA, 1997.
[72] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Proc. Speechreading by Man and Machine: Models, Systems and Applications Workshop*, Aug./Sept. 1995.
[73] M. Minsky, "A framework for representing knowledge," in *The Psychology of Computer Vision*, P. H. Winston, Ed. New York: McGraw-Hill, 1975.
[74] F. Lehman, Ed., *Semantic Networks in Artificial Intelligence.* Oxford, England: Pergamon, 1992.
[75] J. Guan and D. A. Bell, *Evidence Theory and its Applications*, vol. 1. Amsterdam, The Netherlands: North-Holland, 1991.
[76] R. A. Bolt, "Put that there: Voice and gesture at the graphics interface," *ACM Comput. Graph.*, vol. 14, no. 3, pp. 262–270, 1980.
[77] M. T. Vo and C. Wood, "Building an application framework for speech and pen input integration in multimodal learning interfaces," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1996, pp. 3545–3548.
[78] R. Kober, U. Harz, and J. Schiffers, "Fusion of visual and acoustic signals for command-word recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1997.
[79] J. M. Bradshaw, "An introduction to software agents," in *Software Agents*, J. M. Bradshaw, Ed. Cambridge, MA: AAAI Press/MIT Press, 1997.
[80] Y. Shoham, "An overview of agent-oriented programming," in *Software Agents*, J. M. Bradshaw, Ed. AAAI Press/MIT Press, 1997.
[81] N. Negroponte, "Agents: From direct manipulation to delegation," in *Software Agents*, J. M. Bradshaw, Ed. AAAI Press/MIT Press, 1997.
[82] D. B. Moran, A. J. Cheyer, L. E. Julia, D. L. Martin, and S. Park, "Multimodal user interface in the open agent architecture," in *Proc. ACM Int. Conf. Intelligent User Interfaces*, Orlando, FL, 1997, pp. 61–68.
[83] J. A. Pittman, I. Smith, P. Cohen, S. Oviatt, and T.-C. Yang, "QuickSet: A multimodal interface for military simulation," in *Proc. 6th Conf. Computer-Generated Forces and Behavioral Representation*, Orlando, FL, 1996, pp. 217–224.
[84] M. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L. Kalé, R. Skeel, K. Schulten, and R. Kufrin, "MDScope—A visual computing environment for structural biology," *Comput. Phys. Commun.*, vol. 91, no. 1/2/3, pp. 111–134, 1995.
[85] P. R. Cohen, M. Johnston, D. McGee, S. Oviatt, and J. Pittman, "QuickSet: Multimodal interaction for simulation set-up and control," in *Proc. 5th Applied Natural Language Processing Meeting*, Washington, DC, 1997.
[86] B. Suhm, P. Geunter, T. Kemp, A. Lavie, L. Mayfield, A. McNair, I. Rogina, T. Schultz, T. Sloboda, W. Ward, M. Woszczyna, and A. Waibel, "Janus: Toward multilingual spoken language translation," in *Proc. ARPA SLT Workshop*, Austin, TX, 1995.
[87] J. Wang, "Integration of eye-gaze, voice and manual response in multimodal user interface," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, 1995, pp. 3938–3942.
[88] M. Fukumoto, Y. Suenaga, and K. Mase, "'Finger-pointer': Pointing interface by image processing," *Comput. Graph.*, vol. 18, no. 5, pp. 633–642, 1994.
[89] C. Codella, R. Jalili, L. Koved, *et al.*, "Interactive simulation in a multi-person virtual world," in *ACM Conf. Human Factors in Computing Systems (CHI'92)*, pp. 329–334.
[90] P. Maes, T. Darrell, B. Blumberg, and A. Pentland, "The ALIVE system: Wireless, full-body interaction with autonomous agents," *ACM Multimedia Syst.*, 1996.
[91] A. Pentland, "Smart rooms," *Sci. Amer.*, pp. 54–62, Apr. 1996.
[92] MIT Media Laboratory, Perceptual Intelligence Group. [Online]. Available: http://casr.www.media.mit.edu/groups/casr/pentland.html.
[93] N. Tosa. (1996). Neuro-Baby. [Online]. Available: http://www.mic.atr.co.jp/tosa.
[94] M. Kakimoto, N. Tosa, J. Mori, and A. Sanada, "Tool of Neuro-Baby," *Inst. Television Eng. Jpn. Tech. Rep.*, vol. 16, pp. 7–12, June 1992.

Rajeev Sharma (Member, IEEE) received the Ph.D. degree in computer science from the University of Maryland, College Park, in 1993.

For three years, he was with the University of Illinois at Urbana-Champaign as a Beckman Fellow and Adjunct Assistant Professor in the Department of Electrical and Computer Engineering. He currently is an Assistant Professor in the Department of Computer Science and Engineering, Pennsylvania State University, University Park. His research interests lie in studying the role of computer vision in robotics and advanced human–computer interfaces.

