I. Wagner, H. Tellioglu, E. Balka, C. Simone, and L. Ciolfi (eds.).
ECSCW'09: Proceedings of the 11th European Conference on Computer Supported Cooperative Work,
7-11 September 2009, Vienna, Austria

Analyzing Multimodal Communication around a Shared Tabletop Display

Anne Marie Piper and James D. Hollan
Department of Cognitive Science, University of California, San Diego, USA
ampiper@hci.ucsd.edu, hollan@hci.ucsd.edu

Abstract. Communication between people is inherently multimodal. People employ speech, facial expressions, eye gaze, and gesture, among other facilities, to support communication and cooperative activity. Complexity of communication increases when a person is without a modality such as hearing, often resulting in dependence on another person or an assistive device to facilitate communication. This paper examines communication about medical topics through *Shared Speech Interface*, a multimodal tabletop display designed to assist communication between a hearing and deaf individual by converting speech-to-text and representing dialogue history on a shared interactive display surface. We compare communication mediated by a multimodal tabletop display and by a human sign language interpreter. Results indicate that the multimodal tabletop display (1) allows the deaf patient to watch the doctor when she is speaking, (2) encourages the doctor to exploit multimodal communication such as co-occurring gesture-speech, and (3) provides shared access to persistent, collaboratively produced representations of conversation. We also describe extensions of this communication technology, discuss how multimodal analysis techniques are useful in understanding the affects of multiuser multimodal tabletop systems, and briefly allude to the potential of applying computer vision techniques to assist analysis.

Introduction

Loss of hearing is a common problem that can result from a variety of factors (e.g., noise, aging, disease, and heredity). Approximately 28 million Americans have significant hearing loss, and of that group, almost six million are profoundly deaf (NIDCD, 2008). A primary form of communication within the United States deaf

community is American Sign Language (ASL). ASL interpreters play a central role in enabling face-to-face communication between deaf and hearing individuals. For the deaf population fluent in ASL, communicating through an interpreter is an optimal choice for many situations. Interpreters, however, are expensive and in many situations not available. Furthermore, though interpreters are bound by a confidentiality agreement, the presence of a third person in a private conversation may reduce a deaf person's comfort and inhibit their willingness to speak candidly. These factors are especially relevant for the topic of our current analysis: medical conversations between a deaf patient and a hearing, non-signing doctor.

We designed and evaluated Shared Speech Interface (SSI), a multimodal tabletop application that facilitates communication between a deaf and hearing individual. The application was designed to provide private and independent communication within the context of doctor-patient consultations. While our initial findings indicate that communicating through a multimodal tabletop display is both feasible and desirable for deaf individuals (Piper and Hollan, 2008), it is not yet clear how the tabletop display affects communication on a cognitive and social level. This paper presents a micro-analysis of interaction between deaf and hearing individuals to begin to address questions regarding communication, coordination, and cognition. Our analysis examines speech, gesture, eye gaze, and device interaction involving the doctor, patient, and sign language interpreter. We find that the digital table provides dialogue with properties that are not available in conversation through a human interpreter. Specifically, the digital table transforms ephemeral dialogue into a lasting form that allows the deaf individual to better attend to the speaker, supports co-occurring gesture-speech by the hearing user, and provides a shared visual record of conversation.

Deaf Communication

Deaf individuals living in a hearing world face communication challenges everyday and often rely on other people or devices to assist communication. While not all deaf or hearing impaired individuals use sign language, sources estimate that ASL is the fourth most widely used language in the United States (NIDCD, 2008). Sign language interpreters are a common solution for facilitating communication between deaf and hearing individuals, but access to an interpreter requires foresight and can be expensive. While interpreter services are important, they raise issues of privacy in communication. The Deaf community in many locations is small and well-connected. It is not uncommon for a deaf person to know the interpreter, which creates concern for very personal conversations. The interpreter scheduled on a given day may also be of the opposite gender, making discussion of certain medical issues even more uncomfortable. Face-to-face communication through an interpreter requires the deaf individual to focus their attention on the interpreter rather than the speaker. Taking notes during conversation involving an interpreter is also challenging because the deaf individual must pay close attention to the interpreter and cannot easily look down to make notes on paper. Not all deaf individuals

know how to read and write in a spoken language such as English, but those who are proficient may use hand written notes to communicate in the absence of an interpreter. Communication with the hearing world is further complicated because sign languages are not simply visual forms of spoken languages. Instead, each sign language has its own unique grammatical and syntactical structure, making a spoken language a second language for many deaf individuals.

