1

Multimodal Input for Perceptual User Interfaces

Joseph J. LaViola Jr. , Sarah Buchanan, and Corey Pittman

University of Central Florida

1.1 Introduction

Ever since Bolt's seminal paper, "Put that there: Voice and Gesture at the Graphics Interface", the notion that multiple modes of input could be used to interact with computer applications has been an active area of human computer interaction research (Bolt 1980). This combiniation of different forms of input (e.g., speech, gesture, touch, eye gaze) is known as multimodal interaction and its goal is to support natural user experiences by providing the user with choices in how they can interact with a computer. These choices can help to simplify the interface, provide more robust input when recognition technology is used, and support more realistic interaction scenarios because the interface can be more fine tuned to the human communication system. More formally, multimodal interfaces process two or more input modes in a coordinated manner which aim to recognize natural forms of human language and behavior and typically incorporate more than one recognition-based technology (Oviatt 2003).

With the advent of more powerful perceptual computing technologies, multimodal interfaces that can passively sense what the user is doing are becoming more prominent. These interfaces, also called perceptual user interfaces (Turk and Robertson 2000), provide mechanisms that support unobtrusive interaction where sensors are placed in the physical environment and not on the user. The prior chapters in this book have focused on various input technologies and associated interaction modalities. In this chapter, we will examine how these different technologies and their input modalities, specifically speech, gesture, touch, eye gaze, facial expressions, and brain input can be combined and the types of interactions they afford. We will also examine the strategies for combining these input modes together, otherwise known as multimodal integration or fusion. Finally, we will examine some usability issues with mutlimodal interfaces and methods for handling them. Research in multimodal interfaces spans many fields including psychology, cognitive science, software engineering,

This is a Book Title Name of the Author/Editor

and human computer interaction (Dumas et al. 2009). Our focus in this chapter will be on the types of interfaces that have been created using multimodal input. More comprehensive surveys can be found in Jaimes and Sebe (2007); Oviatt (2007).

1.2 Multimodal Interaction Types

Multimodal interaction can be defned as the combination of multiple input modalities to provide the user with a richer set of interactions compared to traditional unimodal interfaces. The combination of input modalities can be divided into six basic types: complementarity, redundancy, equivalence, specialization, concurrency, and transfer (Martin 1998). In this section, we briefly define each.

Complementarity. Two or more input modalities complement each other when they combine to issue a single command. For example, to instantiate a virtual object, a user makes a pointing gesture and then speaks. Speech and gesture complement each other since the gesture provides the information on where to place the object and the speech command provides the information on what type of object to place.

Redundancy. Two or more input modalities are redundant when they simultaneously send information to the application. By having each modality issue the same command, redundant information can help resolve recognition errors and reinforce what operation the system needs to perform (Oviatt and Vangent 1996). For example, a user issues a speech command to create a visualization tool while also making a hand gesture which signifies the creation of that tool. By providing more than one input stream, the system has a better chance of recognizing the user's intended action.

Equivalence. Two or more input modalities are equivalent when the user has a choice of which modality to use. For example, the user can create a virtual object by either issuing a voice command or picking the object from a virtual palette. The two modalities present equivalent interactions in that the end result is the same. The user can choose which modality to use based on preference (they simply like speech input over the virtual palette) or on frustration (the speech recognition is not accurate enough, thus they move to the palette).

Specialization. A particular modality is specialized when it is always used for a specific task because it is more appropriate and/or natural for that task. For example, a user wants to create and place an object in a virtual environment. For this particular task, it makes sense to have a ”pointing” gesture determine the object's location since the number of possible voice commands for placing the object is too large and a voice command cannot achieve the specifcity of the object placement task.

Concurrency. Two or more input modalities are concurrent when they issue divergent commands that overlap in time. For example, a user is navigating by gesture through a virtual environment and while doing so uses voice commands to ask questions about objects in the environment. Concurrency enables the user to issue commands in parallel; reflecting such real world tasks as talking on the phone while making dinner.

Transfer. Two input modalities transfer information when one receives information from another and uses this information to complete a given task. One of the best examples of transfer in multimodal interaction is the push-to-talk interface (Bowman et al. 2004); the speech modality receives information from a hand gesture telling it that speech should be activated.

1.3 Multimodal Interfaces

In this section, we examine how the different technologies and input modalities discussed in this book have been used as part of multimodal interaction systems. Note that although speech input is a predominant modality in multimodal interfaces, we do not have a dedicated section for it in this chapter. Rather, uses of speech are found as part of each modality's subsection.

1.3.1 Touch Input

Multi-touch devices have become more prevalent in recent years with the growing popularity of multi-touch phones, tablets, laptops, table-top surfaces and displays. As a result multi-touch gestures are becoming part of users' everyday vocabulary such as swipe to unlock or pinch to zoom. However, complex tasks, such as 3D modeling or image editing, can be difficult when using multi-touch input alone. Multimodal interaction techniques are being designed to incorporate multi-touch interfaces with other inputs such as speech to create more intuitive interactions for complex tasks.

3D Modeling and Design

Large multitouch displays and table-top surfaces are marketed as natural interfaces that foster collaboration. However, these products often target commercial customers in public settings and are more of a novelty item. Thus, the question remains on whether they provide utility as well as unique experiences. Since the mouse and keyboard are no longer available, speech can provide context to operations where WIMP paradigms were previously used such as in complex engineering applications (e.g. Auto-CAD). For instance, MozArt (Sharma et al. 2011) combines a tiltable multi-touch table with speech commands to provide an easier interface to create 3D models as shown in Figure 1.1. A study was conducted evaluating MozArt versus a multi-touch CAD program with novice users. The majority of users preferred the multimodal interface, although a larger number of users would need to be tested in order to evaluate efficiency and accuracy. Similar interfaces could be improved upon by using speech and touch, as stated in work on a multitouch only interface for performing 3D CAD operations (Radhakrishnan et al. 2012).

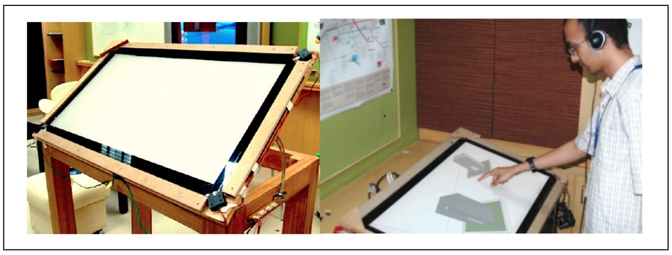


Figure 1.1 MozArt Table hardware prototype. Reproduced by permission of Anirudh Sharma.

Collaboration

Large multi-touch displays and tabletop surfaces are ideal for collaboration since they have a 360 degree touch interface, a large display surface, and allow many input sources. For instance, Tse et al. (2008) created a multimodal multi-touch system that lets users gesture and speak commands to control a design application called The Designer's Environment. The Designer's Environment is based on the KJ creativity method that industrial designers use for brainstorming. The four steps of the KJ creativity method are: (1) create notes, (2) group notes, (3) label groups, and (4) relate groups. In The Designer's Environment multiple users can complete these tasks by using a combination of touch, gesture, and speech input as shown in Figure 1.3. However, there are some obstacles as explored by (Tse et al. 2008): parallel work, mode switching, personal and group territories, and joint multimodal commands. Tse et al. propose solutions to these issues, such as allowing for parallel work by creating personal work areas on the surface.

Tse et al. (2006) also created the GSI Demo (Gesture and Speech Infrastructure created by Demonstration). This system demonstrates multimodal interaction by creating a multi-user speech and gesture input wrapper around existing mouse/keyboard applications. The GSI Demo can effectively convert a single user desktop application to a multi-touch table application, such as maps, command and control simulations, simulation and training, and games. Tse et al. (2007) specifically discuss how playing Blizzard's Warcraft III and The Sims can become a collaborative effort on a multi-touch tabletop system. Their proposed interface allows players to use gesture and speech input to create a new and engaging experience that is more similar to the social aspect provided by arcade gaming, shown in Figure 1.2.

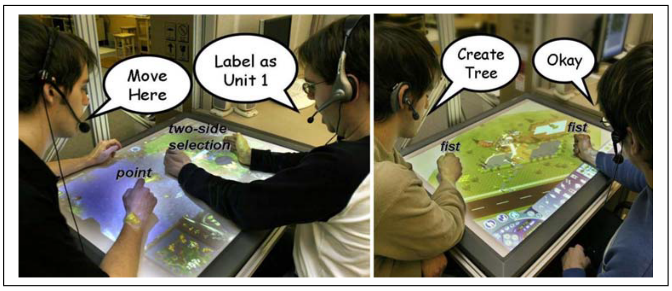


Figure 1.2 Two people interacting with Warcraft III (left) and The Sims game (right). Reproduced by permission of Edward Tse.

Another interesting aspect of collaborative environments is how to track who did or said what during a collaboration effort. Collaboration data can then shed light on the learning or collaboration process. This type of data can also act as input to machine learning or data mining algorithms providing adapted feedback or personalized content. Collaid (Collaborative Learning Aid) is an environment that captures multimodal data about collaboration in a tabletop learning activity (Martínez et al. 2011). Data is collected using a microphone array and a depth sensor, integrated with other parts of the learning system, and finally transformed into visualizations showing the collaboration processes that occurred at

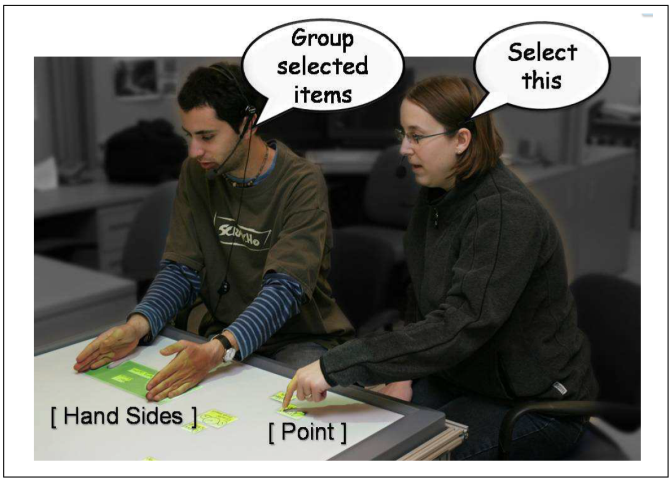


Figure 1.3 A two person grouping hand gesture in the Designer's Environment. Reproduced by permission of Edward Tse.

the table. An example visualization of data collected from a collaborative group versus a less collaborative group is shown in Figure 1.4. Other work on multimodal collaboration using distributed whiteboards can be found in Barthelmess et al. (2005).

