◆ 21 ◆

MULTIMODAL INTERFACES

Sharon Oviatt
Oregon Health & Science University

WHAT ARE MULTIMODAL SYSTEMS, AND WHY ARE WE BUILDING THEM?

Multimodal systems process two or more combined user input modes—such as speech, pen, touch, manual gestures, gaze, and head and body movements—in a coordinated manner with multimedia system output. This class of systems represents a new direction for computing, and a paradigm shift away from conventional WIMP interfaces. Since the appearance of Bolt's (1980) "Put That There" demonstration system, which processed speech in parallel with touch-pad pointing, a variety of new multimodal systems has emerged. This new class of interfaces aims to recognize naturally occurring forms of human language and behavior, which incorporate at least one recognition-based technology (e.g., speech, pen, vision). The development of novel multimodal systems has been enabled by the myriad input and output technologies currently becoming available, including new devices and improvements in recognition-based technologies. This chapter will review the main types of multimodal interfaces, their advantages and cognitive science underpinnings, primary features and architectural characteristics, and general research in the field of multimodal interaction and interface design.

The growing interest in multimodal interface design is inspired largely by the goal of supporting more transparent, flexible, efficient, and powerfully expressive means of human-computer interaction. Multimodal interfaces are expected to be easier to learn and use, and are preferred by users for many applications. They have the potential to expand computing to more challenging applications, to be used by a broader spectrum of everyday people, and to accommodate more adverse usage conditions than in the past. Such systems also have the potential to function in a more robust and stable manner than unimodal recognition systems involving a single recognition-based technology, such as speech, pen, or vision.

The advent of multimodal interfaces based on recognition of human speech, gazes, gestures, and other natural behaviors represents only the beginning of a progression toward computational interfaces capable of relatively humanlike sensory perception. Such interfaces eventually will interpret continuous input from a large number of different visual, auditory, and tactile input modes, which will be recognized as users engage in everyday activities. The same system will track and incorporate information from multiple sensors on the user's interface and surrounding physical environment in order to support intelligent adaptation to the user, task, and usage environment. Future adaptive multimodal-multisensor interfaces have the potential to support new functionality, to achieve unparalleled robustness, and to perform flexibly as a multifunctional and personalized mobile system.

WHAT TYPES OF MULTIMODAL INTERFACES EXIST, AND WHAT IS THEIR HISTORY AND CURRENT STATUS?

Multimodal systems have developed rapidly during the past decade, with steady progress toward building more general and robust systems, as well as more transparent human interfaces than ever before (Benoit, Martin, Pelachaud, Schomaker, & Suhm, 2000; Oviatt et al., 2000). Major developments have occurred in the hardware and software needed to support key component technologies incorporated within multimodal systems, as well as in techniques for integrating parallel input streams. Multimodal systems also have diversified to include new modality combinations, including speech and pen input, speech and lip movements, speech and manual gesturing, and gaze tracking and manual input (Benoit & Le Goff, 1998; Cohen et al., 1997; Stork & Hennecke, 1995; Turk & Robertson, 2000; Zhai, Morimoto, & Ihde, 1999). In addition, the array of multimodal applications has expanded extremely rapidly in recent years. Among other areas, it presently includes multimodal map-based systems for mobile and in-vehicle use; multimodal browsers; multimodal interfaces to virtual reality systems for simulation and training; multimodal person-identification/verification systems for security purposes; multimodal medical, educational, military, and web-based transaction systems; and multimodal access and management of personal information on handhelds and cell phones (Cohen & McGee, 2004; Iyengar, Nock, & Neti, 2003; McGee, 2003; Neti, Iyengar, Potamianos, & Senior, 2000; Oviatt, 2003; Oviatt, Flickner, & Darrell, 2004; Oviatt & Lunsford, 2005; Oviatt et al., 2000; Pankanti, Bolle, & Jain, 2000; Reithinger et al., 2003).

In one of the earliest multimodal concept demonstrations, Bolt had users sit in front of a projection of "Dataland" in "the Media Room" (Negroponte, 1978). Using the "Put That There" interface (Bolt, 1980), they could use speech and pointing on an armrest-mounted touchpad to create and move objects on a 2D large-screen display. For example, the user could issue a command, such as "Create a blue square there," with the intended location of "there" indicated by a 2D cursor mark on the screen. Semantic processing was based on the user's spoken input, and the meaning of the deictic "there" was resolved by processing the x/y coordinate indicated by the cursor at the time "there" was uttered. Since Bolt's early prototype, considerable strides have been made in developing a wide variety of different types of multimodal systems.

Among the earliest and most rudimentary multimodal systems were ones that supported speech input along with a standard keyboard and mouse interface. Conceptually, these multimodal interfaces represented the least departure from traditional graphical user interfaces (GUIs). Their initial focus was on providing richer natural language processing to support greater expressive power for the user when manipulating complex visuals and engaging in information extraction. As speech recognition technology matured during the late 1980s and 1990s, these systems added spoken input as an alternative to text entry via the keyboard. As such, they represent early involvement of the natural language and speech communities in developing the technologies needed to support new multimodal interfaces. Among the many examples of this type of multimodal interface are CUBRICON, Georal, Galaxy, XTRA, Shoptalk, and Miltalk (Cohen et al., 1989; Kobsa et al., 1986; Neal & Shapiro, 1991; Seneff, Goddeau, Pao, & Polifroni, 1996; Siroux, Guyomard, Multon, & Remondeau, 1995; Wahlster, 1991).

Several of these early systems were multimodal-multimedia map systems to which a user could speak or type and point with a mouse to extract tourist information or engage in military situation assessment (Cohen et al., 1989; Neal & Shapiro, 1991; Seneff et al., 1996; Siroux et al., 1995). For example, using the CUBRICON system a user could point to an object on a map and ask, "Is this <point> an air base?" CUBRICON was an expert system with extensive domain knowledge, as well as natural language processing capabilities that included referent identification and dialogue tracking (Neal & Shapiro, 1991). With the Georal system, a user could query a tourist-information system to plan travel routes using spoken input and pointing via a touch-sensitive screen (Siroux et al., 1995). In contrast, the Shoptalk system permitted users to interact with complex graphics representing factory production flow for chip manufacturing (Cohen et al., 1989). Using Shoptalk, a user could point to a specific machine in the production layout and issue the command: "Show me all the times when this machine was down." After the system delivered its answer as a list of time ranges, the user could click on one to ask the follow-up question: "What chips were waiting in its queue then, and were any of them hot lots?" Multimedia system feedback was available in the form of a text answer, or the user could click on the machine in question to view an exploded diagram of the machine queue's contents during that time interval.

More recent multimodal systems have moved away from processing simple mouse or touchpad pointing, and have begun designing systems based on two parallel input streams that each are capable of conveying rich semantic information. These multimodal systems recognize two natural forms of human language and behavior, for which two recognition-based technologies are incorporated within a more powerful bimodal user interface. To date, systems that combine either speech and pen input (Oviatt & Cohen, 2000) or speech and lip movements (Benoit et al., 2000; Stork & Hennecke, 1995; Rubin, Vatikiotis-Bateson, & Benoit, 1998; Potamianos, Neti, Gravier, & Garg, 2003) constitute the two most mature areas within the field of multimodal research. In these cases, the keyboard and mouse have been abandoned. For speech and pen systems, spoken language sometimes is processed along with complex pen-based gestural input involving hundreds of different symbolic interpretations beyond pointing[1] (Oviatt et al., 2000). For speech and lip movement systems, spoken language is processed along with corresponding human lip movements during the natural audio-visual experience of spoken interaction. In both of these sub-literatures, considerable work has been directed toward quantitative modeling of the integration and synchronization characteristics of the two rich input modes being processed, and innovative time-sensitive architectures have been developed to process these patterns in a robust manner.

Multimodal systems that recognize speech and pen-based gestures first were designed and studied in the early 1990s (Oviatt, Cohen, Fong, & Frank, 1992), with the original QuickSet system prototype built in 1994. The QuickSet system is an agent-based collaborative multimodal system that runs on a hand-held PC (Cohen et al., 1997). With QuickSet, for example, a user can issue a multimodal command such as "Airstrips . . . facing this way <draws arrow>, and facing this way <draws arrow>," using combined speech and pen input to place the correct number, length, and orientation (e.g., SW, NE) of aircraft landing strips on a map. Other research-level systems of this type were built in the late 1990s. Examples include the Human-centric Word Processor, Portable Voice Assistant, QuickDoc and MVIEWS (Bers, Miller, & Makhoul, 1998; Cheyer, 1998; Oviatt et al., 2000; Waibel, Suhm, Vo, & Yang, 1997). These systems represent a variety of different system features, applications, information fusion, and linguistic processing techniques. For illustration purposes, a comparison of five different speech and gesture systems is summarized in Fig. 21.1. In most cases, these multimodal systems jointly interpreted speech and pen input based on a frame-based method of information fusion and a late semantic fusion approach, although QuickSet used a statistically-ranked unification process and a hybrid symbolic/statistical architecture (Wu, Oviatt, & Cohen, 1999). Other recent systems also have adopted unification-based multimodal fusion and a hybrid architectural approach for processing multimodal input (Bangalore & Johnston, 2000; Denecke & Yang, 2000; Pfleger, 2004; Wahlster, 2001) and even multimodal output (Kopp, Tepper, & Cassell, 2004).

Multimodal systems that process speech and continuous 3D manual gesturing are emerging rapidly, although these systems remain less mature than ones that process 2D pen input (Encarnacao & Hettinger, 2003; Flanagan & Huang, 2003; Sharma, Pavlovic, & Huang, 1998; Pavlovic, Sharma, & Huang, 1997). This primarily is because of the significant challenges associated with segmenting and interpreting continuous manual movements, compared with a stream of *x/y* ink coordinates. Because of this difference, multimodal speech and pen systems have advanced more rapidly in their architectures, and have progressed further toward commercialization of applications. However, a significant cognitive science literature is available for guiding the design of emerging speech and 3D-gesture prototypes (Condon, 1988; Kendon, 1980; McNeill, 1992), which will be discussed further later in this chapter. Among the earlier systems to begin processing manual pointing or 3D gestures combined with speech were developed by Koons and colleagues (Koons, Sparrell, & Thorisson, 1993), Sharma and colleagues (Sharma et al., 1996), Poddar and colleagues (Poddar, Sethi, Ozyildiz, & Sharma, 1998), and by Duncan and colleagues (Duncan, Brown, Esposito, Holmback, & Xue, 1999).

Historically, multimodal speech and lip movement research has been driven by cognitive science interest in intersensory audio-visual perception and the coordination of speech output with lip and facial movements (Benoit & Le Goff, 1998; Bernstein & Benoit, 1996; Cohen & Massaro, 1993; Massaro & Stork, 1998; McGrath & Summerfield, 1985; McGurk & MacDonald, 1976; McLeod & Summerfield, 1987; Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998; Sumby & Pollack, 1954; Summerfield, 1992; Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzopoulos, 1996). Among the many contributions of this literature has been a detailed classification of human lip movements (visemes) and the viseme-phoneme mappings that occur during articulated speech. Existing systems that have processed combined

---

[1] However, other recent pen/voice multimodal systems that emphasize mobile processing, such as MiPad and the Field Medic Information System (Holzman, 1999; Huang et al., 2000), still limit pen input to pointing.

Multimodal System Characteristics:	**QuickSet**	**Human-Centric Word Processor**	**VR Aircraft Maintenance Training**	**Field Medic Information**	**Portable Voice Assistant**
Recognition of simultaneous or alternative individual modes	Simultaneous & individual modes	Simultaneous & individual modes	Simultaneous & individual modes	Alternative individual modes[1]	Simultaneous & individual modes
Type & size of gesture vocabulary	Pen input, Multiple gestures, Large vocabulary	Pen input, Deictic selection	3D manual input, Multiple gestures, Small vocabulary	Pen input, Deictic selection	Pen input Deictic selection[2]
Size of speech vocabulary[3] & type of linguistic processing	Moderate vocabulary, Grammar-based	Large vocabulary, Statistical language processing	Small vocabulary, Grammar-based	Moderate vocabulary, Grammar-based	Small vocabulary, Grammar-based
Type of signal fusion	Late semantic fusion, Unification, Hybrid symbolic/statistical MTC framework	Late semantic fusion, Frame-based	Late semantic fusion, Frame-based	No mode fusion	Late semantic fusion, Frame-based
Type of platform & applications	Wireless handheld, Varied map & VR applications, digital paper	Desktop computer, Word processing	Virtual reality system, Aircraft maintenance training	Wireless handheld, Medical field emergencies	Wireless handheld, Catalogue ordering
Evaluation status	Proactive user-centered design & iterative system evaluations	Proactive user-centered design	Planned for future	Proactive user-centered design & iterative system evaluations	Planned for future

[1] The FMA component recognizes speech only, and the FMC component recognizes gestural selections or speech. The FMC also can transmit digital speech and ink data, and can read data from smart cards and physiological monitors.