Technology has transformed communication for the Deaf community. Telephone use was impossible for deaf individuals until the adaptation of the Teletype machine (TTY) which allowed individual lines of keyboard entry to be transmitted over phone lines. Adoption of the TTY, its subsequent electronic versions, and now the personal computer, made typing an essential mode of communication within the Deaf community. Researchers have developed a variety of technologies to address communication barriers between the deaf community and hearing world. As early as 1975, researchers began investigating how cooperative computing environments, such as early forms of instant messenger, could facilitate communication between deaf and hearing individuals (Turoff, 1975). More recently, human-computer interaction researchers have examined how mobile devices (e.g., Cavender et al., 2006), tablet computers (Miller et al., 2007), and browser based technologies (Schull, 2006) can augment communication for deaf individuals. While these solutions address various communication challenges for deaf individuals, none address face-to-face communication around a single shared display.

Multimodal Tabletop Displays

Digitally enhanced tabletop displays are growing in appeal and availability. The ability to receive multiple simultaneous touch inputs from a number of people makes tabletop displays a promising technology for facilitating face-to-face group interaction. Within the field of human-computer interaction, substantial attention is given to how tabletop displays can support face-to-face communication and mediate group social dynamics (see Morris, 2006, for a review). Compared to vertical displays such as a computer monitor or wall mounted display, tabletop displays result in more equitable interaction and shared responsibility by group members (Rogers and Lindley, 2004). Recently, there has been growing interest in multimodal multitouch tabletop systems. A multimodal tabletop system accepts touch along with speech and/or eye gaze as input to the system. Tse and his collegues explored how multimodal tabletop systems support gaming, pair interaction around a multimodal tabletop display, and techniques to wrap single-user applications so they include multimodal interaction (2007). Researchers have examined a variety of tabletop group work issues with hearing populations, but until recently with the Shared Speech Interface project (Piper and Hollan, 2008), researchers had yet to examine tabletop computing scenarios with hearing impaired populations.

We developed Shared Speech Interface (SSI), a multimodal tabletop application that enables co-located face-to-face communication and cooperative activity between a hearing and deaf individual. The design of SSI exploits the affordances

of multimodal tabletop displays while addressing communication needs between a deaf patient and a hearing, non-signing medical doctor. Consultations with physicians often involve visuals such as medical records, charts, and scan images. Interactive tabletop displays are effective for presenting visual information to multiple people at once without necessarily designating one person as the owner of the visual. Taking notes while meeting with a physician is problematic for deaf individuals because it requires simultaneously attending to the doctor's facial expressions, the interpreter's visual representation of speech, and notes on paper. A multimodal tabletop display allows the doctor and patient to maintain face-to-face contact while viewing a shared, interactive representation of their conversation and other visual materials.

SSI runs on a MERL DiamondTouch table (Dietz and Leigh, 2001) and uses the DiamondSpin toolkit (Shen et al., 2004). The DiamondTouch table is a multiuser, multitouch top-projected tabletop display. People sit on conductive pads that enable the system to uniquely identify each user and where each user is touching the surface. SSI supports conversational input through standard keyboard entry and a headset microphone. The system is currently English based. Audio captured from the microphone is fed into a speech recognition engine, converted from speech-to-text, and then displayed on the tabletop interface. Currently, SSI works for two users communicating in a face-to-face setting. The hearing user speaks into the headset microphone and the deaf individual enters speech through a standard peripheral keyboard. As the two individuals communicate, their speech appears on the tabletop display in the form of moveable speech bubbles. See Piper and Hollan (2008) for a detailed description of the system design.

Figure 1. A medical doctor and a deaf patient communicate using Shared Speech Interface.