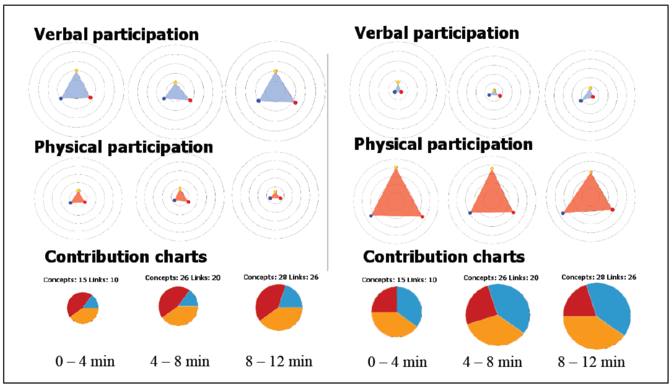


Figure 1.4 A collaborative visualization corresponding to 12 minutes of activity from a communicative group (left) and a less collaborative group (right). Reproduced by permission of Roberto Martinez.

Communication with Disabled or Elderly Patients

Multimodal surface applications have also been shown to support medical communication with hearing impaired patients (Piper and Hollan 2008) and is being explored as way to

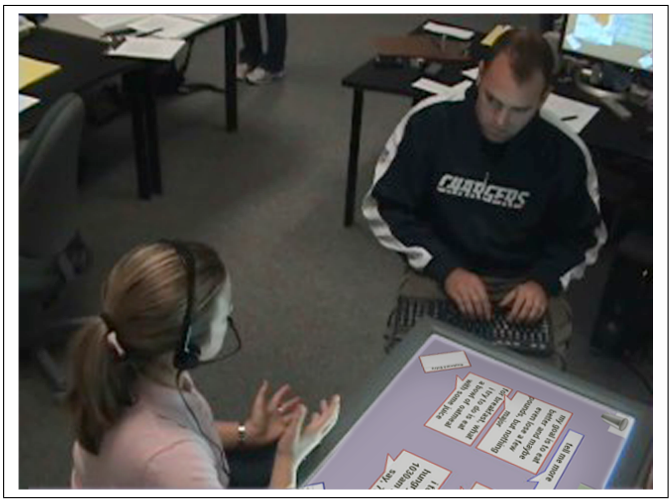


Figure 1.5 A doctor (left) and patient (right) communicate using the Shared Speech Interface (SSI). Meanwhile, movable speech bubbles appear on the multi-touch surface. Reproduced by permission of Ann Marie Piper.

enhance medical communication with the elderly by transcribing speech (Piper 2010). The Shared Speech Interface (SSI) is an application for an interactive multitouch tabletop display designed to facilitate medical communication between a Deaf patient and a hearing, non-signing physician. The Deaf patient types on a keyboard, and the hearing doctor speaks into a headset microphone. As the two people communicate, their speech is transcribed and appears on the display in the form of movable speech bubbles, shown in Figure 1.5. Multimodal surface technology can also benefit other populations with different communication needs. For example, a student learning a foreign language could access text-based representations of speech along with audio clips of the instructor speaking the phrases.

Mobile Search

Mobile users are becoming more tech-savy with their devices and expect to be able to use them while multitasking or on-the-go, such as when driving or for quick access of information. In addition, contemporary mobile devices incorporate a wide range of input technologies such as a multi-touch interface, microphone, camera, GPS, and accelerometer. Mobile applications need to leverage these alternate input modalities in a way that doesn’t require users to stop what they are doing, but instead provides fast, on-the-go input and output capabilities. Mobile devices also face their own challenges such as slower data transfer, smaller displays, smaller keyboards, and thus have their own design implications, making desktop paradigms even less appropriate. Some believe that voice input is an easy solution to these problems. However, using voice input alone is not feasible, since voice recognition is error-prone especially in noisy environments and does not provide fine-grained control. Many multimodal mobile interfaces are emerging that use voice input combined with other interactions in a clever way. For example, voice input can be used to bring context to the desired operation, while leaving fine grained manipulation to direct touch manipulation. Or voice input can be used at the same time as text entry to ensure text is entered correctly.

Voice input is ideal for mobile search since it is quick and convenient. However, since voice input is prone to errors, error correction should be quick and convenient as well. The Search Vox system uses multimodal error correction for mobile search (Paek et al. 2008). After a query is spoken, the user is presented with an N-best list of recognition results for a given query. The N-best results comprise a word palette that allows the user to conveniently rearrange and construct new queries based on the results by using touch input. Another subset of mobile search, local search, is desirable in mobile contexts, allowing searches to be constrained to the current location. Speak4it is a mobile app that takes voice search a step further by allowing users to sketch with their finger the exact area they wish to search in (Ehlen and Johnston 2011). Speak4it supports multimodal input by allowing users to speak or type search criteria and sketch the desired area with touch input. An example scenario of Speak4it would be a biker searching the closest repair shop along a trail, by using speech and gesture to get a more refined search. For instance, the query by voice “bike repair shops near the golden gate bridge” combined with a route drawn on the display, will return the results along the specified route traced on the display (Figure 1.6). Research prototypes with these capabilities have existed for many years, such as QuickSet (Cohen et al. 1997a). However, they have not been available to everyday users until the widespread adoption of mobile devices that come equipped with touch screen input, speech recognition, and mobile internet access. Other work that has explored multimodal interaction for mobile search is the Tilt and Go system by Ramsay et al. (2010). A detailed analysis of speech and multimodal interaction in mobile search is presented in Feng et al. (2011).

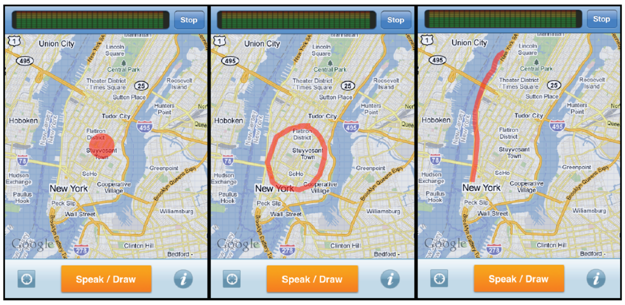


Figure 1.6 Speak4it gesture inputs. Reproduced by permission of Patrick Ehlen.

Mobile Text Entry

Typing on a touchscreen display using a soft keyboard remains the most common, but time consuming, text input method for many users. Two resolutions for faster text entry are gesture keyboards and voice input. Gesture keyboards circumvent typing by allowing the user to swipe over the word path quickly on a familiar QWERTY keyboard, however gestures can be ambiguous for prediction. Voice input offers an attractive alternative that completely eliminates the need for typing. However, voice input relies on automatic speech recognition technology which has poor performance in noisy environments or for non-native users. Speak-As-You-Swipe (SAYS) (Sim 2012) is a multimodal interface that integrates a gesture keyboard with speech recognition to improve the efficiency and accuracy of text entry, shown in Figure 1.7. The swipe gesture and voice inputs provide complementary information for word prediction, allowing the SAYS system to intelligently extract useful cues from the ambient sound to improve the word prediction accuracy. In addition, SAYS builds on previous work (Kristensson and Vertanen 2011) by enabling continuous and synchronous input.

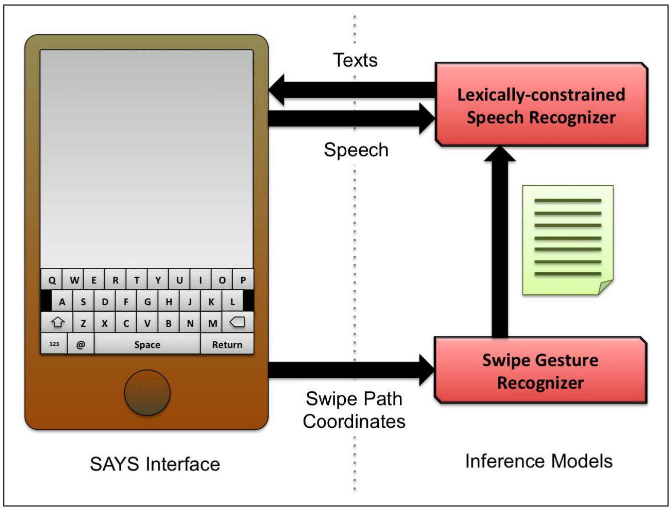


Figure 1.7 Speak As You Swipe (SAYS) interface. Reproduced by permission of Sim Khe Chai.

A similar interface created by Shinoda et al. (2011), allows semi-synchronous speech and pen input for mobile environments, shown in Figure 1.8. There is an inherent time lag between speech and writing, making it difficult to apply conventional multimodal recognition algorithms. To handle this time-lag, they developed a multimodal recognition algorithm that used a segment-based unification scheme and a method of adapting to the time-lag characteristics of individual users. The interface also supports keyboard input and was tested in a variety of different manners: (1) the user writes the initial character of each phrase in a saying, (2) the user writes the initial stroke of the initial character as in (1), (3) the user inputs a pen touch to cue the beginning of each phrase, (4) the user taps the character table to which the first character of each phrase belongs to, and (5) the user inputs the initial character of each phrase using a QWERTY keyboard. These five different pen-input interfaces were evaluated using noisy speech data, and the recognition accuracy of the system was higher than that of speech alone in all five interfaces. They also conducted a usability test for each interface, finding a trade-off between usability and improvement in recognition performance. Other work that compares different multimodal interaction strategies using touch input for mobile text entry can be found in Dearman et al. (2010).