[2] The PVA also performs handwriting recognition.

[3] A small speech vocabulary is up to 200 words, moderate 300–1,000 words, and large in excess of 1,000 words. For pen-based gestures, deictic selection is an individual gesture, a small vocabulary is 2–20 gestures, moderate 20–100, and large in excess of 100 gestures.

FIGURE 21.1. Examples of functionality, architectural features, and general classification of different speech and gesture multimodal applications.

speech and lip movements include the classic work by Petajan (1984), Brooke and Petajan (1986), and others (Adjoudani & Benoit, 1995; Bregler & Konig, 1994; Silsbee & Su, 1996; Tomlinson, Russell, & Brooke, 1996). Additional examples of speech and lip movement systems, applications, and relevant cognitive science research have been detailed elsewhere (Benoit et al., 2000). Researchers in this area have been actively exploring adaptive techniques for improving system robustness, especially in noisy environmental contexts (Dupont & Luettin, 2000; Meier, Hürst, & Duchnowski, 1996; Potamianos, Neti, Gravier, & Garg, 2003; Rogozan & Deglise, 1998), which is an important future research direction. Although this literature has not emphasized the development of applications, nonetheless its quantitative modeling of synchronized phoneme/viseme patterns has been used to build animated characters that generate text-to-speech output and coordinated lip movements. These new animated characters are being used as an interface design vehicle for facilitating users' multimodal interaction with next-generation conversational interfaces (Cassell, Sullivan, Prevost, & Churchill, 2000; Cohen & Massaro, 1993).

While the main multimodal literatures to date have focused on either speech and pen input or speech and lip movements, recognition of other modes also is maturing and beginning to be integrated into new kinds of multimodal systems. In particular, there is growing interest in designing multimodal interfaces that incorporate vision-based technologies, such as interpretation of gaze, facial expressions, head nodding, gesturing, and large body movements (Flanagan & Huang, 2003; Morency, Sidner, Lee, & Darrell, 2005; Morimoto, Koons, Amir, Flickner, & Zhai, 1999; Pavlovic, Berry, & Huang, 1997; Turk & Robertson, 2000; Zhai et al., 1999). These technologies unobtrusively or *passively* monitor user behavior and need not require explicit user commands to a "computer." That contrasts with *active input modes*, such as speech or pen, which the user deploys intentionally as a command issued to the system (see Fig. 21.2). While passive modes may be "attentive" and less obtrusive, active modes generally are more reliable indicators of user intent.

As vision-based technologies mature, one important future direction will be the development of *blended* multimodal interfaces that combine both passive and active modes. These interfaces typically will be *temporally cascaded*, so one goal in designing new prototypes will be to determine optimal processing strategies for using advance information from the first mode (e.g., gaze) to constrain accurate interpretation of the following modes (e.g., gesture, speech). This kind of blended multimodal interface potentially can provide users with greater transparency and control, while also supporting improved robustness and broader application functionality (Oviatt & Cohen, 2000; Zhai et al., 1999). As this collection of technologies matures, there also is strong interest in designing new types of pervasive and mobile interfaces, including ones capable of adaptive processing to the user and environmental context.

As multimodal interfaces gradually evolve toward supporting more advanced recognition of users' natural activities in context, they will expand beyond rudimentary bimodal systems to ones that incorporate three or more input modes, qualitatively different modes, and more sophisticated models of multimodal interaction. This trend already has been initiated within biometrics research, which has combined recognition of multiple behavioral input modes (e.g., voice, handwriting) with physiological ones (e.g., retinal scans, fingerprints) to achieve reliable person identification and verification in challenging field conditions (Choudhury, Clarkson, Jebara, & Pentland, 1999; Jain et al., 1999; Jain & Ross, 2002; Pankanti et al., 2000).

Apart from these developments within research-level systems, multimodal interfaces also are being commercialized as products, especially in areas like personal information access and management on handhelds and cell phones. Microsoft's handheld Mipad for personal information management, and Kirusa's cell phone interface for directory assistance, messaging, and so on, are just two examples of the many mobile commercial products that are being developed with multimodal interfaces. Both include spoken language processing and a stylus for tapping on fields to constrain and guide the natural language processing. In some cases, keyboard input is supported as a third option, as well as multimedia output in the form of visualizations and text-to-speech. Another visible growth area for multimodal interfaces involves in-vehicle control of navigation, communication, and entertainment systems, which has emerged in both domestic and import cars. Mobile map-based systems and systems for safety-critical medical and military applications also are being commercialized by companies like Natural Interaction Systems (e.g., Rasa, NISMap, and NISChart applications), which places an emphasis on developing tangible multimodal interfaces that preserve users' existing work practice, minimize cognitive load, and provide backups in case of system failure (Cohen & McGee, 2004; McGee, 2003).

WHAT ARE THE GOALS AND ADVANTAGES OF MULTIMODAL INTERFACE DESIGN?

Over the past decade, numerous advantages of multimodal interface design have been documented. Unlike a traditional keyboard-and-mouse interface or a unimodal recognition-based interface, multimodal interfaces permit flexible use of input modes. This includes the choice of which modality to use for conveying different types of information, to use combined input modes, or to alternate between modes at any time. Since individual input modalities are well suited in some situations, and less ideal or even inappropriate in others, modality choice is an important design issue in a multimodal system. As systems become more complex and multifunctional, a single modality simply does not permit all users to interact effectively across all tasks and environments.

Since there are large individual differences in ability and preference to use different modes of communication, a multimodal interface permits diverse user groups to exercise selection and control over how they interact with the computer (Fell et al., 1994; Karshmer & Blattner, 1998). In this respect, multimodal interfaces have the potential to accommodate a broader range of users than traditional interfaces, including users of different ages, skill levels, native language status, cognitive styles, sensory impairments, and other temporary illnesses or permanent handicaps. For example, a visually impaired user or one with repetitive stress injury may prefer speech input and text-to-speech output. In contrast, a user with a hearing impairment

Multimodal interfaces process two or more combined user input modes— such as speech, pen, touch, manual gestures, gaze, and head and body movements— in a coordinated manner with multimedia system output. They are a new class of interfaces that aim to recognize naturally occurring forms of human language and behavior, and which incorporate one or more recognition-based technologies (e.g., speech, pen, vision).

Active input modes are ones that are deployed by the user intentionally as an explicit command to a computer system (e.g., speech).

Passive input modes refer to naturally occurring user behavior or actions that are recognized by a computer (e.g., facial expressions, manual gestures). They involve user input that is unobtrusively and passively monitored, without requiring any explicit command to a computer.

Blended multimodal interfaces are ones that incorporate system recognition of at least one passive and one active input mode. (e.g., speech and lip movement systems).

Temporally-cascaded multimodal interfaces are ones that process two or more user modalities that tend to be sequenced in a particular temporal order (e.g., gaze, gesture, speech), such that partial information supplied by recognition of an earlier mode (e.g., gaze) is available to constrain interpretation of a later mode (e.g., speech). Such interfaces may combine only active input modes, only passive ones, or they may be blended.

Mutual disambiguation involves disambiguation of signal or semantic-level information in one error-prone input mode from partial information supplied by another. Mutual disambiguation can occur in a multimodal architecture with two or more semantically rich recognition-based input modes. It leads to recovery from unimodal recognition errors within a multimodal architecture, with the net effect of suppressing errors experienced by the user.

Simultaneous integrator refers to a user who habitually presents two input signals (e.g., speech, pen) in a temporally overlapped manner when communicating multimodal commands to a system

Sequential integrator refers to a user who habitually separates their multimodal signals, presenting one before the other with a brief pause intervening

Multimodal hypertiming refers to the fact that both sequential and simultaneous integrators will further accentuate their basic multimodal integration pattern when under duress (e.g., as task difficulty or system recognition errors increase)

Visemes refers to the detailed classification of visible lip movements that correspond with consonants and vowels during articulated speech. A *viseme-phoneme mapping* refers to the correspondence between visible lip movements and audible phonemes during continuous speech.

Feature-level fusion is a method for fusing low-level feature information from parallel input signals within a multimodal architecture, which has been applied to processing closely synchronized input such as speech and lip movements.

Semantic-level fusion is a method for integrating semantic information derived from parallel input modes in a multimodal architecture, which has been used for processing speech and gesture input.

FIGURE 21.2. Multimodal interface terminology.

or accented speech may prefer touch, gesture, or pen input. The natural alternation between modes that is permitted by a multimodal interface also can be effective in preventing overuse and physical damage to any single modality, especially during extended periods of computer use (Markinson[2], personal communication, 1993).

Multimodal interfaces also provide the adaptability that is needed to accommodate the continuously changing conditions of mobile use. In particular, systems involving speech, pen, or touch input are suitable for mobile tasks and, when combined, users can shift among these modalities from moment to moment as environmental conditions change (Holzman, 1999; Oviatt, 2000b, 2000c). There is a sense in which mobility can induce a state of temporary disability, such that a person is unable to use a particular input mode for some period. For example, the user of an in-vehicle application may frequently be unable to use manual or gaze input, although speech is relatively more available. In this respect, a multimodal interface permits the modality choice and switching that is needed during the changing environmental circumstances of actual field and mobile use.

A large body of data documents that multimodal interfaces satisfy higher levels of user preference when interacting with

---

[2] R. Markinson, University of California at San Francisco Medical School, 1993.

simulated or real computer systems. Users have a strong preference toward interacting multimodally, rather than unimodally, across a wide variety of different application domains, although this preference is most pronounced in spatial domains (Hauptmann, 1989; Oviatt, 1997). For example, 95% to 100% of users preferred to interact multimodally when they were free to use either speech or pen input in a map-based spatial domain (Oviatt, 1997). During pen/voice multimodal interaction, users preferred speech input for describing objects and events, sets and subsets of objects, out-of-view objects, conjoined information, and past and future temporal states, and for issuing commands for actions or iterative actions (Cohen & Oviatt, 1995; Oviatt & Cohen, 1991). However, their preference for pen input increased when conveying digits, symbols, graphic content, and especially when conveying the location and form of spatially oriented information on a dense graphic display such as a map (Oviatt & Olsen, 1994; Oviatt, 1997; Suhm, 1998). Likewise, 71% of users combined speech and manual gestures multimodally, rather than using one input mode, when manipulating graphic objects on a CRT screen (Hauptmann, 1989).

During the early design of multimodal systems, it was assumed that efficiency gains would be the main advantage of designing an interface multimodally, and that this advantage would derive from the ability to process input modes in parallel. It is true that multimodal interfaces sometimes support improved efficiency, especially when manipulating graphical information. In simulation research comparing speech-only with multimodal pen/voice interaction, empirical work demonstrated that multimodal interaction yielded 10% faster task-completion time during visual-spatial tasks, but no significant efficiency advantage in verbal or quantitative task domains (Oviatt, 1997; Oviatt, Cohen, & Wang, 1994). Likewise, users' efficiency improved when they combined speech and gestures multimodally to manipulate 3D objects, compared with unimodal input (Hauptmann, 1989). In another early study, multimodal speech-and-mouse input improved efficiency in a line-art drawing task (Leatherby & Pausch, 1992). Finally, in a study that compared task-completion times for a graphical interface versus a multimodal pen/voice interface, military domain experts averaged four times faster at setting up complex simulation scenarios on a map when they were able to interact multimodally (Cohen, McGee, & Clow, 2000). This latter study was based on testing of a fully functional multimodal system, and it included time required to correct recognition errors.