Dr. Sharma received the ACM Samuel Alexander Doctoral Dissertation Award, an IBM Pre-Doctoral Fellowship, and an NSF CAREER award.

Vladimir I. Pavlović (Student Member, IEEE) was born in Paris, France, in 1966. He received the Dipl.Eng. degree in electrical engineering from the University of Novi Sad, Yugoslavia, in 1991 and the M.S. degree in electrical engineering and computer science from the University of Illinois at Chicago in 1993. He currently is pursuing the Ph.D. degree in electrical engineering at the Beckman Institute and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign.

His research interests include vision-based human–computer interaction, multimodal signal fusion, and image coding.

Thomas S. Huang (Fellow, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., and the M.S. and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

From 1963 to 1973, he was a Member of the Faculty of the Department of Electrical Engineering at MIT. From 1973 to 1980, he was a Member of the Faculty of the School of Electrical Engineering and Director of the Laboratory for Information and Signal Processing at Purdue University, West Lafayette, IN. In 1980, he joined the University of Illinois at Urbana-Champaign, where he is now a William L. Everitt Distinguished Professor of Electrical and Computer Engineering, a Research Professor at the Coordinated Science Laboratory, and Head of the Image Formation and Processing Group at the Beckman Institute for Advanced Science and Technology. During sabbatical leaves, he has worked at the MIT Lincoln Laboratory, the IBM T. J. Watson Research Center, and the Rheinishes Landes Museum in Bonn, West Germany. He was a Visiting Professor with the Swiss Institutes of Technology in Zurich and Lausanne, the University of Hannover in West Germany, INRS-Telecommunications of the University of Quebec, Montreal, Canada, and the University of Tokyo, Japan. He has been a Consultant to numerous industrial firms and government agencies both in the United States and abroad. His professional interests lie in the broad area of information technology, especially the transmission of processing of multidimensional signals. He has published 12 books and more than 300 papers on network theory, digital filtering, image processing, and computer vision. He is a Founding Editor of the *International Journal Computer Vision, Graphics, and Image Processing* and Editor of the Springer Series in Information Sciences published by Springer-Verlag.

Dr. Huang is a Fellow of the International Association of Pattern Recognition and the Optical Society of America. He has received a Guggenheim Fellowship, an A. V. Humboldt Foundation Senior U.S. Scientist Award, and a Fellowship from the Japan Association for the Promotion of Science. He received the IEEE Acoustics, Speech, and Signal Processing Society's Technical Achievement Award in 1987 and the Society Award in 1991.

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].

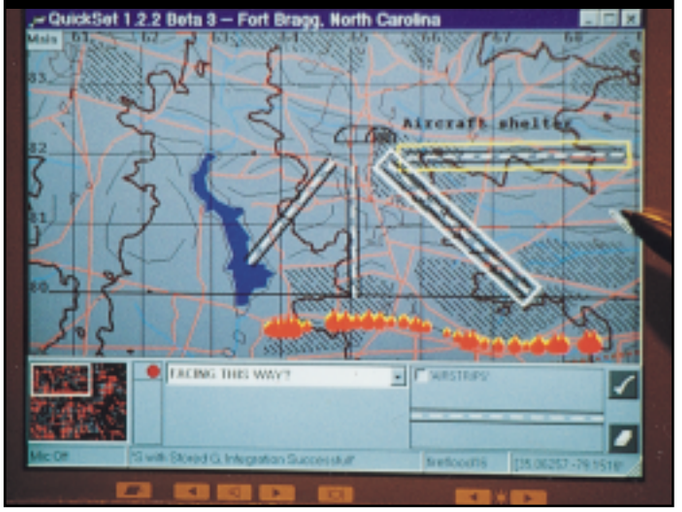
Predicting whether a user will express a command multimodally also depends on the type of action they are performing. In particular, they almost always express commands multimodally when describing spatial information about the location, number, size, orientation, or shape of an object. In the data summarized 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].