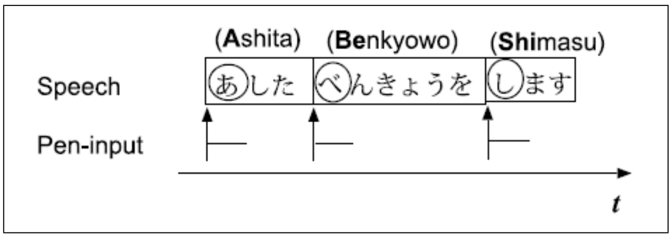


Figure 1.8 Relationship between speech and pen input. Reproduced by permission of Koichi Shinoda.

Mobile Image Editing

Another mobile app that incorporates multimodal inputs in novel way is PixelTone (Laput et al. 2013). PixelTone is a multimodal photo editing interface that combines speech and direct manipulation to make photo editing easier for novices and when using mobile devices. PixelTone uses natural language for expressing desired changes to an image, and direct manipulation to localize these changes to specific regions as shown in Figure 1.9. PixelTone does more than just provide a convenient interface for editing photos. The interface allows for ambiguous commands that a novice user would make such as “Make it look good”, as well as advanced commands such as “Sharpen the midtones at the top”. Although users performed the same using the multimodal interface versus just a touch interface, they preferred the multimodal interface overall and were able to use it effectively for a realistic workload.

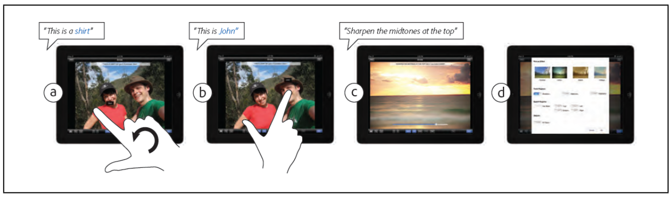


Figure 1.9 PixelTone. Reproduced by permission of Gierad Laput.

Automobile Control

Although a majority of US states ban text messaging for all drivers, there are still a variety of activities drivers attempt to complete during their commute. Research is being conducted to help drivers be able to achieve the primary task of driving effectively while still completing tertiary tasks such as navigating, communicating, adjusting music and controlling climate (Muller and Weinberg 2011). The Society of Automotive Engineers recommends that any tertiary task taking more than 15 seconds to carry out while stationary be prohibited while the vehicle is in motion. Voice controls are exempt from the 15 second rule since they don’t require the user to take there eyes off the road and may appear to be an obvious solution. However, some data suggests that certain kinds of voice interfaces impose high cognitive loads and can negatively affect driving performance. This negative affect is due to the technical limitations of the speech recognition, or to usability flaws such as confusing or inconsistent command sets and unnecessarily deep and complex dialog structures. Multimodal interfaces may be a way to address these problems by combining the best modes of input depending on the driving situation.

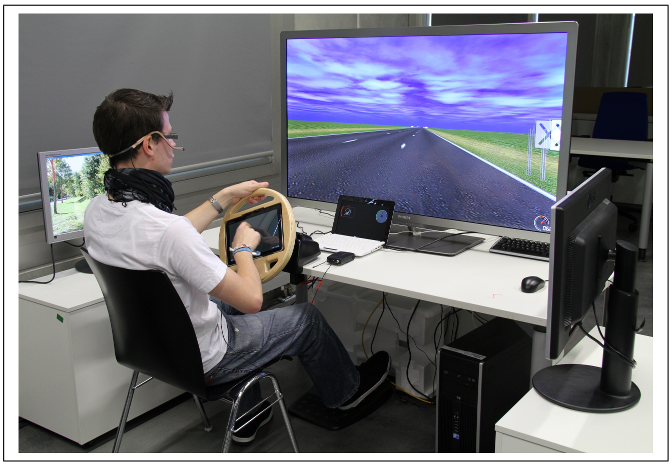


Figure 1.10 Multimodal automobile interface combining speech with gestures on the steering wheel. Reproduced by permission of Bastian Pfleging.

Incorporating speech, touch, gesture, and haptic touchpad input have been explored separately as input to a driving interface. But none of these inputs alone create an ideal solution. A multimodal interface created by Pfleging et al. (2012), combines speech with gestures on the steering wheel to minimize driver distraction, shown in Figure 1.10. Pfleging et al. point out that speech input alone doesn't have fine grained control, while touch input alone requires too much visual interaction, and gesture input alone doesn't scale well (Pfleging et al. 2011). They propose a multimodal interaction style that combines speech and gesture where voice commands select visible objects or functions (mirror, window, etc.) and simple touch gestures are used to control these functions. With this approach recalling voice commands is simpler as the users see what they need to say. By using a simple touch gesture, the interaction style lowers the visual demand and provides at the same time immediate feedback and easy means for undoing actions. Other work that looks at multimodal input for autombile control can be found in Gruenstein et al. (2009).

1.3.2 3D Gesture

Devices like Microsoft's Kinect and Intel's Perceptual Computing camera have steadily become more widespread and have provided for new interaction techniques based on 3D gestures. Using a combination of depth cameras and standard RGB cameras, the devices allow for accurate skeletal tracking and gesture detection. These 3D gestures can be used to accomplish a number of tasks in a more natural way than WIMP interfaces. In order to enrich the user's experiences, these gestures can be combined with different modalities, such as speech and face tracking. Interestingly enough, the Kinect and Perceptual Computing camera have built in microphones, making them ideal for designing multimodal interfaces. Another common technique for gesture detection is to use stereo cameras to detect gestures while using machine learning and filtering to properly classify them. Prior to the development of these technologies, the use of 3D gesture in multimodal interfaces was limited to detecting deictic gestures for selection or similar tasks using simple cameras. Speech is one of the more common modalities to be combined with gesture, as modern 3D gestures tend to involve a large portion of the body and limit what other modalities could realistically be used simultaneously (Jaimes and Sebe 2007). Numerous applications of the Kinect sensor have been realized, with a number of games developed that have used both speech and gesture developed by Microsoft's first party developers. Outisde of the game industry, sensors like the Kinect have been used for simulation and as a means to interface with technology using more natural movements. Some work has been done in the fields of human-robot interaction (HRI) and medicine for the utility of hands-free, gesture-controlled applications.

Games and Simulation

Interacting with a simulation using body gestures is commonly used in virtual environments when combined with speech to allow for simultaneous inputs. Williamson et al. (2011) developed a system combining the Kinect, speech commands, and a Sony Playstation Move controller to make a full-body training simulation for soldiers 1.11. The RealEdge prototype allows users to march through an environment using a marching gesture and make small movements of their virtual avatar by leaning. The user can also look around the environment using the Move controller, which is attached to a weapon-like apparatus. The user may also use voice commands to give commands to virtual characters. One shortcoming of depth camera based gesture recognition devices currently available is that they generally require the user to be facing them to accurately track the skeleton of the user. The RealEdge Fusion System, an extension of the RealEdge prototype, allows for accurate 360 degree turning by adding multiple Kinects around the user and fusing retrieved skeletal information at the data level to allow for robust tracking of the user within the range of the sensors regardless of the users orientation (Williamson et al. 2012). The skeletal tracking information is passed from a Kinect depth sensor through a client laptop to a server where the data is then fused The system only requires the addition of more Kinect sensors, laptops, and a head mounted display to give the user the proper view of their environment.

A large amount of research has been done emphasizing the addition of speech to 3D interfaces for segmentation and selection. Budhiraja et al. specify that a problem with deictic gesture alone is that large numbers of densely packed or occluded objects make selection difficult (Budhiraja and Madhvanath 2012). In order to solve this problem, the authors add speech as a modality to help specify defining attributes of the desired object such as spatial location, location relative to other objects, or physical properties of the object. This allowed for specific descriptions, such as "the blue one to the left," to be used to clarify what object the user aims to select. Physical attributes and locations must be clearly defined for proper identification.

There are instances where 3D gesture is not the primary method of control in an interface.

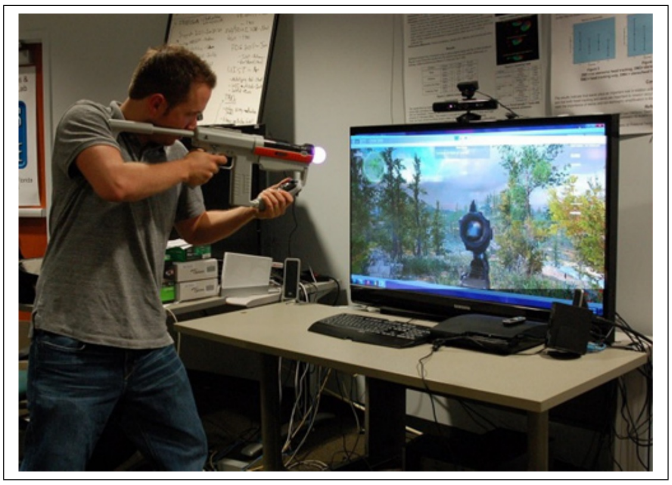


Figure 1.11 RealEdge prototype with Kinect and PS Move Controller. Reproduced by permission of Brian Williamson.

The SpeeG input system is a gesture based keyboard replacement that uses speech augmented with a 3D gesture interface (Hoste et al. 2012). The interface is based on the Dasher interface (Ward et al. 2000) with the addition of speech and hand gestures, which replace the mouse. The system uses intermediate speech recognition to allow the user to guide the software to the correct phrase using their hand. Figure 1.12 shows an example of the virtual environment and the pointing gesture. Though the prototype did not allow for real-time input due to problems with speech recognition latency, users found that SpeeG was ”the most efficient interface” when compared to an onscreen keyboard controlled with a Microsoft Xbox 360 controller, speech control only, and Microsoft Kinect keyboard only.