Analysis of Multimodal Human Interaction

While a tabletop display is considered multimodal when it has multiple modalities of input (i.e., touch and speech, or touch and eye tracking), interaction with other people around a tabletop display is inherently multimodal. In this paper we use

video analysis techniques to closely examine the interplay between speech, gesture, and eye gaze as well as interaction with the device. Video analysis is routinely used to understand activity within naturalistic settings (e.g., Heath, 1986), but some laboratory studies also include analysis of multimodal human interaction data (e.g., Bekker et al., 1995; Kraut et al., 2003; Kirk et al., 2005). From a methodological perspective, Kirk et al. (2005) note the importance of studying laboratory data in an "ethnographic fashion." Furthermore, Hollan et al. (2000) argue more directly for an integrated approach to human-computer interaction research based on theories of distributed cognition and a combination of ethnographic and experimental techniques.

Gesture in Co-located and Remote Interaction

There is a growing interest in co-located gestural interaction and its relevance to the design of cooperative computing systems. Tang (1991) noted the pervasive nature of hand gestures in a group drawing activity and indicated the need to better understand this activity in relation to the people and artifacts in a co-located workspace. Bekker et al. (1995) studied gestures as a way of informing the design of cooperative systems. Kraut et al. (2003) examined how visual information, especially deictic reference, enabled situational awareness and conversational grounding in face-to-face, video-based, and audio-based interaction.

The horizontal form factor of tables has unique affordances for group work compared to vertically mounted displays. Work by Rogers and Lindley (2004) noted an increased use of gesture when groups interacted around a tabletop display compared to a whiteboard display. In another study, Rogers et al. (2004) found that touching a display with fingers has ancillary benefit for group work such as supporting turn-taking. With respect to gesture, Tse et al. (2007) provided similar observations of pairs interacting around a multimodal tabletop display. They noted that "speech and gesture commands serve double duty as both commands to the computer and as implicit communication to others."

A number of systems examined how representing nonverbal behaviors such as gesture and eye gaze across remote environments affects interaction (e.g., Tang and Minneman, 1990, as an early example). Related to gesture analysis, Kirk et al. (2005) examined how specific hand gestures within the context of remote cooperative activity promote awareness and coordinate object focused actions. Similarly, Luff et al. (2006) examined how people working remotely use pointing gestures to coordinate and align themselves around objects of interest.

Gesture Analysis

The term *gesture* is polysemous for human-computer interaction researchers interested in touch-sensitive surfaces. On one hand, gestures are commands to a computer system administered by touching or moving an object, finger, or hand on an interactive surface. In a more traditional sense, the term gesture refers to the way

in which people move or use their body as a means of communication or expression with oneself or others. This section focuses on this latter meaning of gesture. Recently there has been a growing interest in using gesture analysis to understand communication between people (McNeill, 1992; Kendon and Muller, 2001) and within cooperative work environments (Goodwin and Goodwin, 1996; Hindmarsh and Heath, 2000; Zemel et al., 2008). This is largely driven by a theoretical shift from considering gesture as peripheral to human interaction to viewing gesture as central to communication and thought. Kendon (1980) was one of the first to articulate the perspective that speech and gesture are inextricably linked. McNeill proposed a theory that speech and gesture involve a single conceptual source (McNeill, 1985, 1992). He posits that speech and gesture acts develop together. This and related work (McNeill, 1992; Goldin-Meadow, 2003) provide a foundation for using speech and gesture as a way to understand cognitive activity. Furthermore, gesture can indicate underlying reasoning processes that a speaker may not be able to articulate (Goldin-Meadow, 2003), and thus a better understanding of gesture promises to play a crucial role in teaching and learning (see Roth, 2001, for a review).

For the purposes of our discussion and in agreement with practices of gesture researchers, we examine gesture as spontaneous movements of body or hands that are often produced in time with speech but may also occur in the absence of verbal utterances (see McNeill, 1992). Actions such as head scratching or moving an object in an environment are not considered gestures. In our analysis we pay particular attention to gestures that communicate and mediate activity. We classify gestures into David McNeill's widely accepted categories of beat, deictic, iconic, and metaphoric gesture (1992). Examining the frequency and patterns of various gesture types provides potential insight into how people exploit their bodies and environment to assist communication during multimodal tabletop interaction.

Within gesture research, sign language is considered a separate class of communication. Each sign language has a specific syntactical and grammatical structure, and specific gestural forms within a sign language take on linguistic meaning. Communicating through sign language, however, does not preclude the use of spontaneous gestures as described above. In fact, signers use the same proportion of meaningful gesture as speaking individuals use in verbal dialogue (Liddell and Metzger, 1998). There is growing evidence that people – both hearing and hearing impaired – attend to and interpret information in gestures (Goldin-Meadow, 2003; Cassell et al., 1999; Beattie and Shovelton, 1999).