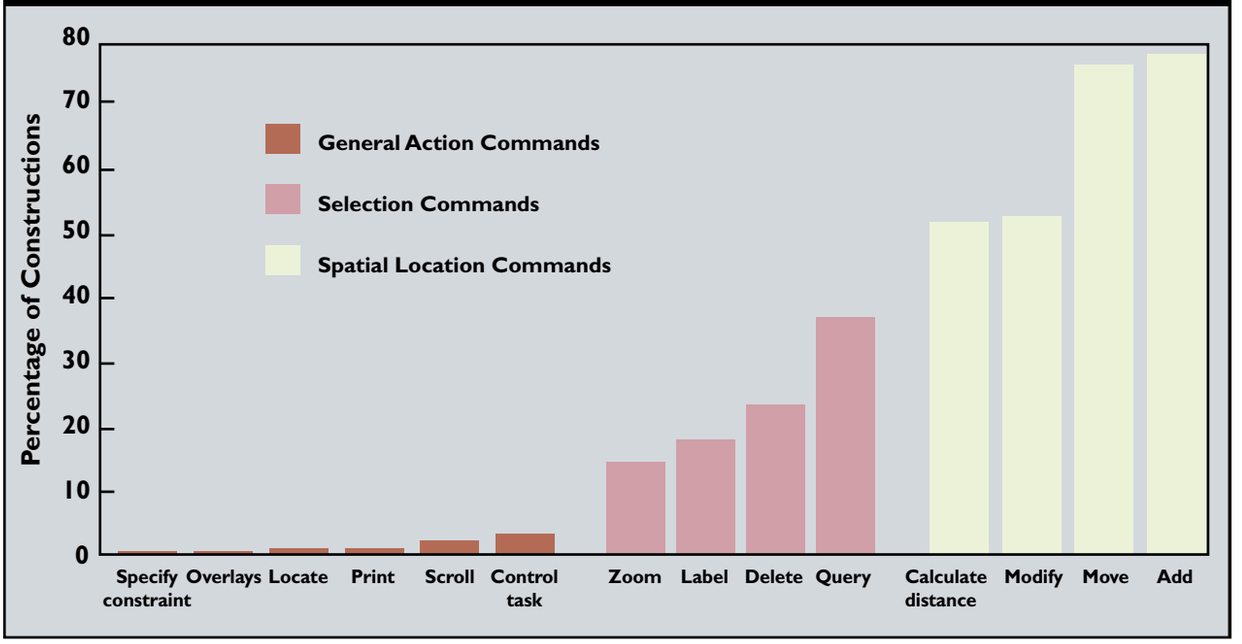
However, this concept of multimodal interaction as point-and-speak makes only limited use of new input modes for *selection* of objects—just as the mouse does. In this respect, it represents the persistence of an old mouse-oriented metaphor. In contrast, modes that transmit written input, manual gesturing, 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).

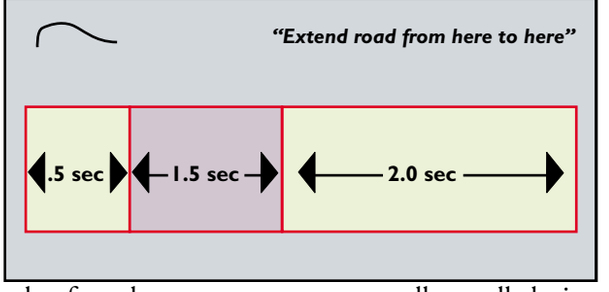


co-occur temporally. This temporal overlap then determines which signals to combine during system processing. For example, successful processing of the deictic term "that square" would rely on interpretation of pointing when the word "that" is spoken in order to extract the intended referent. However, 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.

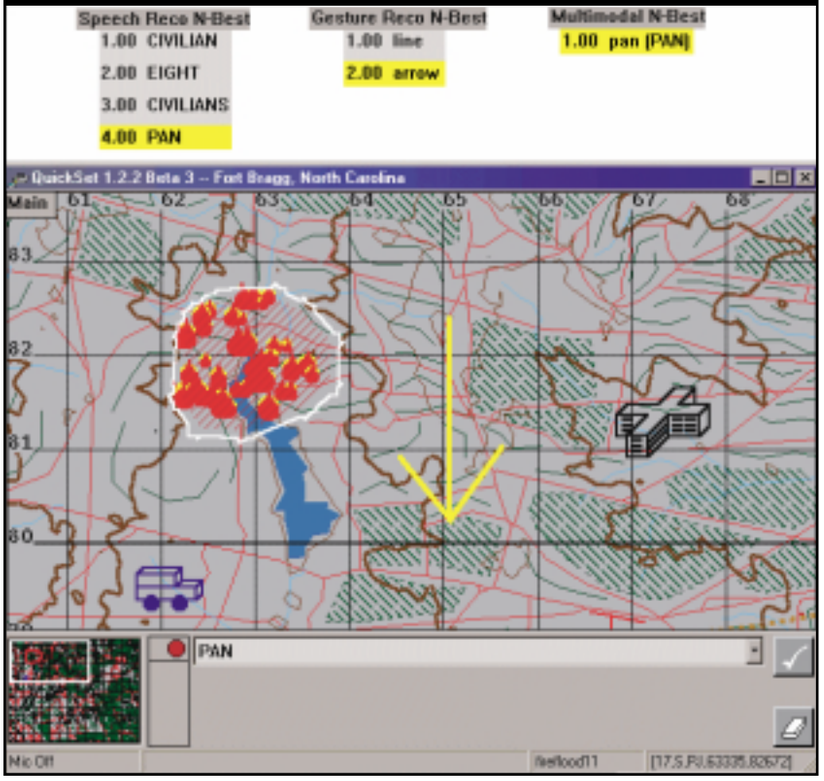
Due to mutual disambiguation, the parallel recognition and semantic interpretation that occurs in a multimodal architecture can yield a higher likelihood of correct interpretation than recognition based on either single input mode. This improvement is a direct result of the disambiguation between signals 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.