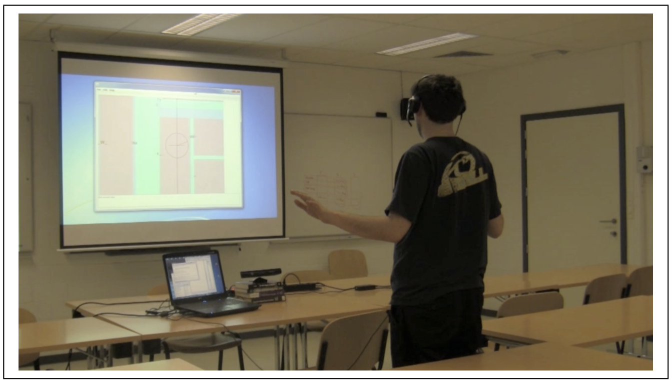
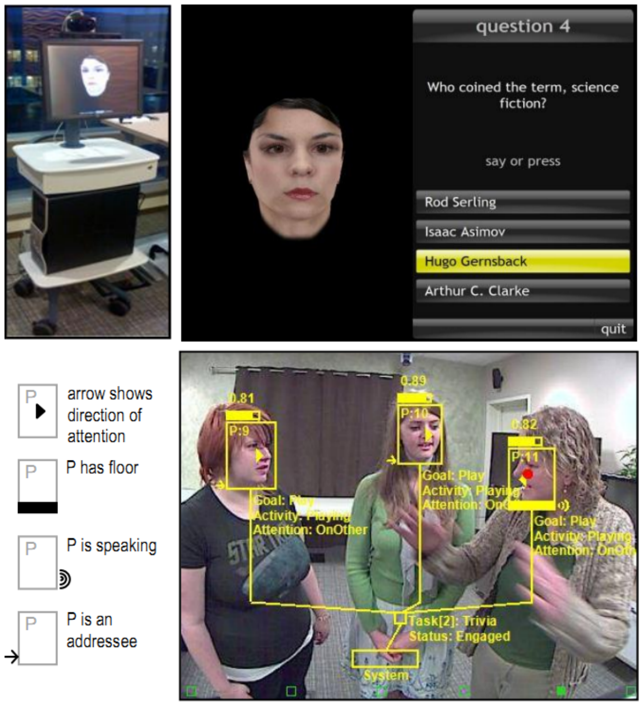


Figure 1.12 SpeeG interface and example environment. Reproduced by permission of Lode Hoste.

3D gestures are not limited to full body movements. Bohus and Horvitz (2009) developed a system to detect head pose, facial gestures, and a limited amount of natural language in a conversational environment. Head position tracking and gaze estimation were done using a basic wide-angle camera and commercially available software. A linear microphone is used to determine the sources of user voices. These modalities were fused and analyzed and a suitable response was given to the users. This multiparty system was used for observational studies of dialogues, in which the system asks quiz questions and waits for responses from the users. Upon receiving a response from the user, the system will vocally ask for confirmation from the users. Depending on their responses, the system will respond appropriately and either continue on or ask for another answer. The system's behavior is based on a turn-taking conversational model, for which the system has four behaviors: Hold, Release, Take, and Null. An example of this system functioning can be seen in Figure 1.13.



P_9, P_{10}, and P_{11} are all engaged in playing a trivia game with the system. The system is currently looking towards P_{11}, as shown by the red dot. The participants' focus of attention is directed towards each other. P_{11} has the floor and is currently speaking to P_9 and P_{10}.

Figure 1.13 Example of system for Turn-taking model for multiparty dialogues. Used with permission from Microsoft Corporation.

Hrúz et al. (2011) designed a system for multimodal gesture and speech regcognition for communication between two disabled users, one hearing impaired and one vision impaired.

The system recognizes signed language from one of the users using a pretrained sign recognizer. The recognizer uses a single camera for capture of the hand signs, so the user must wear dark clothes to create contrast between the hands and the background and allow for more accurate detection of the signs The signs are converted to text and then played for the other user using a language specific text-to-speech systems. The other user speaks to the system, which uses automatic speech recognition to translate the spoken words to text, which is then displayed for the other user. Each one of the users represents a separate input modality for the system architecture, which is illustrated in Figure 1.14. The system allows for communication between two users who otherwise would be unable to find a medium to speak to one another.

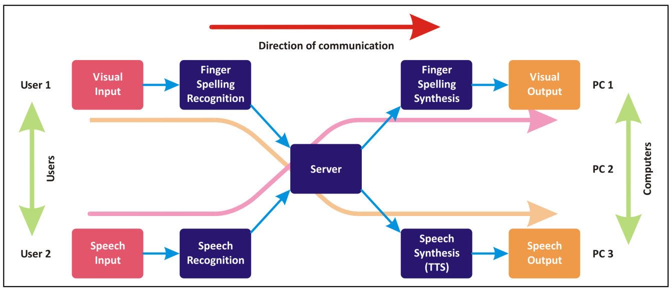


Figure 1.14 Flow of system for communication between two disabled users. Reproduced by permission of Marek Hrúz.

Medical Applications

Gallo et al. (2011) developed a system for navigating medical imaging data using a Kinect. The user is able to traverse the data as though it were a virtual environment using hand gestures for operations like zooming, translating, rotating, and pointing. The user is also able to select and extract regions-of-interest from the environment. The system supports a number of commonly used imaging techniques, including CT, MRI, and PET images. The benefit of having a 3D gesture interface in medicine is that there is no contact with any devices, maintaining a sterile environment. This interface can therefore be used in surgical environments to reference images without having to repeat a sterilization process. The sterile environment problem arises often in computerized mechanical systems. Typically, there is an assistant in the operating room who controls the terminals that contain information about the surgical procedure and images of the patient. These assistants typically do not have the same level of training as the surgeon in charge and may misinterpret information the surgeon could have correctly deciphered.

Human-Robot Interaction

Perzanowski et al. (2001) designed a method of communicating with robots in a way that was natural for humans already: using natural language and gestures. A stereo camera attached to the robot watched the hands of the user and determined if the gestures were meaningful. Combining these gestures with the voice input from the user, the robot would carry out the desired commands from the user. The speech commands could be as simple as "Go over there" or "Go this far." Along with the information the gesture gives, which could be something like a position in the environment from a pointing gesture or a distance from a holding hands spread apart, the robot is able to make intelligent decisions on how far to move. The robot was also capable of being controlled using a PDA that mapped out the environment. The user could combine the PDA commands with speech commands or gesture commands or be used in place of speech and/or gesture. The tasks given to robots could be interrupted and the robot would eventually return to it's original task to complete it. Similar interfaces that combine 3D gestures and speech for controlling robots include (Burger et al. 2012; Stiefelhagen et al. 2004).

Consumer Electronics

Multimodal input technologies have matured to the point where they are becoming interfaces to consumer electronics devices, particularly large screen displays and televisions. One commercial example is Samsung's Smart TV series which includes 3D gestures, speech input, and facial recognition. In the research community, Lee et al. (2013) combined 3D hand gestures and face recognition to interact with Smart TVs. Facial recognition was used as a viewer authentication system and the 3D hand gestures are used to control volume and change channels. Takahashi et al. (2013) also explored 3D hand gestures coupled with face tracking using a ToF depth camera to assist users in TV viewing. Krahnstoever et al. (2002) constructed a similar system for large screen displays that combined 3D hand gestures, head tracking, speech input, and facial recognition. This system was deployed in a shopping mall to assist shoppers in finding appropriate stores. As the technology continues to improve and get smaller and less expensive, these types of multimodal interfaces will be integrated into more consumer electronic devices, such as PCs and laptops, as their main user interface.

1.3.3 Eye Tracking and Gaze

The ability to track a user's eyes in order to know their gaze location has been an important area of research over the last several years (Duchowski 2007) as there are a variety of commercial eye tracking and gaze determination devices used in many different applications from video games to advertising analysis. One of the more common uses of eye gaze in a multimodal input system has been for object selection by either coupling eye gaze with keyboard button presses (Kumar et al. 2007; Zhang and MacKenzie 2007) or using eye gaze as a course selection technique coupled with a mouse for finer pointing (Hild et al. 2013; Zhai et al. 1999).

Eye gaze in multimodal interfaces include integration with hand gestures for multi-display environments with large area screens (Cha and Maier 2012). Coupling eye gaze with a variety of different modalities including a mouse scroll wheel, tilting a handheld device, and touch input on a smart phone have also be used for pan and zoom operations for large information spaces (Stellmach and Dachselt 2012). Text entry has been explored using eye

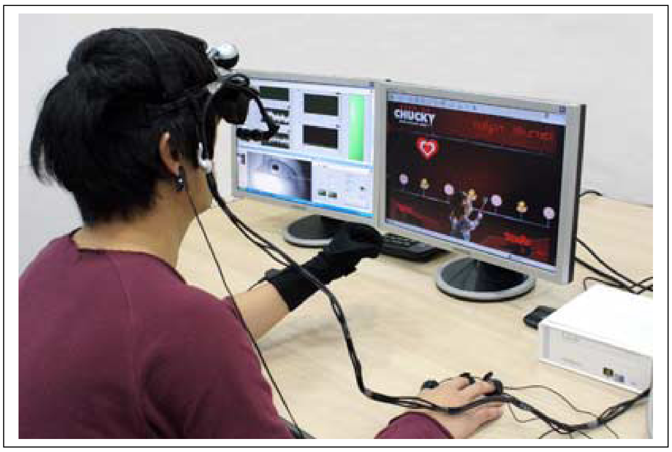


Figure 1.15 A video game that combines eye gaze, hand gestures, and bio-feedback. Reproduced by permission of Hwan Heo.

gaze by combining it with speech input (Beelders and Blignaut 2012). In this interface, users would focus on a character of interest and issue a voice command to type the character in a document, although in usability testing a traditional keyboard was found to be faster and more accurate. Multimodal interfaces making use of eye gaze have also be developed in the entertainment domain. For example, Heo et al. (2010) developed a video game (see Figure 1.16) that combined eye gaze, hand gestures, and bio-signal analysis and showed the multimodal interface was more compelling than traditional keyboard and mouse controls. Finally, eye gaze has recently been explored in combination with Brain-Computer input (BCI) to support disabled users (Vilimek and Zander 2009; Zander et al. 2010a). In this interface, eye gaze was used to point to objects with the BCI component was used to simulate dwell time for selection. Integrating eye gaze with BCI for multimodal interaction makes intuitive sense, given that movement is often limited when using BCI (see Section 1.3.5).