Eye Gaze Analysis

In addition to gesture, other nonverbal interaction such as eye gaze can provide insight into communication. Early work by Kendon (1967) gives a history of gaze research and describes the function of gaze as "an act of perception by which one interactant can monitor the behavior of another, and as an expressive sign and regulatory signal by which he may influence the behavior of the other." Change in gaze

direction such as looking away while speaking and then back to the listener at the end of an utterance gives listeners information about turn-taking (Duncan, 1972, 1974; Duncan and Fiske, 1977). Eye gaze is also used to demonstrate engagement (Goodwin, 2000, 1981) as well as indicate attention and show liking (Argyle and Cook, 1976; Kleinke, 1986) during face-to-face interaction. Eye gaze, accompanied with or without gesture, is also used in pointing acts (Kita, 2003).

When working with deaf populations, understanding patterns of eye gaze is especially important. Direction of gaze indicates whether or not an individual is attending to visual forms of speech. In conversation, a deaf individual reading sign will maintain relatively steady gaze towards the person signing (Baker and Padden, 1978; Siple, 1978). Eye contact with the signer is a signal that the signer has the floor, and shifting gaze away from the signer can indicate a turn request (Baker, 1977). In American Sign Language, the direction of gaze can also be used for deictic reference (Baker and Padden, 1978; Engberg-Pedersen, 2003), and monitoring gaze direction may provide insight into accompanying interaction. Signers tend to shift gaze from the face of their listener to their own hands when they want to call attention to gestures, and it is common for the signer to look back up at their listener to ensure that they too are looking at the gesture (Gullberg and Holmqvist, 2006). Work by Emmorey et al. (2008) found that people reading sign language do in fact follow gaze down to the hands when a signer looks at his or her hands. In summary, eye gaze is an important aspect of multimodal interaction and understanding it may lead to innovation in cooperative multimodal technology design.

Experimental Setup

Eight deaf adults (mean age=33, stdev=11.4, range=[22,52]; 3 males) and one medical doctor (age=28, female) participated in a laboratory study. All eight deaf participants were born deaf or became deaf before the age of one. Three participants identified English as their native language and five identified ASL. All participants were fluent in ASL and proficient at reading and writing in English. The medical doctor had prior experience treating deaf patients but does not know ASL. None of the participants had used a tabletop display prior to participating in this study.

Deaf participants were given sample medical issues (e.g., about routine vaccinations for travel abroad or advice on losing or gaining weight) to discuss with the doctor. Each deaf participant worked with the same doctor, which resembles the real-world scenario where one doctor has similar conversations with multiple patients throughout the day. The patient and doctor discussed a medical issue using either the multimodal tabletop system (digital table condition) or a professional American Sign Language interpreter (interpreter condition). Each discussion prompt had a corresponding medical visual that was preloaded into the tabletop system (e.g., a map for discussion about foreign travel). A paper version of the visual was provided for the interpreter condition. Medical professionals helped to ensure that the discussion prompts reflected authentic conversations that might occur in normal patient interaction but whose content did not require participants to

discuss information that might be too personal. Deaf participants experienced both the digital table and interpreter condition. The order of conditions and discussion prompts was randomized between subjects. Each session was video taped by two cameras from different angles to capture participants' interactions with each other and the digital table. All sessions were conducted around a DiamondTouch table to keep the environment consistent; the tabletop display was turned off for interpreter condition. Three researchers were present for the testing sessions and took notes. Each conversation with the doctor lasted from seven to nine minutes.

Our research team reviewed over two hours of video data, and together we transcribed and coded key segments of interaction. We were careful to select segments of activity that are representative of behavioral patterns. Video data were transcribed using notation techniques by Goodwin (2000) and McNeill (1992). Brackets surround speech that is co-timed with a gesture, and bold face speech indicates the stroke of the gesture. Transcriptions involving the interpreter indicate the interpreter's speech on behalf of the deaf individual and are not a transcription of sign language used.