One particularly advantageous feature of multimodal interface design is its superior error handling, both in terms of error avoidance and graceful recovery from errors (Oviatt & van Gent, 1996; Oviatt, Bernard, & Levow, 1999; Oviatt, 1999a; Rudnicky & Hauptmann, 1992; Suhm, 1998; Tomlinson et al., 1996). There are user-centered and system-centered reasons why multimodal systems facilitate error recovery, when compared with unimodal recognition-based interfaces. For example, in a multimodal speech and pen-based gesture interface users will select the input mode that they judge to be less error prone for particular lexical content, which tends to lead to error avoidance (Oviatt & van Gent, 1996). They may prefer speedy speech input, but will switch to pen input to communicate a foreign surname. Secondly, users' language often is simplified when interacting multimodally, which can substantially reduce the complexity of natural language processing and thereby reduce recognition errors, as described later in this chapter (Oviatt & Kuhn, 1998). In one study, users' multimodal utterances were documented to be briefer, to contain fewer complex locative descriptions, and 50% fewer spoken disfluencies, when compared with a speech-only interface. Thirdly, users have a strong tendency to switch modes after system recognition errors, which facilitates error recovery. This error resolution occurs because the confusion matrices differ for any given lexical content for the different recognition technologies involved in processing (Oviatt et al., 1999; Oviatt, 2002).

In addition to these user-centered reasons for better error avoidance and resolution, there also are system-centered reasons for superior error handling. A well-designed multimodal architecture with two semantically rich input modes can support *mutual disambiguation* of input signals. 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 can result in recovery of the correct plural interpretation. This recovery can occur in a multimodal architecture even though the speech recognizer initially ranks the plural interpretation "ditches" as a less preferred choice on its n-best list. Mutual disambiguation involves recovery from unimodal recognition errors within a multimodal architecture, because semantic information from each input mode supplies partial disambiguation of the other mode, thereby leading to more stable and robust overall system performance (Oviatt, 1999a, 2000a, 2002). Another example of mutual disambiguation is shown in Fig. 21.3. To achieve optimal error handling, a multimodal interface ideally should be designed to include complementary input modes, and

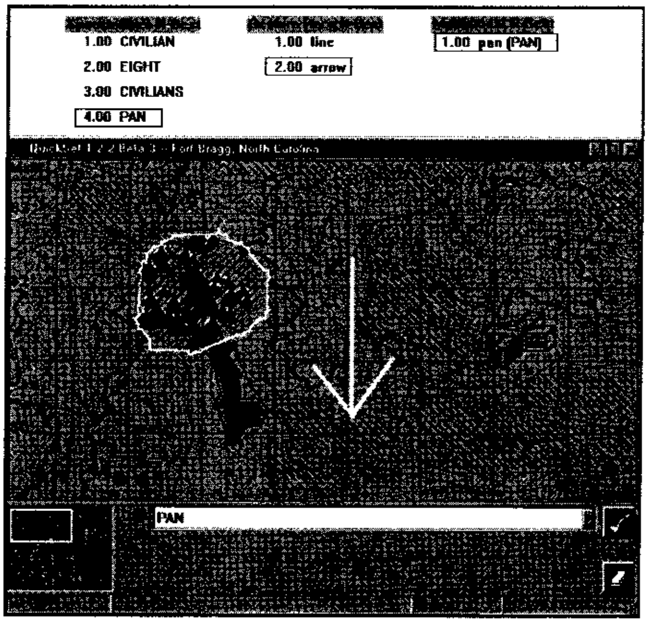

FIGURE 21.3. Multimodal command to "pan" the map, which illustrates mutual disambiguation occurring between incoming speech and gesture information, such that lexical hypotheses were pulled up on both *n*-best lists to produce a correct final multimodal interpretation.

the alternative input modes should provide duplicate functionality such that users can accomplish their goals using either mode.

In two recent studies involving over 4,600 multimodal commands, a multimodal architecture was found to support mutual disambiguation and error suppression ranging between 19% and 41% (Oviatt, 1999a, 2000a, 2002). Improved robustness also was greater for "challenging" user groups (accented vs. native speakers) and usage contexts (mobile vs. stationary use). These results indicate that a well-designed multimodal system not only can perform more robustly than a unimodal system, but also in a more stable way across varied real-world users and usage contexts. Finally, during audio-visual perception of speech and lip movements, improved speech recognition also has been demonstrated for both human listeners (McLeod & Summerfield, 1987) and multimodal systems (Adjoudani & Benoit, 1995; Tomlinson et al., 1996).

Another recent focus has been on the advantages of multimodal interface design for minimizing users' cognitive load. As task complexity increases, there is evidence that users self-manage their working memory limits by distributing information across multiple modalities, which in turn enhances their task performance during both perception and production (Calvert, Spence, & Stein, 2004; Mousavi, Low, & Sweller, 1995; Oviatt, 1997; Oviatt, Coulston, & Lunsford, 2004; Tang, McLachlan, Lowe, Saka, & MacLean, 2005). These predictions and findings are based on Wickens and colleagues' cognitive resource theory and Baddeley's theory of working memory (Baddeley, 1992; Wickens, Sandry, & Vidulich, 1983). The latter maintains that short-term or working memory consists of multiple independent processors associated with different modes. This includes a visual-spatial "sketch pad" that maintains visual materials such as pictures and diagrams in one area of working memory, and a separate phonological loop that stores auditory-verbal information. Although these two processors are believed to be coordinated by a central executive, in terms of lower-level modality processing they are viewed as functioning largely independently, which is what enables the effective size of working memory to expand when people use multiple modalities during tasks (Baddeley, 1992). So with respect to management of cognitive load, the inherent flexibility of multimodal interfaces is well suited to accommodating the high and changing load con ditions typical of realistic mobile use.

WHAT METHODS AND INFORMATION HAVE BEEN USED TO DESIGN NOVEL MULTIMODAL INTERFACES?

The design of new multimodal systems has been inspired and organized largely by two things. First, the cognitive science literature on intersensory perception and intermodal coordination during production is beginning to provide a foundation of information for user modeling, as well as information on what systems must recognize and how multimodal architectures should be organized. For example, the cognitive science literature has provided knowledge of the natural integration patterns that typify people's lip and facial movements with speech output (Benoit, Guiard-Marigny, Le Goff, & Adjoudani, 1996; Ekman, 1992; Ekman & Friesen, 1978; Fridlund, 1994; Hadar, Steiner, Grant, & Rose, 1983; Massaro & Cohen, 1990; Stork & Hennecke, 1995; Vatikiotis-Bateson et al., 1996), and their coordinated use of manual or pen-based gestures with speech (Kendon, 1980; McNeill, 1992; Oviatt, DeAngeli, & Kuhn, 1997). Given the complex nature of users' multimodal interaction, cognitive science has played and will continue to play an essential role in guiding the design of robust multimodal systems. In this respect, a multidisciplinary perspective will be more central to successful multimodal system design than it has been for traditional GUI design. The cognitive science underpinnings of multimodal system design are described later in this chapter.

Secondly, high-fidelity automatic simulations also have played a critical role in prototyping new types of multimodal systems (Dahlbäck, Jëonsson, & Ahrenberg, 1992; Oviatt et al., 1992). When a new multimodal system is in the planning stages, design sketches and low-fidelity mock-ups may initially be used to visualize the new system and plan the sequential flow of human-computer interaction. These tentative design plans then are rapidly transitioned into a higher-fidelity simulation of the multimodal system, which is available for proactive and situated data collection with the intended user population. High-fidelity simulations have been the preferred method for designing and evaluating new multimodal systems, and extensive data collection with such tools preferably is completed before a fully functional system ever is built.

During high-fidelity simulation testing, a user interacts with what she believes is a fully functional multimodal system although the interface is actually a simulated front-end designed to appear and respond as the fully functional system would. During the interaction, a programmer assistant at a remote location provides the simulated system responses. As the user interacts with the front end, the programmer tracks her multimodal input and provides system responses as quickly and accurately as possible. To support this role, the programmer makes use of automated simulation software that is designed to support interactive speed, realism with respect to the targeted system, and other important characteristics. For example, with these automated tools, the programmer may be able to make a single selection on a workstation field to rapidly send simulated system responses to the user during a data-collection session.

High-fidelity simulations have been the preferred method for prototyping multimodal systems for several reasons. Simulations are relatively easy and inexpensive to adapt, compared with building and iterating a complete system. They also permit researchers to alter a planned system's characteristics in major ways (e.g., input and output modes available), and to study the impact of different interface features in a systematic and scientific manner (e.g., type and base-rate of system errors). In comparison, a particular system with its fixed characteristics is a less flexible and suitable research tool, and the assessment of any single system basically amounts to an individual case study. Using simulation techniques, rapid adaptation and investigation of planned system features permit researchers to gain a broader and more principled perspective on the potential of newly emerging technologies. In a practical sense, simulation research can assist in the evaluation of critical performance tradeoffs and in making decisions about alternative system designs, which designers must do as they strive to create more usable multimodal systems.

To support the further development and commercialization of multimodal systems, additional infrastructure that will be needed in the future includes (a) simulation tools for rapidly building and reconfiguring multimodal interfaces, (b) automated tools for collecting and analyzing multimodal corpora, and (c) automated tools for iterating new multimodal systems to improve their performance (see Oviatt et al., 2000, for further discussion).

WHAT ARE THE COGNITIVE SCIENCE UNDERPINNINGS OF MULTIMODAL INTERFACE DESIGN?

This section discusses the growing cognitive science literature that provides the empirical underpinnings needed to design next-generation multimodal interfaces. The ability to develop multimodal systems depends on knowledge of the natural integration patterns that typify people's combined use of different input modes. In particular, the design of new multimodal systems depends on intimate knowledge of the properties of different modes and the information content they carry, the unique characteristics of multimodal language and its processability, and the integration and synchronization characteristics of users' multimodal interaction. It also relies on accurate prediction of when users are likely to interact multimodally, and how alike different users are in their specific integration patterns. The relevant cognitive science literature on these topics is very extensive, especially when consideration is given to all of the underlying sensory perception and production capabilities involved in different input modes currently being incorporated in new multimodal interfaces. As a result, this section will be limited to introducing the main cognitive science themes and findings that are relevant to the more common types of multimodal system.

This cognitive science foundation also has played a key role in identifying computational "myths" about multimodal interaction, and replacing these misconceptions with contrary empirical evidence. Figure 21.4 summarizes 10 common myths about multimodal interaction, which are addressed and discussed in more detail elsewhere (Oviatt, 1999b). As such, the literature summarized in this section aims to provide a more accurate foundation for guiding the design of next-generation multimodal systems.

Ten Myths of Multimodal Interaction

Myth #1: *If you build a multimodal system, users will interact multimodally*

Myth #2: *Speech & pointing is the dominant multimodal integration pattern*

Myth #3: *Multimodal input involves simultaneous signals*

Myth #4: *Speech is the primary input mode in any multimodal system that includes it*

Myth #5: *Multimodal language does not differ linguistically from unimodal language*

Myth #6: *Multimodal integration involves redundancy of content between modes*

Myth #7: *Individual error-prone recognition technologies combine multimodally to produce even greater unreliability*

Myth #8: *All users' multimodal commands are integrated in a uniform way*

Myth #9: *Different input modes are capable of transmitting comparable content*

Myth #10: *Enhanced efficiency is the main advantage of multimodal systems*

(taken from Oviatt, 1999b)

FIGURE 21.4. Ten myths of multimodal interaction: Separating myth from empirical reality.

When Do Users Interact Multimodally?

During natural interpersonal communication, people are always interacting multimodally. Of course, in this case the number of information sources or modalities that an interlocutor has available to monitor is essentially unlimited. However, all multimodal systems are constrained in the number and type of input modes they can recognize. Also, a user can compose active input during human-computer interaction that either is delivered multimodally or that is delivered entirely using just one mode. That is, although users in general may have a strong preference to interact multimodally rather than unimodally, this is no guarantee that they will issue every command to a system multimodally, given the particular type of multimodal interface available. Therefore, the first nontrivial question that arises during system processing is whether a user is communicating unimodally or multimodally.

In the case of speech- and pen-based multimodal systems, users typically mix unimodal and multimodal expressions. In one study involving a visual-spatial domain, users' commands were expressed multimodally 20% of the time, with others just spoken or written (Oviatt et al., 1997). In contrast, in other spatial domains, the ratio of users' multimodal interaction often is 65% to 70% (Oviatt, 1999b; Oviatt et al., 2004). Predicting whether a user will express a command multimodally also depends on the type of action she is performing. In particular, users usually express commands multimodally when describing spatial information about the location, number, size, orientation, or shape of an object. In one study, 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 (Oviatt et al., 1997). 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 expressed themselves multimodally less than 1% of the time. These data emphasize that future multimodal systems will need to distinguish between instances in which 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. They also suggest that knowledge of the type of actions to be included in an application, such as whether the application entails manipulating spatial information, should influence the basic decision of whether to build a multimodal interface at all.