1.3.4 Facial Expressions

Recognizing facial expressions is a challenging and still open research problem that can be an important component of perceptual computing applications. There has been a significant amount of research on facial expression recognition in the computer vision community (Sandbach et al. 2012). In terms of multimodal interaction, facial expressions are commonly used in two ways. In the first way, other modalities are integrated with facial expression recognizers to better improve facial expression recognition accuracy which ultimately supports human emotion recognition. For example, De Silva et al. (1997) combined both video and audio information to determine what emotions are better detected with one type of information or another. They found that anger, happiness, surprise and dislike were better recognized by humans with visual information while sadness and fear tended to be easier to recognize using audio. Busso et al. (2004) found similar results by integrating speech and facial expressions together in a emotion recognition detection system.

Kessous et al. (2010) used facial expressions and speech as well as body gestures to detect human emotions in a multimodal-based recognizer. Another example of an emotion detection system that couples facial expressions and speech input can be found in (Wöllmer et al. 2010).

The second way facial expression recognition has been used in the context of multimodal interfaces is to build affective computing systems which determine emotional state or mood in an attempt to adapt an application's interface, level of difficulty, and other parameters to improve user experience. For example, Lisetti and Nasoz (2002) developed the MAUI system, a multimodal affective user interface that perceives a user's emotional state combining facial images, speech and biometric feedback. In another example, Caridakis et al. (2010) developed an affective state recognizer using both audio and visual cues using a recurrent neural network. They were able to achieve recognition rates as high as 98%, which shows promise for multimodal perceptual computing systems that can not only observe and understand the user when they are actively issuing commands, but when they are being passively monitored as well.

1.3.5 Brain Computer Input

Modern brain-computer interfaces (BCI) are capable of monitoring our mental state using nothing more than an electroencephalography (EEG) connected to a computer. In order for signals to be tracked using today's technologies, a number of electrodes must be connected to the user's head in specific locations. These connections limit the possibilities for the addition of certain modalities to systems that the user can interact with. If the user moves while wearing a BCI, there is typically some noise in the signal, decreasing the accuracy of the signal. This is often not a large issue, as BCIs have typically been used to facilitate communication and movement from disabled individuals. A number of companies, including Emotiv (Figure 1.16) and Neurosky, have begun developing low cost BCIs for previously unconventional applications, one of which is video games. With the decrease in the cost of consumer BCIs, a number of multimodal applications using them have been proposed. Currently the most commonly used modalities with BCIs are speech and eye gaze, as these don't require significant movement of the body. Gürkök and Nijholt (2012) cite numerous examples of BCIs being used to improve user experience and task performance by using a brain-control interface as a modality within multimodal interfaces.

Electroencephalography (EEG) is often combined with additional neuroimaging methods, such as electromyography (EMG), which measures muscle activity. Leeb et al. (2010) showed significant improvement in recognition performance when EEG is used in conjunction with EMG when compared to either EEG or EMG alone. Bayesian fusion of the two signals was used to generate the combined signal. EEG combined with NIRS (Near-infrared spectroscopy) has also been shown to significantly improve classification accuracy of signals (Fazli et al. 2012). NIRS creates a problem for real-time BCIs because of its significant latency.

Gürkök et al. (2011) studied the effect of redundant modalities on a user's performance in a video game of their own design. The game, titled "Mind the Sheep!," required the player to move a group of dogs around to herd sheep into a pen. The system setup is shown in Figure 1.17. The game used a mouse in conjunction with one or two other modalities. The

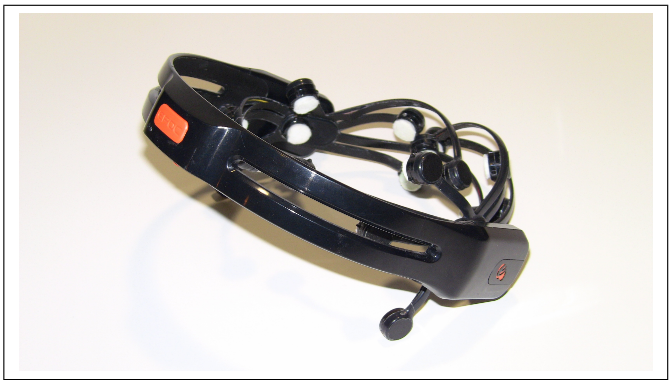


Figure 1.16 Emotiv EEG neuroheadset. Reproduced by permission of Corey Pittman.

participant controlled which dog was selected using either speech or a BCI. To select a dog with speech, participants needed only to say the name of the dog. To control the dog with the BCI, the participant was required to focus on the desired dog's location and then release the mouse button on the location where the dog was to move. Participants were asked to play the game under multiple conditions: with automatic speech recognition (ASR), with BCI, and lastly with both in a multimodal mode. The study found that being given the opportunity to select the current modality did not give a significant performance increase when compared to either of the single modality modes, with some participants not even changing modalities once.

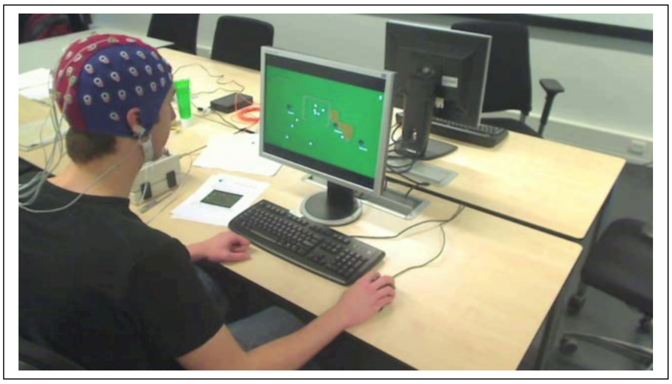


Figure 1.17 "Mind the Sheep!" system interface. Reproduced by permission of Hayrettin Gürkök.

One further application for BCI's is modeling interfaces. Sree et al. (2013) designed a framework for using BCI as an additional modality to assist in 3D modeling. EEG and EMG are again combined, this time in conjunction with a keyboard and mouse to control the modeling operations, with the Emotiv EEG neuroheadset as the primary device. The software accompanying the Emotiv is used to set some parameters for the device's signals and calibrate the device for each particular user. Facial movements are detected using the EMG component, with gestures like looking left controlling drawing an arc or blinking to left click on the mouse. Mouse movement is controlled using the EEG component, which detects the intent of the user. The Emotiv API allows for interpretation of twelve movements, six directional movements and six rotational movements which can all be used in the CAD environment. Participant fatigue was a common problem with the system, along with some problems with EEG signal strength. Additional input modalities, such as speech, could be added to this system to allow for improved usability.

Zander et al. (2010b) allowed users the freedom to control a BCI using either imagined movement, visual focus, or a combination of both. The authors proposed that since BCI does not always work for all users when there is only one method of control, providing alternatives or hybrid techniques allows for significant improvement in accuracy. Maye et al. (2011) present a method for increasing the number of distinct stimuli (different tactile and visual stimuli) when using a BCI increases the number of the dimensions the user can control, while maintaining similar mental effort. Brain activity can be more easily classified by asking the user to shift focus between different stimuli. Zander et al. (2010b) separate BCIs that are used in HCI into three categories: active, reactive, and passive.

1.4 Multmodal Integration Strategies

One of the most critical parts of a multimodal interface is the integration component, often called the fusion engine, that combines the various input modalities together to create a cohesive interface that supports meaningful commands. (Lalanne et al. 2009). There are many technical challenges with building multimodal integration engine for several reasons. First, the different input modalities may have very different characteristics including data formats, frequencies, and semantic meaning, making it difficult to combine them together. Second, timing plays a significant role in integration since different modalities may come at different times during an interaction sequence, requiring the integration engine to be flexible in its ability to process the input streams. Third, and related to timing, resolving ambiguities is a significant challenge when the integration engine is under-constrained, the engine does not have enough information to make a fusion decision, or over-constrained, the engine has conflicting information and has more than one possible fusion decision to make. Finally, the input modalities used in multimodal interfaces often stem form natural communication channels (e.g., 3D gesture, speech, facial expressions) where recognition technologies are required to segment and classify the incoming data. Thus, levels of probabilistic uncertainty will exist with these modalities, making the integration engine's job more complex.

There are two basic approaches to performing multimodal integration in the fusion engine. The first is early integration and the second is late integration and within these two approaches there are several different integration strategies (Turk 2013). The premise behind early integration is the data is first integrated before any major processing is done (aside for any low-level processing). In contrast to early integration, late integration processes the data from each mode separately and interprets it as if it is unimodal before integration occurs. The advantages of late integration is that since the input modalities can be analyzed individually, time synchronization is not an issue and software development tends to be simpler. However, late integration suffers from a potential loss of information in that it can miss potential cross-modal interaction. For example, in early integration the results from a gesture recognizer could be improved based on the results from a speech recognizer or vice versa. However, with late integration, each individual recognizer must make a decision about its modality independently. The choice of whether to use early or late integration is still an open research question and depends on the types of modalities used as well as the multimodal interaction styles supported by an application. Note that in some cases, a compromise between early and late integration can be used to perform the multimodal integration. For example, taking an early integration of 3D gestures and eye gaze with late integration with speech.

Within the context of early and late integration, there are three different integration levels for any incoming data stream (Sharma et al. 1998). Data-level and feature-level integration both fit into the early integration approach. Data-level integration focuses on low-level processing and is often used when the input modalities are similar such as lip and facial expressions. This type of processing is also used when there is minimal integration required. Since it is closest to the raw data source, it can provide the highest level of detail but is susceptible to noise. Feature-level integration is used when the modalities are closely coupled or time synchronized. Example modalities would be speech recognition from voice and lip movements and example strategies include neural networks and hidden Markov models. Feature-level integration is also less sensitive to noise but does not provide as much detail as low-level integration. Decision-level integration (i.e., dialog level fusion (Lalanne et al. 2009)) is a late integration approach and is the most common type of multimodal integration. Its main strength is its ability to handle loosely coupled modalities (e.g., touch input and speech) but relies on the accuracy of the processing already done on each individual modality. Frame-based, unification-based, procedural, and symbolic/statistical integration are the most common integration strategies under decision-level integration.

1.4.1 Frame-based Integration

Frame-based integration focuses on data structures that support attribute-value pairs. These frames collect these pairs from the various modalities to make an overall interpretation. For example, for speech input, an attribute-value pair could be "operation" with possible values of "delete", "add entry", "modifiy entry", etc... Each frame supports an individual input modality and the integration occurs as the union of the sets of values in the individual frames. Scores are assigned to each attribute and the overall score from the integration determine the best course of action to take. Koons et al. (1993) were one of the first groups to explore feature-based integration that combined 3D gestures, eye gaze, and speech. More recently, Dumas et al. (2008) developed the HephaisTK multimodal interface toolkit which includes the frame-based approach to multimodal integration. Other multimodal interfaces that make use of frame-based integration using a variety of different input modalities include (Bouchet et al. 2004; Nigay and Coutaz 1995; Vo and Wood 1996).

1.4.2 Unification-based Integration

The main idea behind unification-based integration is the use of the unification operator. Taken from natural language processing (Calder 1987), it determines the consistency of two pieces of partial information, and if consistent, combines them into a single result (Johnston et al. 1997). For example, Cohen et al. (1997b) was the first to use unification coupled with typed feature structures to integrate pen gesture and speech input in the QuickSet system. More recently Taylor et al. (2012) chose a unification integration scheme for combining speech and 3D pointing gestures and speech with touch gestures to support interaction with an unmanned robotic vehicle while Sun et al. (2006) also used unification coupled with a multimodal grammar syntax in a traffic management tool that used 3D gestures and speech. Unification tends to work well when fusing only two modalities at any one time and most unification-based integration research tends to focus on input pairs. Other examples of multimodal integration based on unification can be found in (Holzapfel et al. 2004; Pfleger 2004).

1.4.3 Procedural Integration

Procedural integration techniques explicitly represent the multimodal state space through algorithmic management (Lalanne et al. 2009). Common example representations using procedural integration include augmented transition networks and finite state machines. For example, both Neal et al. (1989) and Latoschik (2002) made use of augmented transition networks and Johnston and Bangalore (2005) and Bourguet (2002) used finite state automata for procedural integration. Other approaches to procedural integration include Petri nets (Navarre et al. 2005) as well as guided propagation networks (Martin 1998). In these systems speech input was combined with either the mouse, keyboard, pen input, touch input, or 3D gestures.

1.4.4 Symbolic/Statistical Integration

Symbolic/statistical integration takes more traditional unification-based approaches and combines them with statistical processing to form hybrid multimodal integration strategies (Wu et al. 1999). These strategies also bring in concepts from machine learning. Although machine learning has been primarily used with feature level integration, it also have been explored at the decision level (Dumas et al. 2009). As an example, Pan et al. (1999) used Bayesian inference to derive a formula for estimating joint probabilities of multisensory signals and uses appropriate mapping functions to reflect signal dependencies. Mapping selection is guided by maximum mutual information.

Another early example of a symbolic/statistical integration technique is the Member Team Committee (MTC) architecture (Wu et al. 2002) used in the QuickSet application. In this approach, modalities are integrated based on their posterior probabilities. Recognizers from individual modes become members of an MTC statistical integrator and multiple teams are then trained to coordinate and weight the output from the different modes. Each team established a posterior estimate for a multimodal command, given the current input received. The committee of the MTC integrator analyzes the empirical distribution of the posteriors and then establishes an n-best ranking for each possible command. Flippo et al. (2003) used a similar approach to MTC as part of a multimodal interaction framework where they use parallel agents to estimate a posterior probability for each modal recognition result and then weights them to come to an overall decision on the appropriate multimodal command. More recently, Dumas et al. (2012) developed a statistical multimodal integration scheme that uses hidden Markov models to select relevant modalities at the semantic level via temporal relationship properties. More information on using machine learning in multimodal integration can be found in (Damousis and Argyropoulos 2012; Huang et al. 2006). Although these methods can be very powerful at modeling the uncertainty in multimodal integration, their main drawback is the need for an adequate amount of training data.

1.5 Usability Issues with Multimodal Interaction

Given the very nature of multimodal interaction in the context of perceptual computing, usability becomes a critical part of multimodal interface design because of the attempt to tightly couple different input modalities together in a way that provides an intuitive and powerful user experience (Reeves et al. 2004). To discuss some of the usability issues with multimodal input, we use Oviatt's 10 myths of multimodal interaction as starting points for discussion (Oviatt 1999). Although written several years ago, they are still applicable today.

If you build a multimodal system, users will interact multimodally. Just because an application supports multiple modes of input, does not mean that users will make use of them for all commands. Thus, flexibility in the command structure is critical to support natural forms of communication between the human and the computer. In other words, multimodal interfaces should be designed with flexibility for the user in mind so they can issue commands in different ways. For example, a user should be able to use both speech and 3D gesture simultaneously to issue a command as well as providing support for using speech and eye gaze or using 3D gesture and eye gaze or using speech in isolation. This design choice makes the overall mutimodal user interface more complex in terms of how the input modes are integrated, but provides the most generality.

Speech and pointing is the dominant multimodal integration pattern. From a usability perspective, speech and pointing make for an intuitive multimodal input combination, especially when the user wants to select some virtual object and then perform some operation on it (e.g., paint [this] cylinder blue). However, as we have seen in this chapter, there are a variety of different multimodal input combinations that are possible. They key question from a usability perspective is does a particular input combination make sense for a particular task. As a general guideline, it is important to provide multimodal input combinations that will support a simple and natural interaction metaphor for any given task. For example, speech and pointing may not be the best input combination in mobile settings where touch input or 3D gestures may be more appropriate.

Multimodal input involves simultaneous signals. Not all multimodal input strategies require users to perform the individual inputs at the same time. Certain modalities certainly afford such temporal integration, but in many cases individual input modes are used consecutively (e.g., saying something then performing a 3D gesture and vise versa) but they still act as complementary modes of input. In fact, multimodal input strategies can also use one mode for certain tasks and a second or third mode for other tasks. Thus, from a usability perspective, it is import to recognize that there are a number of different ways individual inputs can be combined and not all of them need to support simultaneous input.

Speech is the primary input mode in any multimodal system that includes it. While speech is a dominant input channel that humans use to communicate, it many cases it does not need to be the primary modality in a multimodal interface. Unfortunately, speech recognition can be compromised in noisy environments, making it a less robust input mechanism. In addition, it may be the case that users do not want to use voice input for privacy concerns. In other cases, speech may simply be a backup modality or other modalities may be better suited to combine for a given task. Thus, when designing multimodal interaction scenarios it is not a requirement to make speech the primary input mode and should only be used when it makes the most sense.

Multimodal language does not differ linguistically from unimodal language. One of the benefits of multimodal interaction is that it tends to simplify input patterns. As an example, consider a user who wants to move an object from one location to another. Using speech in isolation would require the user to not only describe the object of interest, but also describe the location that the objects needs to be placed. However, with a combination of speech and gesture, the user can simplify the description of the object because they are also using a second modality (in this case pointing) to both identify the object and place it in a different location. This input combination implies that one can use simpler input commands for each individual mode of input when performing concurrent multimodal interaction. In terms of usability, it is important to understand that unimodal language can be more complex than multimodal language and that multimodal input can take away this complexity making for an easier to use interface.

Multimodal integration involves redundancy of content between modes. One of the key ideas behind multimodal integration is that providing redundant modes of input can help to support a better user experience because each individual mode can be used to reinforce one another. This is certainly true from a computational point of view and does have its place in multimodal integration. However, the complementary nature of multimodal input should not be ignored for its benefits from a usability perspective. Thus, ensuring proper multimodal integration to support complementarity is important from the user's perspective.

Individual error-prone recognition technologies combine multimodally to produce even greater unreliability. One of the interesting challenges with multimodal input, especially with perceptual computing, is that the input modalities used (e.g., speech, 3D gesture, 2D gesture) require recognition technologies used to understand the input. Unfortunately recognition is error-prone based on the accuracy of the individual recognizers. However, combining multiple recognition-based input together actually can help improve accuracy for the overall command, producing a more reliable interface. Multimodal integration strategies are key to this improvement. In addition, users, if given a choice, will tend to work with modalities that they believe have higher accuracies. Thus, from a usability perspective, this usage pattern is another reason to ensure the multimodal interface is flexible.

All users' multimodal commands are integrated in a uniform way. Users of multimodal interfaces tend to identify an integration pattern for how they will use the interface fairly early on and they will stick with this pattern. However, as we have seen, there are many different ways that people can use a multimodal interface. Thus, it is important for the multimodal integration scheme to be flexible and try to identify the dominant integration pattern on a per user basis. This approach then can support improved recognition rates because the fusion engine would be aware of how the user is interacting with the different modalities.

Different input modes are capable of transmitting comparable content. Not all input modalities are created equal. In other words, different modalities have strengths and weaknesses depending on the type of information that the user wants to convey. For example, eye gaze will produce very different types of information than speech. Thus, from a usability perspective it is important to understand what modalities are available and what each one is ideal for. Trying to use a given modality as input for tasks that it is not suited for will only make the interface difficult to use.

Enhanced efficiency is the main advantage of multimodal systems. Finally, speed and efficiency is not the only advantage of multimodal interfaces. Reducing errors from individual recognition systems as well as providing more flexibility to interact with an application in the way that users want to are also important advantages of multimodal interaction. In addition, a properly designed multimodal interface will provide a level of generality to the user to support a variety of different tasks, applications, and environments.

1.6 Conclusion

In this chapter, we have explored how combining different input modalites can form natural and expressive multimodal interfaces. We have examined the types of multimodal input strategies and presented a variety of different multimodal interfaces that offer different combinations of touch input, speech, 3D gesture, eye gaze and tracking, facial expressions, and brain computer input. We have also examined multimodal integration or fusion, a critical component of a multimodal architecture that integrates different modalities together to form a cohesive interface by examining the different approaches and levels of integration. Finally, we have presented a number of usability issues as they relate to multimodal input. Clearly, mulitmodal interfaces have come a long way from Bolt's Put that there system (Bolt 1980). However, more work is still needed in a variety of areas including multimodal integration, recognition technology, and usability to fully support perceptual computing applications that provide powerful, efficient, and expressive human computer interaction.