Findings from a more recent study reveal that multimodal interface users spontaneously respond to dynamic changes in their own cognitive load by shifting to multimodal communication as load increases with task difficulty and communicative complexity (Oviatt, Coulston, & Lunsford, 2004). Given a flexible multimodal interface, users' ratio of multimodal (versus unimodal) interaction increased substantially from 18.6% when referring to established dialogue context to 77.1% when required to establish a new context, a +315% relative increase. Likewise, the ratio of users' multimodal interaction increased significantly as the tasks became more difficult, from 59.2% during low difficulty tasks, to 65.5% at moderate difficulty, 68.2% at high, and 75.0% at very high difficulty, an overall relative increase of +27%. These adaptations in multimodal interaction levels reflect users' efforts to self-manage limitations in their working memory as discourse-level demands and task complexity increased. As discussed earlier, users accomplished this by distributing communicative information across multiple modalities in a manner compatible with a cognitive load theory of multimodal interaction. This interpretation is consistent with Baddeley's theory of working memory (Baddeley, 1992), as well as the growing literatures within education (Mousavi, Low, & Sweller, 1995; Sweller, 1988), linguistics (Almor, 1999), and multisensory perception (Calvert, Spence, & Stein, 2004; Ernst & Bulthoff, 2004). Recent work on visual and haptic processing under workload also indicates that presentation of haptic feedback during a complex task can augment users' ability to handle visual information overload (Tang, McLachlan, Lowe, Saka, & MacLean, 2005).

In a multimodal interface that processes passive or blended input modes, there always is at least one passively tracked input source providing continuous information (e.g., gaze tracking, head position). In these cases, all user input would by definition be classified as multimodal, and the primary problem would become segmentation and interpretation of each continuous input stream into meaningful actions of significance to the application. In the case of blended multimodal interfaces (e.g., gaze tracking and mouse input), it still may be opportune to distinguish active forms of user input that might be more accurately or expeditiously handled as unimodal events.

What Are the Integration and Synchronization Characteristics of Users' Multimodal Input?

The past literature on multimodal systems has focused largely on simple selection of objects or locations in a display, rather than considering the broader range of multimodal integration patterns. Since the development of Bolt's (1980) "Put That There" system, speak-and-point has been viewed as the prototypical form of multimodal integration. In Bolt's 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. Since that time, other multimodal systems also have attempted to resolve deictic expressions using a similar approach—for example, using gaze location instead of manual pointing (Koons et al., 1993).

Unfortunately, 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 much more richly expressive than simple pointing or selection. In fact, studies of users' integrated pen/voice input indicate that a speak-and-point pattern only comprises 14% of all spontaneous multimodal utterances (Oviatt et al., 1997). Instead, pen input more often is used to create graphics, symbols and signs, gestural marks, digits, and lexical content. During interpersonal multimodal communication, linguistic analysis of spontaneous manual gesturing also indicates that simple pointing accounts for less than 20% of all gestures (McNeill, 1992). Together, these cognitive science and user-modeling data highlight 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. It is clear that a broader set of multimodal integration issues needs to be addressed in future work. Future research also should explore typical integration patterns between other promising modality combinations, such as speech and gaze.

It also is commonly assumed that any signals involved in a multimodal construction will co-occur temporally. The presumption is that this temporal overlap then determines which signals to combine during system processing. In the case of speech and manual gestures, successful processing of the deictic term "that square" in Bolt's original system relied on interpretation of pointing when the word "that" was 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 (Oviatt et al., 1997).

Beyond the issue of deixis, a series of studies has shown that users' input frequently does not overlap at all during multimodal commands to a computer (Oviatt, 1999b; Oviatt et al., 2003; Oviatt et al., 2005; Xiao, Girand, & Oviatt, 2002; Xiao, Lunsford, Coulston, Wesson, & Oviatt, 2003). In fact, there are two distinct types of user with respect to integration patterns: *simultaneous* integrators and *sequential* ones. A user who habitually integrates her speech and pen input in a *simultaneous* manner overlaps them temporally, whereas a sequential integrator finishes one mode before beginning the second, as summarized in Fig. 21.2. These two types of user integration patterns occur across the lifespan from children through the elderly (Oviatt et al., 2005; Xiao et al., 2002; Xiao et al., 2003). They also can be detected almost immediately during multimodal interaction, usually on the very first input. Users' habitual integration pattern remains strikingly highly consistent during a session, as well as resistant to change following explicit instructions or attempts at training (Oviatt et al., 2003; Oviatt et al., 2005). This bimodal distribution of user integration patterns has been observed in different task domains (e.g., map-based real estate selection, crisis management, educational applications with animated characters), and also when using different types of interface (e.g., conversational, command style) (Oviatt, 1999b; Xiao et al., 2002; Xiao et al., 2003). In short, empirical studies have demonstrated that this bimodal distinction between users in their fundamental integration pattern generalizes widely across different age groups, task domains, and types of interface.

One interesting discovery in recent work is the phenomenon of ***multimodal hypertiming***, which refers to the fact that both sequential and simultaneous integrators will entrench further or accentuate their habitual multimodal integration pattern (*e.g.*, increasing their intermodal *lag* during sequential integrations, or *overlap* during simultaneous integrations, as summarized in 21.2) during system error handling or when completing increasingly difficult tasks. In fact, users will progressively increase their degree of entrenchment by 18% as system errors increase, and by 59% as task difficulty increases (Oviatt et al., 2003). As such, changes in the degree of users' multimodal hypertiming provide a potentially sensitive means of evaluating their cognitive load during real-time interactive exchanges. In the context of system error handling, the phenomenon of multimodal hypertiming basically replaces the hyperarticulation that is typically observed in users during error-prone speech-only interactions.

Given the bimodal distribution of user integration patterns, ***adaptive temporal thresholds*** potentially could support more tailored and flexible approaches to fusion. Ideally, an adaptive multimodal system would detect, automatically learn, and adapt to a user's dominant multimodal integration pattern, which could result in substantial improvements in system processing speed, accuracy of interpretation, and synchronous interchange with the user. For example, it has been estimated that system delays could be reduced to approximately 40% to 50% of what they currently are by adopting user-defined thresholds (Oviatt et al., 2005). Recent research has begun comparing different learning-based models for adapting a multimodal system's temporal thresholds to an individual user in real time (Huang & Oviatt, in press).

Unfortunately, users' multimodal integration patterns have not been studied as extensively or systematically for other input modes, such as speech and manual gesturing. Linguistics research on interpersonal communication patterns has revealed that both spontaneous gesturing and signed language often precede their spoken lexical analogues during human communication (Kendon, 1980; Naughton, 1996), when considering word-level integration pattern. In fact, the degree to which gesturing precedes speech is greater in topic-prominent languages such as Chinese than it is in subject-prominent ones like Spanish or English (McNeill, 1992). Even in the speech and lip-movement literature, close but not perfect temporal synchrony is typical, with lip movements occurring a fraction of a second before the corresponding auditory signal (Abry, Lallouache, & Cathiard, 1996; Benoit, 2000). However, when considering the whole user utterance as the unit of analysis, some other studies of speech and manual gesturing have found a higher rate of simultaneity for these modes (Epps, Oviatt, & Chen, 2004). Learning-based approaches that are capable of accurately identifying and adapting to different multimodal integration patterns, whether due to differences among users, modality combinations, or applications and usage contexts, will be required in order to generalize and speed up multimodal system development in the future.

In short, although two input modes may be 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, multimodal system designers cannot necessarily count on conveniently overlapped signals in order to achieve successful processing in the multimodal architectures they build. Future research needs to explore the integration patterns and temporal cascading that can occur among three or more input modes—such as gaze, gesture, and speech—so that more advanced multimodal systems can be designed and prototyped.

In the design of new multimodal architectures, it is important to note that data on the order of input modes and average time lags between input modes has been used to determine the likelihood that an utterance is multimodal versus unimodal, and to establish temporal thresholds for fusion of input. In the future, weighted likelihoods associated with different utterance segmentations, for example, that an input stream containing speech, writing, speech should be segmented into [S/W S] rather than [S W/S], and with intermodal time lag distributions, will be used to optimize correct recognition of multimodal user input (Oviatt, 1999b). In the design of future time-critical multimodal architectures, data on users' integration and synchronization patterns will need to be collected for other mode combinations during realistic interactive tasks, so that temporal thresholds can be established for performing multimodal fusion.

What Individual Differences Exist in Multimodal Interaction, and What Are the Implications for Designing Systems for Universal Access?

There are large individual differences in users' multimodal interaction patterns, beginning with their overall preference to interact unimodally versus multimodally, and which mode they generally prefer (e.g., speaking versus writing) (Oviatt et al., 2004). As outlined above, there likewise are striking differences among users in adopting either a sequential or simultaneous multimodal integration pattern. Recent research has revealed that these two patterns are associated with behavioral and linguistic differences between the groups (Oviatt et al., 2005). Whereas in an interactive task context their performance speed was comparable, sequential integrators were far less error-prone and excelled during new or complex tasks. Although their speech rate was no slower, sequential integrators also had more precise articulation (e.g., less disfluency). Finally, sequential integrators were more likely to adopt terse and direct command-style language, with a smaller and less varied vocabulary, which appeared focused on achieving error-free communication. These user differences in interaction patterns have been interpreted as deriving from fundamental differences among users in their reflective-impulsive cognitive style (Oviatt et al., 2005). Based on this work, one goal of future multimodal interface design will be to support the poorer attention span and higher error rate of impulsive users—especially for mobile in-vehicle, military, and similar application contexts in which the cost of committing errors is unacceptably high.

Apart from these individual differences, cultural differences also have been documented between users in modality integration patterns. For example, substantial individual differences have been reported in the temporal synchrony between speech and lip movements (Kricos, 1996) and, in addition, lip movements during speech production are known to be less exaggerated among Japanese speakers than Americans (Sekiyama & Tohkura, 1991). In fact, extensive inter-language differences have been observed in the information available from lip movements during audio-visual speech (Fuster-Duran, 1996). These findings have implications for the degree to which disambiguation of speech can be achieved through lip movement information in noisy environments or for different user populations. Finally, nonnative speakers, the hearing impaired, and elderly listeners all are more influenced by visual lip movement than auditory cues when processing speech (Fuster-Duran, 1996; Massaro, 1996). These results have implications for the design and expected value of audio-visual multimedia output for different user groups in animated character interfaces. With respect to support for universal access, recent work also has shown the advantage of combined audio-visual processing for recognition of impaired speech (Potamianos & Neti, 2001).

Finally, gender, age, and other individual differences are common in gaze patterns, as well as speech and gaze integration (Argyle, 1972). As multimodal interfaces incorporating gaze become more mature, further research will need to explore these gender and age-specific patterns, and to build appropriately adapted processing strategies. In summary, considerably more research is needed on multimodal integration and synchronization patterns for new mode combinations, as well as for diverse and disabled users for whom multimodal interfaces may be especially suitable for ensuring universal access.

Is Complementarity or Redundancy the Main Organizational Theme That Guides Multimodal Integration?

It frequently 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. 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 (Oviatt et al., 1997). In fact, a major complementarity between speech and manually oriented pen input involves visual-spatial semantic content, which is one reason these modes are an opportune combination for visual-spatial applications. Whereas spatial information is uniquely and clearly indicated via pen input, the strong descriptive capabilities of speech are better suited for specifying temporal and other nonspatial information. Even during multimodal correction of system errors, when users are highly motivated to clarify and reinforce their information delivery, speech and pen input express redundant information less than 1% of the time. Finally, during interpersonal communication linguists also have documented that spontaneous speech and manual gesturing involve complementary rather than duplicate information between modes (McNeill, 1992).

Other examples of primary multimodal complementarities during interpersonal and human-computer communication have been described in past research (McGurk & MacDonald, 1976; Oviatt & Olsen, 1994; Wickens et al., 1983). For example, in the literature on multimodal speech and lip movements, natural feature-level complementarities have been identified between visemes and phonemes for vowel articulation, with vowel rounding better conveyed visually, and vowel height and backness better revealed auditorally (Massaro & Stork, 1998; Robert-Ribes et al., 1998).

In short, actual data highlight 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, although in certain contexts (such as teaching) a greater percentage of duplicate content than usual may be expected to exist. In multimodal systems involving both speech and pen-based gestures and speech and lip movements, one explicit goal has been to integrate complementary modalities in a manner that yields a synergistic blend, such that each mode can be capitalized upon and used to overcome weaknesses in the other mode (Cohen et al., 1989). This approach to system design has promoted the philosophy of using modes and component technologies to their natural advantage, and of combining them in a manner that permits mutual disambiguation. One advantage of achieving such a blend is that the resulting multimodal architecture can function more robustly than an individual recognition-based technology or a multimodal system based on input modes lacking natural complementarities.