References

Barthelmess P, Kaiser E, Huang X and Demirdjian D 2005 Distributed pointing for multimodal collaboration over sketched diagrams *Proceedings of the 7th international conference on Multimodal interfaces*, pp. 10–17 ICMI '05. ACM, New York, NY, USA.

Beelders TR and Blignaut PJ 2012 Measuring the performance of gaze and speech for text input *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 337–340 ETRA '12. ACM, New York, NY, USA.

Bohus D and Horvitz E 2009 Dialog in the open world: platform and applications *Proceedings of the 2009 international conference on Multimodal interfaces*, pp. 31–38 ACM.

Bolt RA 1980 "put-that-there": Voice and gesture at the graphics interface *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, pp. 262–270 SIGGRAPH '80. ACM, New York, NY, USA.

Bouchet J, Nigay L and Ganille T 2004 Icare software components for rapidly developing multimodal interfaces *Proceedings of the 6th international conference on Multimodal interfaces*, pp. 251–258 ICMI '04. ACM, New York, NY, USA.

Bourguet M 2002 A toolkit for creating and testing multimodal interface designs *Proceedings of User Interface Software and Technology (UIST 2002) Companion proceedings*, pp. 29–30.

Bowman DA, Kruijff E, LaViola JJ and Poupyrev I 2004 *3D User Interfaces: Theory and Practice*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA.

Budhiraja P and Madhvanath S 2012 The blue one to the left: enabling expressive user interaction in a multimodal interface for object selection in virtual 3d environments *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 57–58 ACM.

Burger B, Ferrané I, Lerasle F and Infantes G 2012 Two-handed gesture recognition and fusion with speech to command a robot. *Autonomous Robots* **32**(2), 129–147.

Busso C, Deng Z, Yildirim S, Bulut M, Lee CM, Kazemzadeh A, Lee S, Neumann U and Narayanan S 2004 Analysis of emotion recognition using facial expressions, speech and multimodal information *Proceedings of the 6th international conference on Multimodal interfaces*, pp. 205–211 ICMI '04. ACM, New York, NY, USA.

Calder J 1987 Typed unification for natural language processing In *Categories, Polymorphism, and Unification* (ed. Kahn G, MacQueen D and Plotkin G) Centre for Cognitive Science University of Edinburgh, Edinburgh, Scotland†.

Caridakis G, Karpouzis K, Wallace M, Kessous L and Amir N 2010 Multimodal users affective state analysis in naturalistic interaction. *Journal on Multimodal User Interfaces* **3**(1-2), 49–66.

Cha T and Maier S 2012 Eye gaze assisted human-computer interaction in a hand gesture controlled multi-display environment *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, pp. 13:1–13:3 Gaze-In '12. ACM, New York, NY, USA.

Cohen PR, Johnston M, McGee D, Oviatt S, Pittman J, Smith I, Chen L and Clow J 1997a Quickset: Multimodal interaction for distributed applications *Proceedings of the fifth ACM international conference on Multimedia*, pp. 31–40 ACM.

Cohen PR, Johnston M, McGee D, Oviatt S, Pittman J, Smith I, Chen L and Clow J 1997b Quickset: multimodal interaction for distributed applications *Proceedings of the fifth ACM international conference on Multimedia*, pp. 31–40 MULTIMEDIA '97. ACM, New York, NY, USA.

Damousis IG and Argyropoulos S 2012 Four machine learning algorithms for biometrics fusion: a comparative study. *Appl. Comp. Intell. Soft Comput.* **2012**, 6:6–6:6.

De Silva L, Miyasato T and Nakatsu R 1997 Facial emotion recognition using multi-modal information *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*, vol. 1, pp. 397–401 vol.1.

Dearman D, Karlson A, Meyers B and Bederson B 2010 Multi-modal text entry and selection on a mobile device *Proceedings of Graphics Interface 2010*, pp. 19–26 Canadian Information Processing Society.

Duchowski AT 2007 *Eye Tracking Methodology: Theory and Practice*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Dumas B, Lalanne D and Oviatt S 2009 Multimodal interfaces: A survey of principles, models and frameworks In *Human Machine Interaction* (ed. Lalanne D and Kohlas J) vol. 5440 of *Lecture Notes in Computer Science* Springer Berlin Heidelberg pp. 3–26.

Dumas B, Lalanne D, Guinard D, Koenig R and Ingold R 2008 Strengths and weaknesses of software architectures for the rapid creation of tangible and multimodal interfaces *Proceedings of the 2nd international conference on Tangible and embedded interaction*, pp. 47–54 TEI '08. ACM, New York, NY, USA.

Dumas B, Signer B and Lalanne D 2012 Fusion in multimodal interactive systems: an hmm-based algorithm for user-induced adaptation *Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems*, pp. 15–24 EICS '12. ACM, New York, NY, USA.

Ehlen P and Johnston M 2011 Multimodal local search in speak4it *Proceedings of the 16th international conference on Intelligent user interfaces*, pp. 435–436 ACM.

Fazli S, Mehnert J, Steinbrink J, Curio G, Villringer A, Müller KR and Blankertz B 2012 Enhanced performance by a hybrid nirs–eeg brain computer interface. *Neuroimage* **59**(1), 519–529.

Feng J, Johnston M and Bangalore S 2011 Speech and multimodal interaction in mobile search. *Signal Processing Magazine, IEEE* **28**(4), 40–49.

Flippo F, Krebs A and Marsic I 2003 A framework for rapid development of multimodal interfaces *Proceedings of the 5th international conference on Multimodal interfaces*, pp. 109–116 ICMI '03. ACM, New York, NY, USA.

Gallo L, Placitelli AP and Ciampi M 2011 Controller-free exploration of medical image data: Experiencing the kinect *Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on*, pp. 1–6 IEEE.

Gruenstein A, Orszulak J, Liu S, Roberts S, Zabel J, Reimer B, Mehler B, Seneff S, Glass J and Coughlin J 2009 City browser: developing a conversational automotive hmi *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, pp. 4291–4296 CHI EA '09. ACM, New York, NY, USA.

Gürkök H and Nijholt A 2012 Brain–computer interfaces for multimodal interaction: a survey and principles. *International Journal of Human-Computer Interaction* **28**(5), 292–307.

Gürkök H, Hakvoort G and Poel M 2011 Modality switching and performance in a thought and speech controlled computer game *Proceedings of the 13th international conference on multimodal interfaces*, pp. 41–48 ICMI '11. ACM, New York, NY, USA.

Heo H, Lee EC, Park KR, Kim CJ and Whang M 2010 A realistic game system using multi-modal user interfaces. *Consumer Electronics, IEEE Transactions on* **56**(3), 1364–1372.

Hild J, Muller E, Klaus E, Peinsipp-Byma E and Beyerer J 2013 Evaluating multi-modal eye gaze interaction for moving object selection *ACHI 2013 : The Sixth International Conference on Advances in Computer-Human Interactions*, pp. 454–459.

Holzapfel H, Nickel K and Stiefelhagen R 2004 Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures *Proceedings of the 6th international conference on Multimodal interfaces*, pp. 175–182 ICMI '04. ACM, New York, NY, USA.

Hoste L, Dumas B and Signer B 2012 Speeg: a multimodal speech-and gesture-based text input solution *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 156–163 ACM.

Hrúz M, Campr P, Dikici E, Kındıroğlu AA, Krňoul Z, Ronzhin A, Sak H, Schorno D, Yalçın H, Akarun L *et al.* 2011 Automatic fingersign-to-speech translation system. *Journal on Multimodal User Interfaces* **4**(2), 61–79.

Huang X, Oviatt S and Lunsford R 2006 Combining user modeling and machine learning to predict users multimodal integration patterns In *Machine Learning for Multimodal Interaction* (ed. Renals S, Bengio S and Fiscus J) vol. 4299 of *Lecture Notes in Computer Science* Springer Berlin Heidelberg pp. 50–62.

Jaimes A and Sebe N 2007 Multimodal humancomputer interaction: A survey. *Computer Vision and Image Understanding* **108**(12), 116 – 134. Special Issue on Vision for Human-Computer Interaction.

Johnston M and Bangalore S 2005 Finite-state multimodal integration and understanding. *Nat. Lang. Eng.* **11**(2), 159–187.

Johnston M, Cohen PR, McGee D, Oviatt SL, Pittman JA and Smith I 1997 Unification-based multimodal integration *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 281–288 ACL '98. Association for Computational Linguistics, Stroudsburg, PA, USA.

Kessous L, Castellano G and Caridakis G 2010 Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces* **3**(1-2), 33–48.

Koons DB, Sparrell CJ and Thorisson KR 1993 Intelligent multimedia interfaces American Association for Artificial Intelligence Menlo Park, CA, USA chapter Integrating simultaneous input from speech, gaze, and hand gestures, pp. 257–276.

Krahnstoever N, Kettebekov S, Yeasin M and Sharma R 2002 A real-time framework for natural multimodal interaction with large screen displays *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, p. 349 IEEE Computer Society.

Kristensson PO and Vertanen K 2011 Asynchronous multimodal text entry using speech and gesture keyboards. *INTERSPEECH*, pp. 581–584.

Kumar M, Paepcke A and Winograd T 2007 Eyepoint: practical pointing and selection using gaze and keyboard *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 421–430 CHI '07. ACM, New York, NY, USA.

Lalanne D, Nigay L, Palanque p, Robinson P, Vanderdonckt J and Ladry JF 2009 Fusion engines for multimodal input: a survey *Proceedings of the 2009 international conference on Multimodal interfaces*, pp. 153–160 ICMI-MLMI '09. ACM, New York, NY, USA.

Laput GP, Dontcheva M, Wilensky G, Chang W, Agarwala A, Linder J and Adar E 2013 Pixeltone: a multimodal interface for image editing *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2185–2194 CHI '13. ACM, New York, NY, USA.

Latoschik M 2002 Designing transition networks for multimodal vr-interactions using a markup language *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pp. 411–416.