What Are the Primary Features of Multimodal Language?

Communication channels can be tremendously influential in shaping the language transmitted within them. From past research, there now is cumulative evidence that many linguistic features of multimodal language are qualitatively very different from that of spoken or formal textual language. In fact, it can differ in features as basic as brevity, semantic content, syntactic complexity, word order, disfluency rate, degree of ambiguity, referring expressions, specification of determiners, anaphora, deixis, and linguistic indirectness. In many respects, multimodal language is simpler linguistically than spoken language. In particular, comparisons have revealed that the same user completing the same map-based task communicates significantly fewer words, briefer sentences, and fewer complex spatial descriptions and disfluencies when interacting multimodally, compared with using speech alone (Oviatt, 1997). One implication of these findings is that multimodal interface design has the potential to support more robust future systems than a unimodal design approach. The following is an example of a typical user's spoken input while attempting to designate an open space using a map system: "Add an open space on the north lake to b—include the north lake part of the road and north." In contrast, the same user accomplished the same task multimodally by encircling a specific area and saying, "Open space."

In previous research, hard-to-process, disfluent language has been observed to decrease by 50% during multimodal interaction with a map, compared with a more restricted speech-only interaction (Oviatt, 1997). This drop occurs mainly because people have difficulty speaking spatial information, which precipitates disfluencies. In a flexible multimodal interface, they instead use pen input to convey spatial information, thereby avoiding the need to speak it. Further research is needed to establish whether other forms of flexible multimodal communication also generally ease users' cognitive load, which may be reflected in a reduced rate of disfluencies.

During multimodal pen/voice communication, the linguistic indirection that is typical of spoken language frequently is replaced with more direct commands (Oviatt & Kuhn, 1998). In the following example, a study participant made a disfluent indirect request using speech input while requesting a map-based distance calculation: "What is the distance between the Victorian Museum and the, uh, the house on the east side of Woodpecker Lane?" When requesting distance information multimodally, the same user encircled the house and museum while speaking the following brief direct command: "Show distance between here and here." In this research, the briefer and more direct multimodal pen/voice language also contained substantially fewer referring expressions, with a selective reduction in co-referring expressions that instead were transformed into deictic expressions. This latter reduction in coreference would simplify natural language processing by easing the need for anaphoric tracking and resolution in a multimodal interface. Also consistent with fewer referring expressions, explicit specification of definite and indefinite reference is less common in multimodal language (Oviatt & Kuhn, 1998). Current natural language processing algorithms typically rely heavily on the specification of determiners in definite and indefinite references in order to represent and resolve noun-phrase reference. One unfortunate byproduct of the lack of such specifications is that current language processing algorithms are unprepared for the frequent occurrence of elision and deixis in multimodal human-computer interaction.

In other respects, multimodal language clearly is different from spoken language, although not necessarily simpler. For example, users' multimodal pen/voice language departs from the canonical English word order of S-V-O-LOC (e.g., Subject-Verb-Object-Locative constituent), which is observed in spoken language and formal textual language. Instead, users' multimodal constituents shift to a LOC-S-V-O word order. A recent study reported that 95% of locative constituents were in sentence-initial position during multimodal interaction. However, for the same users completing the same tasks while speaking, 96% of locatives were in sentence-final position (Oviatt et al., 1997). It is likely that broader analysis of multimodal communication patterns, which could involve gaze and manual gesturing to indicate location rather than pen-based pointing, would reveal a similar reversal in word order.

One implication of these many differences is that new multimodal corpora, statistical language models, and natural language-processing algorithms will need to be established before multimodal language can be processed optimally. Future research and corpus-collection efforts also will be needed on different types of multimodal communication, and in other application domains, so that the generality of previously identified multimodal language differences can be explored.

WHAT ARE THE BASIC WAYS IN WHICH MULTIMODAL INTERFACES DIFFER FROM GRAPHICAL USER INTERFACES?

Multimodal research groups currently are rethinking and redesigning basic user interface architectures because a whole new range of architectural requirements has been posed. First, graphical user interfaces typically assume that a single event stream controls the underlying event loop, with any processing sequential in nature. For example, most GUIs ignore typed input when a mouse button is depressed. In contrast, multimodal interfaces typically can process continuous and simultaneous input from parallel incoming streams. Secondly, GUIs assume that the basic interface actions, such as selection of an item, are atomic and unambiguous events. In contrast, multimodal systems process input modes using recognition-based technologies, which are designed to handle uncertainty and entail probabilistic methods of processing. Thirdly, GUIs often are built to be separable from the application software that they control, although the interface components usually reside centrally on one machine. In contrast, recognition-based user interfaces typically have larger computational and memory requirements, which often makes it desirable to distribute the interface over a network so that separate machines can handle different recognizers or databases. For example, cell phones and networked PDAs may extract features from speech input, but transmit them to a recognizer that resides on a server. Finally, multimodal interfaces that process two or more recognition-based input streams require time stamping of input, and the development of temporal constraints on mode fusion operations. In this regard, they involve uniquely time-sensitive architectures.

WHAT BASIC ARCHITECTURES AND PROCESSING TECHNIQUES HAVE BEEN USED TO DESIGN MULTIMODAL SYSTEMS?

Many early multimodal interfaces that handled combined speech and gesture, such as Bolt's "Put That There" system (Bolt, 1980), have been based on a control structure in which multimodal integration occurs during the process of parsing spoken language. As discussed earlier, when the user speaks a deictic expression such as "here" or "this," the system searches for a synchronized gestural act that designates the spoken referent. While such an approach is viable for processing a point-and-speak multimodal integration pattern, as discussed earlier, multimodal systems must be able to process richer input than just pointing, including gestures, symbols, graphic marks, lip movements, meaningful facial expressions, and so forth. To support more broadly functional multimodal systems, general processing architectures have been developed since Bolt's time. Some of these recent architectures handle a variety of multimodal integration patterns, as well as the interpretation of both unimodal and combined multimodal input. This kind of architecture can support the development of multimodal systems in which modalities are processed individually as input alternatives to one another, or those in which two or more modes are processed as combined multimodal input.

For multimodal systems designed to handle joint processing of input signals, there are two main subtypes of multimodal architecture. One subtype integrates signals at the *feature level* (e.g., "early fusion"). The other integrates information at a *semantic level* (e.g., "late fusion"). Examples of systems based on an early feature-fusion processing approach include those developed by Bregler and colleagues (Bregler, Manke, Hild, & Waibel, 1993), Vo and colleagues (Vo et al., 1995), and Pavlovic and colleagues (Pavlovic, Sharma, & Huang, 1997; Pavlovic & Huang, 1998). In feature-fusion architecture, the signal-level recognition process in one mode influences the course of recognition in the other. Feature fusion is considered more appropriate for closely temporally synchronized input modalities, such as speech and lip movements (Stork & Hennecke, 1995; Rubin et al., 1998).

In contrast, multimodal systems using the late semantic fusion approach have been applied to processing multimodal speech and pen input or manual gesturing, for which the input modes are less coupled temporally. These input modes provide different but complementary information that typically is integrated at the utterance level. Late semantic integration systems use individual recognizers that can be trained using unimodal data, which are easier to collect and are already publicly available for speech and handwriting. In this respect, systems based on semantic fusion can be scaled up more easily in number of input modes or vocabulary size. Examples of systems based on semantic fusion include Put That There (Bolt, 1980), ShopTalk (Cohen et al., 1989), QuickSet (Cohen et al., 1997), CUBRICON (Neal & Shapiro, 1991), Virtual World (Codella et al., 1992), Finger-Pointer (Fukumoto, Suenaga, & Mase, 1994), VisualMan (Wang, 1995), Human-Centric Word Processor, Portable Voice Assistant (Bers et al., 1998), the VR Aircraft Maintenance Training System (Duncan et al., 1999) and Jeanie (Vo & Wood, 1996).

As an example of multimodal information processing flow in a late-stage semantic architecture, Fig. 21.5 illustrates two input modes (e.g., speech and manual or pen-based gestures) recognized in parallel and processed by an understanding component. The results involve partial meaning representations that are fused by the multimodal integration component, which also is influenced by the system's dialogue management and interpretation of current context. During the integration process, alternative lexical candidates for the final multimodal interpretation

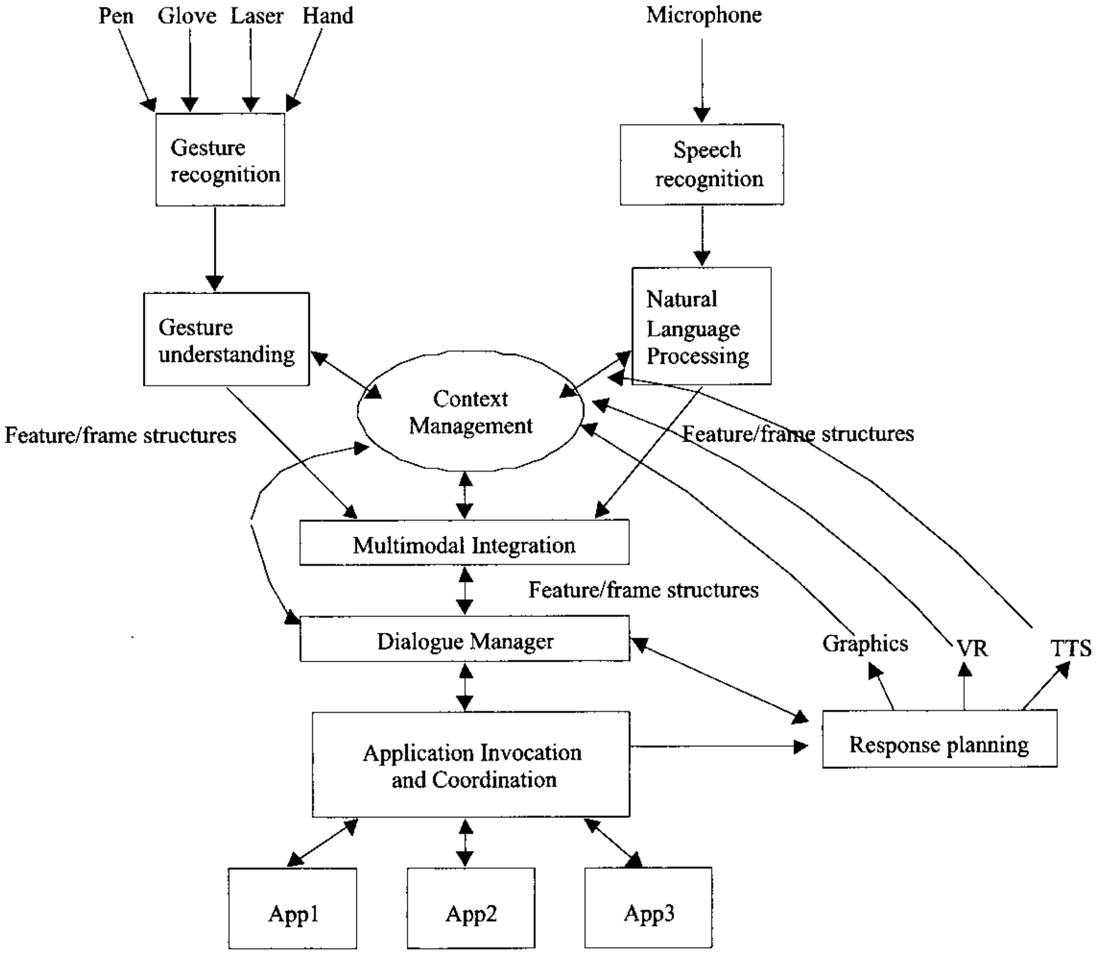

FIGURE 21.5. Typical information processing flow in a multimodal architecture designed for speech and gesture.

are ranked according to their probability estimates on an n-best list. The best-ranked multimodal interpretation then is sent to the application invocation and control component, which transforms this information into a series of commands to one or more back-end application systems. System feedback typically includes multimedia output, which may incorporate text-to-speech and non-speech audio; graphics and animation; and so forth. For examples of feature-based multimodal processing flow and architectures, especially as applied to multimodal speech and lip movement systems, see Benoit et al. (2000).

There are many ways to realize this information processing flow as architecture. One common infrastructure that has been adopted by the multimodal research community involves *multi-agent architectures*, such as the Open Agent Architecture (Cohen, Cheyer, Wang, & Baeg, 1994; Martin, Cheyer, & Moran, 1999) and Adaptive Agent Architecture (Kumar & Cohen, 2000). In a multi-agent architecture, the many components needed to support the multimodal system (e.g., speech recognition, gesture recognition, natural language processing, multimodal integration) may be written in different programming languages, on different machines, and with different operating systems. Agent communication languages are being developed that can handle asynchronous delivery, triggered responses, multi-casting, and other concepts from distributed systems, and that are fault-tolerant (Kumar & Cohen, 2000). Using a multi-agent architecture, for example, speech and gestures can arrive in parallel or asynchronously via individual modality agents, with the results recognized and passed to a facilitator. These results, typically an n-best list of conjectured lexical items and related time-stamp information, then are routed to appropriate agents for further language processing. Next, sets of meaning fragments derived from the speech and pen signals arrive at the multimodal integrator. This agent decides whether and how long to wait for recognition results from other modalities, based on the system's temporal thresholds. It fuses the meaning fragments into a semantically- and temporally-compatible whole interpretation before passing the results back to the facilitator. At this point, the system's final multimodal interpretation is confirmed by the interface, delivered as multimedia feedback to the user, and executed by any relevant applications. In summary, multi-agent architectures provide essential infrastructure for coordinating the many complex modules needed to implement multimodal system processing, and they permit doing so in a distributed manner that is compatible with the trend toward mobile computing.

The core of multimodal systems based on semantic fusion involves algorithms that integrate common meaning representations derived from speech, gesture, and other modalities into a combined final interpretation. The semantic fusion operation requires a common meaning-representation framework for all modalities, and a well-defined operation for combining partial meanings that arrive from different signals. To fuse information from different modalities, various research groups have independently converged on a strategy of recursively matching and merging attribute/value data structures, although using a variety of different algorithms (Vo & Wood, 1996; Cheyer & Julia, 1995; Pavlovic & Huang, 1998; Shaikh et al., 1997). This approach is considered a *frame-based integration* technique. An alternative logic-based approach derived from computational linguistics (Carpenter, 1990, 1992; Calder, 1987) involves the use of *typed feature structures* and *unification-based integration*, which is a more general and well-understood approach. Unification-based integration techniques also have been applied to multimodal system design (Cohen et al., 1997; Johnston et al., 1997; Wu et al., 1999). Feature-structure unification is considered well suited to multimodal integration, because unification can combine complementary or redundant input from both modes, but it rules out contradictory input. Given this foundation for multimodal integration, more research still is needed on the development of canonical meaning representations that are common among different input modes which will need to be represented in new types of multimodal systems.

When statistical processing techniques are combined with a symbolic unification-based approach that merges feature structures, then the multimodal architecture that results is a *hybrid symbolic/statistical* one. Hybrid architectures represent one major new direction for multimodal system development. Multimodal architectures also can be hybrids in the sense of combining Hidden Markov Models (HMMs) and Neural Networks (NNs). New hybrid architectures potentially are capable of achieving very robust functioning, compared with either an early- or late-fusion approach alone. For example, the Members-Teams-Committee (MTC) hierarchical recognition technique, which is a hybrid symbolic/statistical multimodal integration framework trained over a labeled multimodal corpus, recently achieved 95.26% correct recognition performance, or within 1.4% of the theoretical system upper bound (Wu et al., 1999). Other architectural approaches and contributions to processing multimodal information have been summarized elsewhere (Oliver & Horvitz, 2005; Potamianos et al., 2003).

WHAT ARE THE MAIN FUTURE DIRECTIONS FOR MULTIMODAL INTERFACE DESIGN?

The computer science community is just beginning to understand how to design innovative, well integrated, and robust multimodal systems. To date, most multimodal systems remain bimodal, and recognition technologies related to several human senses (e.g., haptics, smell, taste) have yet to be well represented within multimodal interfaces. The design and development of new types of systems that include such modes will not be achievable through intuition. Rather, it will depend on knowledge of the natural integration patterns that typify people's combined use of various input modes. This means that the successful design of multimodal systems will continue to require guidance from cognitive science on the coordinated human perception and production of natural modalities. In this respect, multimodal systems only can flourish through multidisciplinary cooperation, as well as teamwork among those representing expertise in the component technologies.

Most of the systems outlined in this chapter have been built during the past 15 years, and they are research-level systems. However, in some cases they have developed well beyond the prototype stage, and are being integrated with other software at academic and federal sites, or appearing as newly shipped products. To achieve wider commercialization of multimodal interfaces, such systems will need to develop more powerful and general methods of natural language and dialogue processing, and temporal modeling and processing of incoming signals. In addition, multimodal datasets and tools are very much needed to build applications more rapidly in a wide range of domains, including for newly emerging collaborative multimodal applications such as meeting support and education (Barthelmess, Kaiser, Huang, & Demirdjian, 2005; Cohen & McGee, 2004; Danninger et al., 2005; Gatica-Perez, Lathoud, Odobez, & McCowan, 2005; McGee, 2003; Pentland, 2005). The many mobile multimodal interfaces currently being built also will require active adaptation to the user, task, ongoing dialogue, and environmental context, which is another very active area of recent work (Gorniak & Roy, 2005; Gupta, 2004; Huang & Oviatt, 2005; Jain & Ross, 2002; Potamianos et al., 2003; Xiao et al., 2003). To facilitate the speed and generality of multimodal interface adaptation to these important variables, future work will need to integrate new machine learning techniques that are now being developed to handle asynchronous and heterogeneous data (Bengio, 2004; McCowan et al., 2005). Finally, in the future a coherent theoretical framework needs to be developed to account for multimodal interaction patterns. This will be invaluable for proactively guiding the design of new multimodal interfaces to be compatible with human capabilities and limitations. Current work in cognitive neuroscience and multisensory perception are beginning to provide an empirical and theoretical basis for this future interface design (Calvert et al., 2004; Ernst & Bulthoff, 2004).

In conclusion, multimodal interfaces are just beginning to model human-like sensory perception. They are recognizing and identifying actions, language, and people that have been seen, heard, or in other ways experienced in the past. They literally reflect and acknowledge the existence of human users, empower them in new ways, and create for them a "voice." They also can be playful and self-reflective interfaces that suggest new forms of human identity as we interact face to face with animated personas representing our own kind. In all of these ways novel multimodal interfaces, as primitive as their early bimodal instantiations may be, represent a new multidisciplinary science, a new art form, and a socio-political statement about our collective desire to humanize the technology we create.

ACKNOWLEDGMENTS

I'd like to thank the National Science Foundation for their support over the past decade, which has enabled me to pursue basic exploratory research on many aspects of multimodal interaction, interface design, and system development. The preparation of this chapter has been supported by NSF Grants IRI-9530666, IIS-0117868, and NSF Special Extension for Creativity (SEC) Grant IIS-9530666. This work also has been supported by contracts DABT63-95-C-007, N66001-99-D-8503, and NBCHD030010 from DARPA's Information Technology and Information Systems Office, and Grant No. N00014-99-1-0377 from ONR. Any opinions, findings or conclusions expressed in this chapter are those of the author and do not necessarily reflect the views of the federal agencies sponsoring this work.

I'd also like to thank Phil Cohen and others in the Center for Human-Computer Communication for many insightful discussions, and Dana Director and Rachel Coulston for expert assistance with manuscript preparation. Finally, I wish to acknowledge LEA, Inc. for giving permission to reprint Figs. 21.1 and 21.5, and to acknowledge ACM for allowing the reprint of Figs. 21.3 and 21.4.

References

Abry, C., Lallouache, M.-T., & Cathiard, M.-A. (1996). How can coarticulation models account for speech sensitivity to audio-visual desynchronization? In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines: Models, Systems and Applications* (pp. 247–255). New York: Springer Verlag.

Adjoudani, A., & Benoit, C. (1995). Audio-visual speech recognition compared across two architectures. *Proc. of the Eurospeech Conference Vol. 2* (pp. 1563–1566). Madrid, Spain.

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review, 106*, 748–765.

Argyle, M. (1972). nonverbal communication in human social interaction. In R. Hinde (Ed.), *Nonverbal Communication* (pp. 243–267). Cambridge, UK: Cambridge University Press.

Baddeley, A. (1992). Working memory. *Science, 255*, 556–559.

Bangalore, S., & Johnston, M. (2000). Integrating multimodal language processing with speech recognition. In B. Yuan, T. Huang & X. Tang (Eds.), *Proceedings of the International Conference on Spoken Language Processing (ICSLP'2000) Vol. 2* (pp. 126–129). Beijing: Chinese Friendship Publishers.

Barthelmess, P., Kaiser, E., Huang, X., & Demirdjian, D. (2005). Distributed pointing for multimodal collaboration over sketched diagrams, *Proceedings of the Seventh International Conference on Multimodal Interfaces* (pp. 10–17) New York, N.Y.: ACM Press.

Bengio, S. (2004). Multimodal speech processing using asynchronous Hidden Markov Models, *Information Fusion, 5*(2), 81–89.

Benoit, C., Guiard-Marigny, T., Le Goff, B., & Adjoudani, A. (1996). Which components of the face do humans and machines best speechread? In D. G. Stork & M. E. Hennecke, (Eds.), *Speechreading by Humans and Machines: Models, Systems, and Applications: Vol. 150 of NATO ASI Series. Series F: Computer and Systems Sciences* (pp. 315–325). Berlin, Germany: Springler-Verlag.

Benoit, C., & Le Goff, B. (1998). Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. *Speech Communication, 26*, 117–129.

Benoit, C., Martin, J.-C., Pelachaud, C., Schomaker, L., & Suhm, B. (2000). Audio-visual and multimodal speech-based systems. In D. Gibbon, I. Mertins, & R. Moore (Eds.), *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation* (pp. 102–203). Amsterdam, the Netherlands: Kluwer.

Bernstein, L., & Benoit, C. (1996). For speech perception by humans or machines, three senses are better than one. *Proceedings of the International Conference on Spoken Language Processing, (ICSLP 96) Vol. 3* (pp. 1477–1480). New York: IEEE press.

Bers, J., Miller, S., & Makhoul, J. (1998). Designing conversational interfaces with multimodal interaction. *DARPA Workshop on Broadcast News Understanding Systems*, 319–321.

Bolt, R. A. (1980). Put-that-there: Voice and gesture at the graphics interface. *Computer Graphics, 14*(3), 262–270.

Bregler, C., & Konig, Y. (1994). Eigenlips for robust speech recognition. *Proceedings of the International Conference on Acoustics Speech and Signal Processing* (*IEEE-ICASSP*) *Vol. 2*, 669–672.

Bregler, C., Manke, S., Hild, H., & Waibel, A. (1993). Improving connected letter recognition by lipreading. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (*IEEE-ICASSP*) *Vol.1*, 557–560.

Brooke, N. M., & Petajan, E. D. (1986). Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics. *Proceedings International Conference Speech Input and Output: Techniques and Applications, 258*, 104–109.

Calder, J. (1987). Typed unification for natural language processing. In E. Klein & J. van Benthem (Eds.), *Categories, Polymorphisms, and Unification* (pp. 65–72). Edinburgh, Scotland: Center for Cognitive Science, University of Edinburgh.

Calvert, G., Spence, C., & Stein, B. E. (Eds.). (2004). *The Handbook of Multisensory Processing*. Cambridge, MA: MIT Press.

Carpenter, R. (1990). Typed feature structures: Inheritance, (in)equality, and extensionality. *Proceedings of the ITK Workshop: Inheritance in Natural Language Processing*, 9–18. Tilburg the Netherlands: Institute for Language Technology and Artificial Intelligence, Tilburg University.

Carpenter, R. (1992). *The logic of typed feature structures.* Cambridge, England: Cambridge University Press.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (Eds.). (2000). *Embodied conversational agents.* Cambridge, MA: MIT Press.

Cheyer, A. (1998, January). MVIEWS: Multimodal tools for the video analyst. *International Conference on Intelligent User Interfaces* (*IUI'98*), (pp. 55–62). New York: ACM Press.

Cheyer, A., & Julia, L. (1995, May). Multimodal maps: An agent-based approach. *International Conference on Cooperative Multimodal Communication,* (*CMC'95*), (pp. 103–113). Eindhoven, The Netherlands.

Choudhury, T., Clarkson, B., Jebara, T., & Pentland, S. (1999). Multimodal person recognition using unconstrained audio and video. *Proceedings of the 2nd International Conference on Audio-and-Video-based Biometric Person Authentication* (pp. 176–181). Washington, DC:.