Lee SH, Sohn MK, Kim DJ, Kim B and Kim H 2013 Smart tv interaction system using face and hand gesture recognition *Consumer Electronics (ICCE), 2013 IEEE International Conference on*, pp. 173–174 IEEE.

Leeb R, Sagha H, Chavarriaga R and del R Millan J 2010 Multimodal fusion of muscle and brain signals for a hybrid-bci *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pp. 4343–4346 IEEE.

Lisetti CL and Nasoz F 2002 Maui: a multimodal affective user interface *Proceedings of the tenth ACM international conference on Multimedia*, pp. 161–170 MULTIMEDIA '02. ACM, New York, NY, USA.

Martin JC 1998 Tycoon: Theoretical framework and software tools for multimodal interfaces *In John Lee (Ed.), Intelligence and Multimodality in Multimedia Interfaces*. AAAI Press.

Martínez R, Collins A, Kay J and Yacef K 2011 Who did what? who said that?: Collaid: an environment for capturing traces of collaborative learning at the tabletop *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, pp. 172–181 ACM.

Maye A, Zhang D, Wang Y, Gao S and Engel AK 2011 Multimodal brain-computer interfaces. *Tsinghua Science & Technology* **16**(2), 133–139.

Muller C and Weinberg G 2011 Multimodal input in the car, today and tomorrow. *Multimedia, IEEE* **18**(1), 98–103.

Navarre D, Palanque P, Bastide R, Schyn A, Winckler M, Nedel LP and Freitas CMDS 2005 A formal description of multimodal interaction techniques for immersive virtual reality applications *Proceedings of the 2005 IFIP TC13 international conference on Human-Computer Interaction*, pp. 170–183 INTERACT'05. Springer-Verlag, Berlin, Heidelberg.

Neal JG, Thielman CY, Dobes Z, Haller SM and Shapiro SC 1989 Natural language with integrated deictic and graphic gestures *Proceedings of the workshop on Speech and Natural Language*, pp. 410–423 HLT '89. Association for Computational Linguistics, Stroudsburg, PA, USA.

Nigay L and Coutaz J 1995 A generic platform for addressing the multimodal challenge *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 98–105 CHI '95. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA.

Oviatt S 1999 Ten myths of multimodal interaction. *Commun. ACM* **42**(11), 74–81.

Oviatt S 2003 Advances in robust multimodal interface design. *Computer Graphics and Applications, IEEE* **23**(5), 62–68.

Oviatt S 2007 Multimodal interfaces In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, Second Edition* (ed. Sears A and Jacko J) CRC Press pp. 413–432.

Oviatt S and Vangent R 1996 Error resolution during multimodal human-computer interaction. pp. 204–207.

Paek T, Thiesson B, Ju YC and Lee B 2008 Search vox: Leveraging multimodal refinement and partial knowledge for mobile voice search *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pp. 141–150 ACM.

Pan H, Liang ZP and Huang T 1999 Exploiting the dependencies in information fusion *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2, pp. –412 Vol. 2.

Perzanowski D, Schultz AC, Adams W, Marsh E and Bugajska M 2001 Building a multimodal human-robot interface. *Intelligent Systems, IEEE* **16**(1), 16–21.

Pfleger N 2004 Context based multimodal fusion *Proceedings of the 6th international conference on Multimodal interfaces*, pp. 265–272 ICMI '04. ACM, New York, NY, USA.

Pfleging B, Kienast M, Schmidt A and Döring T 2011 Speet: A multimodal interaction style combining speech and touch interaction in automotive environments *Adjunct Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI*, pp. 65–66.

Pfleging B, Schneegass S and Schmidt A 2012 Multimodal interaction in the car: combining speech and gestures on the steering wheel *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 155–162 AutomotiveUI '12. ACM, New York, NY, USA.

Piper AM 2010 Supporting medical communication with a multimodal surface computer *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, pp. 2899–2902 CHI EA '10. ACM, New York, NY, USA.

Piper AM and Hollan JD 2008 Supporting medical conversations between deaf and hearing individuals with tabletop displays *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pp. 147–156 CSCW '08. ACM, New York, NY, USA.

Radhakrishnan S, Lin Y, Zeid I and Kamarthi S 2012 Finger-based multitouch interface for performing 3d cad operations. *International Journal of Human-Computer Studies*.

Ramsay A, McGee-Lennon M, Wilson GA, Gray SJ, Gray P and De Turenne F 2010 Tilt and go: exploring multimodal mobile maps in the field. *Journal on Multimodal User Interfaces* **3**(3), 167–177.

Reeves LM, Lai J, Larson JA, Oviatt S, Balaji TS, Buisine S, Collings P, Cohen P, Kraal B, Martin JC, McTear M, Raman T, Stanney KM, Su H and Wang QY 2004 Guidelines for multimodal user interface design. *Commun. ACM* **47**(1), 57–59.

Sandbach G, Zafeiriou S, Pantic M and Yin L 2012 Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing* **30**(10), 683 – 697. ¡ce:title¿3D Facial Behaviour Analysis and Understanding¡/ce:title¿.

Sharma A, Madhvanath S, Shekhawat A and Billinghurst M 2011 Mozart: a multimodal interface for conceptual 3d modeling *Proceedings of the 13th international conference on multimodal interfaces*, pp. 307–310 ICMI '11. ACM, New York, NY, USA.

Sharma R, Pavlovic V and Huang T 1998 Toward multimodal human-computer interface. *Proceedings of the IEEE* **86**(5), 853–869.

Shinoda K, Watanabe Y, Iwata K, Liang Y, Nakagawa R and Furui S 2011 Semi-synchronous speech and pen input for mobile user interfaces. *Speech Communication* **53**(3), 283–291.

Sim KC 2012 Speak-as-you-swipe (says): a multimodal interface combining speech and gesture keyboard synchronously for continuous mobile text entry *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 555–560 ACM.

Sree S, Verma A and Rai R 2013 Creating by imaging: Use of natural and intuitive bci in 3d cad modeling *ASME International Design Engineering Technical Conference ASME/DETC/CIE* ASME.

Stellmach S and Dachselt R 2012 Investigating gaze-supported multimodal pan and zoom *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 357–360 ETRA '12. ACM, New York, NY, USA.

Stiefelhagen R, Fugen C, Gieselmann R, Holzapfel H, Nickel K and Waibel A 2004 Natural human-robot interaction using speech, head pose and gestures *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3, pp. 2422–2427 IEEE.

Sun Y, Chen F, Shi YD and Chung V 2006 A novel method for multi-sensory data fusion in multimodal human computer interaction *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, pp. 401–404 OZCHI '06. ACM, New York, NY, USA.

Takahashi M, Fujii M, Naemura M and Satoh S 2013 Human gesture recognition system for tv viewing using time-of-flight camera. *Multimedia Tools and Applications* **62**(3), 761–783.

Taylor G, Frederiksen R, Crossman J, Quist M and Theisen P 2012 A multi-modal intelligent user interface for supervisory control of unmanned platforms *Collaboration Technologies and Systems (CTS), 2012 International Conference on*, pp. 117–124.

Tse E, Greenberg S and Shen C 2006 Gsi demo: multiuser gesture/speech interaction over digital tables by wrapping single user applications *Proceedings of the 8th international conference on Multimodal interfaces*, pp. 76–83 ACM.

Tse E, Greenberg S, Shen C and Forlines C 2007 Multimodal multiplayer tabletop gaming. *Computers in Entertainment (CIE)* **5**(2), 12.

Tse E, Greenberg S, Shen C, Forlines C and Kodama R 2008 Exploring true multi-user multimodal interaction over a digital table *Proceedings of the 7th ACM conference on Designing interactive systems*, pp. 109–118 DIS '08. ACM, New York, NY, USA.

Turk M 2013 Multimodal interaction: A review. *Pattern Recognition Letters* (0), –.

Turk M and Robertson G 2000 Perceptual user interfaces (introduction). *Commun. ACM* **43**(3), 32–34.

Vilimek R and Zander T 2009 Bc(eye): Combining eye-gaze input with brain-computer interaction In *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments* (ed. Stephanidis C) vol. 5615 of *Lecture Notes in Computer Science* Springer Berlin Heidelberg pp. 593–602.

Vo MT and Wood C 1996 Building an application framework for speech and pen input integration in multimodal learning interfaces *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 6, pp. 3545–3548.

Ward DJ, Blackwell AF and MacKay DJ 2000 Dashera data entry interface using continuous gestures and language models *Proceedings of the 13th annual ACM symposium on User interface software and technology*, pp. 129–137 ACM.

Williamson BM, LaViola JJ, Roberts T and Garrity P 2012 Multi-kinect tracking for dismounted soldier training *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, vol. 2012 NTSA.

Williamson BM, Wingrave C, LaViola JJ, Roberts T and Garrity P 2011 Natural full body interaction for navigation in dismounted soldier training *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, vol. 2011 NTSA.

Wöllmer M, Metallinou A, Eyben F, Schuller B and Narayanan SS 2010 Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. *INTERSPEECH*, pp. 2362–2365.

Wu L, Oviatt SL and Cohen PR 1999 Multimodal integration - a statistical view. *IEEE Transactions on Multimedia* **1**, 334–341.

Wu L, Oviatt SL and Cohen PR 2002 From members to teams to committee-a robust approach to gestural and multimodal recognition. *IEEE Transactions on Neural Networks* **13**(4), 972–982.

Zander TO, Gaertner M, Kothe C and Vilimek R 2010a Combining eye gaze input with a brain–computer interface for touchless human–computer interaction. *Intl. Journal of Human–Computer Interaction* **27**(1), 38–51.

Zander TO, Kothe C, Jatzev S and Gaertner M 2010b Enhancing human-computer interaction with input from active and passive brain-computer interfaces *Brain-Computer Interfaces* Springer pp. 181–199.

Zhai S, Morimoto C and Ihde S 1999 Manual and gaze input cascaded (magic) pointing *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 246–253 CHI '99. ACM, New York, NY, USA.

Zhang X and MacKenzie I 2007 Evaluating eye tracking with iso 9241 - part 9 In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments* (ed. Jacko J) vol. 4552 of *Lecture Notes in Computer Science* Springer Berlin Heidelberg pp. 779–788.