Codella, C., Jalili, R., Koved, L., Lewis, J., Ling, D., Lipscomb, J., et al. (1992). Interactive simulation in a multi-person virtual world. *Proceedings of the Conference on Human Factors in Computing Systems* (*CHI'92*) (pp. 329–334). New York: ACM Press.

Cohen, M. M., & Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. In M. Magnenat-Thalmann & D. Thalmann (Eds.), *Models and Techniques in Computer Animation.* Tokyo: Springer-Verlag.

Cohen, P. R., Cheyer, A., Wang, M., & Baeg, S. C. (1994). An open agent architecture. *AAAI '94 Spring Symposium Series on Software Agents*, 1–8. (Reprinted in Huhns and Singh (Eds.). (1997). *Readings in Agents* (pp. 197–204). San Francisco: Morgan Kaufmann.)

Cohen, P. R., Dalrymple, M., Moran, D. B., Pereira, F. C. N., Sullivan, J. W., Gargan, R. A., et al. (1989). Synergistic use of direct manipulation and natural language. *Proceedings of the Conference on Human Factors in Computing Systems* (*CHI'89*), (pp. 227–234). New York: ACM Press. Reprinted in Maybury & Wahlster (Eds.), (1998). *Readings in Intelligent User Interfaces* (pp. 29–37). San Francisco: Morgan Kaufmann.)

Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., et al. (1997). Quickset: Multimodal interaction for distributed applications. *Proceedings of the Fifth ACM International Multimedia Conference* (pp. 31–40). New York: ACM Press.

Cohen, P. R., & McGee, D. (2004, January). Tangible multimodal interfaces for safety-critical applications, *Communications of the ACM, Vol. 47*. (pp. 41–46). New York, NY: ACM Press.

Cohen, P. R., McGee, D. R., & Clow, J. (2000). The efficiency of multimodal interaction for a map-based task. *Proceedings of the Language Technology Joint Conference* (*ANLP-NAACL 2000*) (pp. 331–338). Seattle, WA: Association for Computational Linguistics Press.

Cohen, P. R., & Oviatt, S. L. (1995). The role of voice input for human-machine communication. *Proceedings of the National Academy of Sciences, 92*(22), (pp. 9921–9927). Washington, DC: National Academy of Sciences Press.

Condon, W. S. (1988). An analysis of behavioral organization. *Sign Language Studies, 58*, 55–88.

Dahlbäck, N., Jëonsson, A., & Ahrenberg, L. (1992, January). Wizard of Oz studies—why and how. In W. D. Gray, W. E. Hefley, & D. Murray (Eds.), *Proceedings of the International Workshop on Intelligent User Interfaces* (pp. 193–200). New York: ACM Press.

Danninger, M., Flaherty, G., Bernardin, K., Ekenel, H., Kohler, T., Malkin, R., et al. (2005). The connector- Facilitaing context-aware communication, *Proceedings of the International Conference on Multimodal Interfaces* (pp. 69–75) New York: ACM.

Denecke, M., & Yang, J. (2000). Partial Information in Multimodal Dialogue. *Proceedings of the International Conference on Multimodal Interaction*, (pp. 624–633). Beijing, China:Friendship Press.

Duncan, L., Brown, W., Esposito, C., Holmback, H., & Xue, P. (1999). Enhancing virtual maintenance environments with speech understanding. Boeing M&CT TechNet.

Dupont, S., & Luettin, J. (2000, September). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia, 2*(3), 141–151. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Ekman, P. (1992, January). Facial expressions of emotion: New findings, new questions. *American Psychological Society, 3*(1), 34–38.

Ekman, P., & Friesen, W. (1978). *Facial Action Coding System.* Mountain View, CA: Consulting Psychologists Press.

Encarnacao, L. M., & Hettinger, L. (Eds.) (2003, September/October) Perceptual Multimodal Interfaces (special issue), *IEEE Computer Graphics and Applications.* Mountain View, CA.

Epps, J., Oviatt, S., & Chen, F. (2004). Integration of speech and gesture input during multimodal interaction. *Proceedings of the Australian International Conference on Computer-Human Interaction* (*OzCHI*).

Ernst, M., & Bulthoff, H. (2004, April). Merging the sense into a robust whole percept. *Trends in Cognitive Science, 8*(4), 162–169.

Fell, H., Delta, H., Peterson, R., Ferrier, L., Mooraj, Z., & Valleau, M. (1994). Using the baby-babble-blanket for infants with motor problems. *Proceedings of the Conference on Assistive Technologies* (*ASSETS'94*), 77–84.

Flanagan, J., & Huang, T. (Eds.). (2003, September). Multimodal Human Computer Interfaces (special issue), *Proceedings of IEEE, Vol. 91* (9).

Fridlund, A. (1994). *Human facial expression: An evolutionary view.* New York: Academic Press.

Fukumoto, M., Suenaga, Y., & Mase, K. (1994). Finger-pointer: Pointing interface by image processing. *Computer Graphics, 18*(5), 633–642.

Fuster-Duran, A. (1996). Perception of conflicting audio-visual speech: an examination across Spanish and German, In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines: Models, Systems and Applications* (pp. 135–143). New York: Springer Verlag.

Gatica-Perez, D., Lathoud, G., Odobez, J.-M., & McCowan, I. (2005). Multimodal multispeaker probabilistic tracking in meetings. *Proceedings of the Seventh International Conference on Multimodal Interfaces*, (pp. 183–190). New York: ACM.

Gorniak, P., & Roy, D. (2005). Probabilistic grounding of situated speech using plan recognition and reference resolution. *Proceedings of the Seventh International Conference on Multimodal Interfaces* (pp. 138–143). New York: ACM.

Gupta, A. (2004). Dynamic time windows for multimodal input fusion, *Proceedings of the International Conference on Spoken Language Processing (ICSLP'04)*. 1009–1012

Hadar, U., Steiner, T. J., Grant, E. C., & Rose, C. F. (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science, 2*, 35–46.

Hauptmann, A. G. (1989). Speech and gestures for graphic image manipulation. *Proceedings of the Conference on Human Factors in Computing Systems (CHI'89), Vol. 1* (pp. 241–245). New York: ACM Press.

Holzman, T. G. (1999). Computer-human interface solutions for emergency medical care. *Interactions*, 6(3), 13–24.

Huang, X., Acero, A., Chelba, C., Deng, L., Duchene, D., Goodman, J., et al. (2000). MiPad: A next-generation PDA prototype. *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000) Vol. 3* (pp. 33–36). Beijing, China: Chinese Military Friendship Publishers.

Huang, X. & Oviatt, S. Toward adaptive information fusion in multimodal systems, *Second Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MIML'05)*, Springer Lecture Notes in Computer Science, ed. by Steve Renals & Samy Bengio, Springer-Verlag GmbH, 2006, vol. 3869, 15–27.

Iyengar, G., Nock, H., & Neti, C. (2003). Audio-visual synchrony for detection of monologues in video archives. *Proceedings of ICASSP*.

Jain, A., Hong, L., & Kulkarni, Y. (1999). A multimodal biometric system using fingerprint, face and speech, *2nd International Conference on Audio- and Video-based Biometric Person Authentication* (pp. 182–187). Washington DC:

Jain, A., & Ross, A. (2002). Learning user-specific parameters in a multibiometric system. *Proceedings of the International Conference on Image Processing (ICIP)*, September 22–25, 2002. (pp. 57–60) Rochester, NY:

Johnston, M., Cohen, P. R., McGee, D., Oviatt, S. L., Pittman, J. A., & Smith, I. (1997). Unification-based multimodal integration. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* (pp. 281–288). San Francisco: Morgan Kaufmann. Karshmer, A. I., & Blattner, M. (editors). (1998). *Proceedings of the 3rd International ACM Proceedings of the Conference on Assistive Technologies (ASSETS'98)*. (URL http://www.acm.org/sigcaph/assets/assets98/assets98index.html).

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. Key (Ed.), *The Relationship of Verbal and Nonverbal Communication* (pp. 207–227). The Hague, Netherlands: Mouton.

Kobsa, A., Allgayer, J., Reddig, C., Reithinger, N., Schmauks, D., Harbusch, K., et al. (1986). Combining Deictic Gestures and Natural Language for Referent Identification. *Proceedings of the 11th International Conference on Computational Linguistics* (pp. 356–361). Bonn, Germany Proceedings of COLING '86, August 25–29, 1986, University of Bonn, Germany. Published by Institut für angewandte Kommunikations-und Sprachforschung e.V. (IKS).

Kopp, S., Tepper, P., & Cassell, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. *Proceedings of the 6th International Conference on Multimodal Interfaces*, (pp. 97–104). New York: ACM.

Koons, D., Sparrell, C., & Thorisson, K. (1993). Integrating simultaneous input from speech, gaze, and hand gestures. In M. Maybury (Ed.), *Intelligent Multimedia Interfaces* (pp. 257–276). Cambridge, MA: MIT Press.

Kricos, P. B. (1996). Differences in visual intelligibility across talkers. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines: Models, Systems and Applications* (pp. 43–53). New York: Springer Verlag.

Kumar, S., & Cohen, P. R. (2000). Towards a fault-tolerant multi-agent system architecture. *Fourth International Conference on Autonomous Agents 2000* (pp. 459–466). Barcelona, Spain: ACM Press.

Leatherby, J. H., & Pausch, R. (1992, July). Voice input as a replacement for keyboard accelerators in a mouse-based graphical editor: An empirical study. *Journal of the American Voice Input/Output Society, 11*(2).

Martin, D. L., Cheyer, A. J., & Moran, D. B. (1999). The open agent architecture: A framework for building distributed software systems. *Applied Artificial Intelligence, 13*, 91–128.

Massaro, D. W. (1996). Bimodal speech perception: A progress report. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines: Models, Systems and Applications* (pp. 79–101). New York: Springer Verlag.

Massaro, D. W., & Cohen, M. M. (1990, January). Perception of synthesized audible and visible speech. *Psychological Science, 1*(1), 55–63.

Massaro, D. W., & Stork, D. G. (1998). Sensory integration and speechreading by humans and machines. *American Scientist, 86*, 236–244.

McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., & Zhang, D. (2005). Automatic analysis of multimodal group actions in meetings, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 27*(3), 305–317.

McGee, D. (2003, June). *Augmenting environments with multimodal interaction*. Unpublished manuscript, Oregon Health & Science University.

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America, 77*(2), 678–685.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices, *Nature, 264*, 746–748.

McLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21*, 131–141.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

Meier, U., Hürst, W., & Duchnowski, P. (1996). Adaptive bimodal sensor fusion for automatic speechreading. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE-ICASSP)* (pp. 833–836). Los Alamitos, CA:IEEE Press

Morency, L.-P., Sidner, C., Lee, C., & Darrell, T. (2005). Contextual recognition of head gestures. *Proceedings of the Seventh International Conference on Multimodal Interfaces* (pp. 18–24). New York: ACM.

Morimoto, C., Koons, D., Amir, A., Flickner, M., & Zhai, S. (1999). Keeping an Eye for HCI. *Proceedings of SIBGRAPI'99, XII Brazilian Symposium on Computer Graphics and Image Processing*, 171–176.

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87*(2), 319–334.

Naughton, K. (1996). Spontaneous gesture and sign: A study of ASL signs co-occurring with speech. In L. Messing (Ed.), *Proceedings of the Workshop on the Integration of Gesture in Language & Speech* (pp. 125–134). Newark, DE:University of Delaware

Neal, J. G., & Shapiro, S. C. (1991). Intelligent multimedia interface technology. In J. Sullivan & S. Tyler (Eds.), *Intelligent User Interfaces* (pp.11–43). New York: ACM Press.

Negroponte, N. (1978, December). *The Media Room*. Cambridge, MA: MIT Press.

Neti, C., Iyengar, G., Potamianos, G., & Senior, A. (2000). Perceptual interfaces for information interaction: Joint processing of audio and visual information for human-computer interaction. In B. Yuan, T. Huang & X. Tang (Eds.), *Proceedings of the International Conference on Spoken Language Processing (ICSLP'2000), Vol. 3* (pp. 11–14). Beijing, China: Chinese Friendship Publishers.

Oliver, N., & Horvitz, E. (2005). S-SEER: Selective perception in a multimodal office activity system. *International Journal of Computer Vision and Image Understanding*.

Oviatt, S. L. (1997). Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction [Special issue on Multimodal Interfaces]*, *12*, 93–129.

Oviatt, S. L. (1999a). Mutual disambiguation of recognition errors in a multimodal architecture. *Proceedings of the Conference on Human Factors in Computing Systems* (*CHI'99*) (pp. 576–583). New York: ACM Press.

Oviatt, S. L. (1999b). Ten myths of multimodal interaction. *Communications of the ACM, 42*(11), 74–81.

Oviatt, S. L. (2000a). Multimodal system processing in mobile environments. *Proceedings of the Thirteenth Annual ACM Symposium on User Interface Software Technology* (*UIST'2000*) (pp. 21–30). New York: ACM Press.

Oviatt, S. L. (2000b). Taming recognition errors with a multimodal architecture. *Communications of the ACM, 43*(9), 45–51.

Oviatt, S. L. (2000c). Multimodal signal processing in naturalistic noisy environments. In B. Yuan, T. Huang, & X. Tang (Eds.), *Proceedings of the International Conference on Spoken Language Processing (ICSLP'2000), Vol. 2* (pp. 696–699). Beijing, China: Chinese Friendship Publishers.

Oviatt, S. L. (2002). Breaking the robustness barrier: Recent progress in the design of robust multimodal systems. *Advances in Computers 56*, 305–341.

Oviatt, S. L. (2003, September/October). Advances in robust multimodal interfaces, *IEEE Computer Graphics and Applications*, 62–68.

Oviatt, S. L., Bernard, J., & Levow, G. (1999). Linguistic adaptation during error resolution with spoken and multimodal systems. *Language and Speech, 41*(3–4), 415–438.

Oviatt, S. L., & Cohen, P. R. (1991). Discourse structure and performance efficiency in interactive and noninteractive spoken modalities. *Computer Speech and Language*, *5*(4), 297–326.

Oviatt, S. L., & Cohen, P. R. (2000, March). Multimodal systems that process what comes naturally. *Communications of the ACM, 43*(3), 45–53.

Oviatt, S. L., Cohen, P. R., Fong, M. W., & Frank, M. P. (1992). A rapid semiautomatic simulation technique for investigating interactive speech and handwriting. *Proceedings of the International Conference on Spoken Language Processing, 2*, 1351–1354.

Oviatt, S. L., Cohen, P. R., & Wang, M. Q. (1994). Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity. *Speech Communication, 15*, 283–300.

Oviatt, S. L., Cohen, P. R., Wu, L., Vergo, J., Duncan, L., Suhm, B., et al. (2000). Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions. *Human Computer Interaction, 15*(4), 263–322.

Oviatt, S. L., Coulston, R., & Lunsford, R. (2004). When do we interact multimodally? Cognitive load and multimodal communication patterns, *Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI'04)* 129–136.

Oviatt, S. L., Coulston, R., Shriver, S., Xiao, B., Wesson, R., Lunsford, R., et al. (2003). Toward a theory of organized multimodal integration patterns during human-computer interaction, *Proceedings of the International Conference on Multimodal Interfaces (ICMI'03)* (pp. 44–51). New York: ACM Press.

Oviatt, S. L., DeAngeli, A., & Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. *Proceedings of Conference on Human Factors in Computing Systems (CHI'97)* (pp. 415–422). New York: ACM Press.

Oviatt, S. L, Flickner, M., & Darrell, T. (Eds.) (2004, January). Multimodal interfaces that flex, adapt and persist, *Communications of the ACM 47*(1), 30–33.

Oviatt, S. L., & Kuhn, K. (1998). Referential features and linguistic indirection in multimodal language. *Proceedings of the International Conference on Spoken Language Processing, 6*, (pp. 2339–2342). Syndey, Australia: ASSTA, Inc.

Oviatt, S. L. & Lunsford, R. (2005). Multimodal interfaces for cell phones and mobile technology, *International Journal of Speech Technology, 8*(2), Springer & Springer-link on-line, 127–132

Oviatt, S. L., Lunsford, R., & Coulston, R. (2005). Individual differences in multimodal integration patterns: What are they and why do they exist? *Proceedings of the Conference on Human Factors in Computing Systems (CHI'05), CHI Letters* (pp. 241–249). New York: ACM Press.

Oviatt, S. L., & Olsen, E. (1994). Integration themes in multimodal human-computer interaction. In Shirai, Furui, & Kakehi (Eds.), *Proceedings of the International Conference on Spoken Language Processing, 2*, 551–554.

Oviatt, S. L., & van Gent, R. (1996). Error resolution during multimodal human-computer interaction. *Proceedings of the International Conference on Spoken Language Processing, 2*, 204–207.

Pankanti, S., Bolle, R.M., & Jain, A. (Eds.). (2000). Biometrics: The future of identification. *Computer, 33*(2), 46–80.

Pavlovic, V., Berry, G., & Huang, T. S. (1997). Integration of audio/visual information for use in human-computer intelligent interaction. *Proceedings of IEEE International Conference on Image Processing*, 121–124.

Pavlovic, V., & Huang, T. S. (1998). Multimodal prediction and classification on audio-visual features. *AAAI'98 Workshop on Representations for Multi-modal Human-Computer Interaction* (pp. 55–59). Menlo Park, CA: AAAI Press.

Pavlovic, V., Sharma, R., & Huang, T. (1997). Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(7) 677–695.

Pentland, S. (2005, March). Socially aware computation and communication, *IEEE Computer*, 63–70.

Potamianos, G., & Neti, C. (2001). Automatic speechreading of impaired speech, *Proceedings of the International Conference on Auditory-Visual Speech Processing*, 177–182.

Potamianos, G., Neti, C., Gravier, G., & Garg, A. (2003, September). Automatic recognition of audio-visual speech: Recent progress and challenges, *Proceedings of the IEEE, Vol. 91* (9). September, 2003.

Petajan, E. D. (1984). *Automatic Lipreading to Enhance Speech Recognition*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Pfleger, N. (2004). Context-based multimodal fusion. *Proceedings of the 6th International Conference on Multimodal Interfaces* (pp. 265–272). New York: ACM.

Poddar, I., Sethi, Y., Ozyildiz, E., & Sharma, R. (1998, November). Toward natural gesture/speech HCI: A case study of weather narration. In M. Turk, (Ed.), *Proceedings 1998 Workshop on Perceptual User Interfaces (PUI'98)* (pp. 1–6). San Francisco, CA.

Reithinger, N., Alexandersson, J., Becker, T., Blocher, A., Engel, R., Lockelt, M., et al. (2003). Multimodal architectures and frameworks: SmartKom: adaptive and flexible multimodal access to multiple applications, *Proceedings of the 5th International Conference on Multimodal Interfaces* (pp. 101–108). New York: ACM.

Robert-Ribes, J., Schwartz, J.-L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual, and auditory-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America, 103*(6) 3677–3689.

Rogozan, A., & Deglise, P. (1998). Adaptive fusion of acoustic and visual sources for automatic speech recognition. *Speech Communication, 26*(1–2), 149–161.

Rubin, P., Vatikiotis-Bateson, E., & Benoit, C. (Eds.). (1998). Audio-visual speech processing *Speech Communication, 26*, 1–2.

Rudnicky, A., & Hauptman, A. (1992). Multimodal interactions in speech systems. In M. Blattner & R. Dannenberg (Eds.), *Multimedia Interface Design* (pp. 147–172). New York: ACM Press.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America, 90*, 1797–1805.

Seneff, S., Goddeau, D., Pao, C., & Polifroni, J. (1996). Multimodal discourse modelling in a multi-user multi-domain environment. In T. Bunnell & W. Idsardi (Eds.), *Proceedings of the International Conference on Spoken Language Processing, Vol. 1* (pp. 192–195).

Shaikh, A., Juth, S., Medl, A., Marsic, I., Kulikowski, C., & Flanagan, J. (1997). An architecture for multimodal information fusion. *Proceedings of the Workshop on Perceptual User Interfaces (PUI'97)*, 91–93.

Sharma, R., Huang, T. S., Pavlovic, V. I., Schulten, K., Dalke, A., Phillips, J., et al. (1996, August). Speech/gesture interface to a visual computing environment for molecular biologists. *Proceedings of 13th International Conference on Pattern Recognition (ICPR 96), Vol. 3*, 964–968.

Sharma, R., Pavlovic, V. I., & Huang, T. S. (1998). Toward multimodal human-computer interface. *Proceedings IEEE, 86*(5), 853–860.

Silsbee, P. L., & Su, Q. (1996). Audiovisual sensory intergration using Hidden Markov Models. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines: Models, Systems and Applications* (pp. 489–504). New York: Springer Verlag.

Siroux, J., Guyomard, M., Multon, F., & Remondeau, C. (1995, May). Modeling and processing of the oral and tactile activities in the Georal tactile system. *International Conference on Cooperative Multimodal Communication, Theory & Applications*. Eindhoven, Netherlands.

Stork, D. G., & Hennecke, M. E. (Eds.). (1995). *Speechreading by Humans and Machines*. New York: Springer Verlag.

Suhm, B. (1998). *Multimodal interactive error recovery for nonconversational speech user interfaces*. Doctoral thesis. Fredericiana University, Germany: Shaker Verlag.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Summerfield, A. Q. (1992). Lipreading and audio-visual speech perception, *Philosophical Transactions of the Royal Society of London, Series B, 335*, 71–78.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.

Tang, A., McLachlan, P., Lowe, K., Saka, C., & MacLean, K. (2005). Perceiving ordinal data haptically under workload, *Proceedings of the Seventh International Conference on Multimodal Interfaces*, (pp. 317–324). New York: ACM.

Tomlinson, M. J., Russell, M. J., & Brooke, N. M. (1996). Integrating audio and visual information to provide highly robust speech recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE-ICASSP) Vol. 2*, 821–824.

Turk, M., & Robertson, G. (Eds.). (2000). Perceptual user interfaces. *Communications of the ACM, 43*(3), 32–70.

Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. V., & Terzopoulos, D. (1996). The dynamics of audiovisual behavior of speech. In D. G. Stork and M. E. Hennecke, (Eds.), *Speechreading by Humans and Machines: Models, Systems, and Applications, Vol. 150 of NATO ASI Series. Series F: Computer and Systems Sciences* (pp. 221–232). Berlin, Germany: Springler-Verlag.

Vo, M. T., & Wood, C. (1996). Building an application framework for speech and pen input integration in multimodal learning interfaces. *Proceedings of the International Conference on Acoustics Speech and Signal Processing (IEEE-ICASSP) Vol. 6*, 3545–3548.

Vo, M. T., Houghton, R., Yang, J., Bub, U., Meier, U., Waibel, A., et al. (1995). Multimodal learning interfaces. *Proceedings of the DARPA Spoken Language Technology Workshop.*

Wahlster, W. (1991). User and discourse models for multimodal communciation. In J. W. Sullivan & S. W. Tyler (Eds.), *Intelligent User Interfaces* (pp. 45–67). New York: ACM Press.

Wahlster, W. (2001, March). SmartKom: multimodal dialogs with mobile web users. *Proceedings of the Cyber Assist International Symposium*, 33–34.

Waibel, A., Suhm, B., Vo, M. T., & Yang, J. (1997, April). Multimodal interfaces for multimedia information agents. *Proceedings of the International Conference on Acoustics Speech and Signal Processing (IEEE-ICASSP) Vol. 1*, 167–170.

Wang, J. (1995). Integration of eye-gaze, voice and manual response in multimodal user interfaces. *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, 3938–3942.

Wickens, C. D., Sandry, D. L., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output. *Human Factors, 25*, 227–248.

Wu, L., Oviatt, S., &. Cohen, P. (1999). Multimodal integration—A statistical view. *IEEE Transactions on Multimedia, 1*(4), 334–341.

Xiao, B., Lunsford, R., Coulston, R., Wesson, R., & Oviatt, S. L. (2003). Modeling multimodal integration patterns and performance in seniors: Toward adaptive processing of individual differences, *Proceedings of the International Conference on Multimodal Interfaces (ICMI'03)* (pp. 265–272). New York: ACM Press.

Xiao, B., Girand, C., & Oviatt, S. (2002, September). Multimodal integration patterns in children. *Proceedings of the International Conference on Spoken Language Processing (ICSLP'02)*, 629–632.

Zhai, S., Morimoto, C., & Ihde, S. (1999). Manual and gaze input cascaded (MAGIC) pointing. *Proceedings of the Conference on Human Factors in Computing Systems (CHI'99)* (pp. 246–253). New York: ACM Press.