Results

Initial findings indicate that Shared Speech Interface is a promising medium for facilitating medical conversations (see Piper and Hollan, 2008, for more details), but how does the multimodal tabletop display shape communication? To answer this question, analysis focuses on four areas of co-located interaction. First, we examine patterns of gaze by the deaf individual as a way to understand their attention during interaction. Second, we present an analysis of gesture by the doctor to identify differences in how she exploits multiple modes of communication depending on the communication medium. Then we discuss how the deaf individual monitors multiple modalities of communication with an emphasis on co-occurring gesture-speech by the doctor. Lastly, we describe how the tabletop display provides persistent, collaboratively produced representations that can aid discussion in cognitively valuable ways.

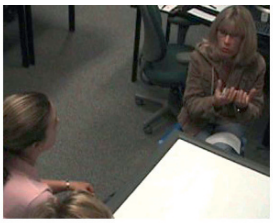
Use of Eye Gaze

Video data reveal distinctly different patterns of eye gaze by the deaf individual when conversation is mediated by an interpreter compared to the multimodal digital table. Eye gaze is a particularly critical channel of communication for deaf individuals, as conversation is purely visual. Examining eye gaze data allows us to infer where the deaf individual is attending during communication. Our results show that when an interpreter is involved in communication, the deaf individual focuses gaze on the interpreter and glances only momentarily at the doctor, as expected per Baker and Padden (1978) and Siple (1978). We found that deaf participants in our study looked at the interpreter when they were reading signs (i.e., "listening") as well as when they were signing (i.e., "speaking"). Consider the following excerpt

of conversation from the interpreter condition. In this interaction, the doctor fixes her gaze on the deaf patient; however, the deaf patient focuses primarily on the interpreter and makes limited eye contact with the doctor. In both conditions, the doctor maintains eye contact with the patient throughout the conversation and uses eye gaze and backchannel communication (e.g., head nodding in center frame of Figure 2) to demonstrate attention and agreement with the patient's speech.

((*patient fixes gaze on interpreter*))
((*doctor fixes gaze on patient*))

D: where do you plan on eating (.)
most of the time (3.2)



P: uh:: (.2) i well (.)
im not exactly sur::e (.2) ((*nods*))
but maybe (.) probably (.)
ill stay with the group that
were going with (.)

D: mmhumm (.5) mhum=((*nods*))

P: and eat in the jungle (.3)
id imagine=

P: ((*shifts gaze to doctor*))
=or im not sure (.3)
((*shifts gaze back to interpreter*))
theres no restaurants there in the jungle (.2)
so i dont know (.6)

Figure 2. Doctor and patient communicating through interpreter. Patient watches interpreter while doctor looks at patient.

To elaborate this point, consider Figure 3 that illustrates the duration and patterns of eye gaze by this same individual. We highlight this case because the pattern illustrated here is typical for interaction. In the interpreter condition the patient fixes her gaze on the interpreter as needed for communication (Figure 3, grey areas in top bar graph). In contrast, communication through the digital table allows her to spend more time watching the doctor (Figure 3, black areas in bottom bar graph). As illustrated by Figure 3, when an interpreter mediates communication, this deaf patient makes quick one-second glances at the doctor and rarely holds gaze for longer than 3 seconds (gaze time on doctor: total=77sec, mean=2.1, stdev=2.0; gaze time on interpreter: total=293sec, mean=8.0, stdev=7.3). This is likely an attempt to demonstrate that she is attending to the doctor without signaling to the interpreter that she would like a turn to speak, as a sustained shift in eye gaze in sign language communication indicates a turn request (Baker, 1977). In the digital table condition, the patient makes frequent shifts in gaze between the doctor and tabletop and looks at the doctor for slightly longer intervals (gaze time on doctor: total=143sec, mean=3.0, stdev=2.6; gaze time on table: total=227sec, mean=4.9, stdev=7.7). The digital table requires the patient to look down for periods of time to type speech on the keyboard. Even with looking down at the keyboard, the doctor in our study noticed a difference in eye gaze by the patient. In a follow-up interview she said:

> The physician patient interaction involves more than just words. Body language is integral to the medical interview and provides key details into the patient's condition

and level of understanding. The inclusion of the interpreter forced the deaf patients to make eye contact with her rather than me, not allowing me to gauge whether information or a question I asked was understood as well as more subtle insights into the patient's overall comfort level.

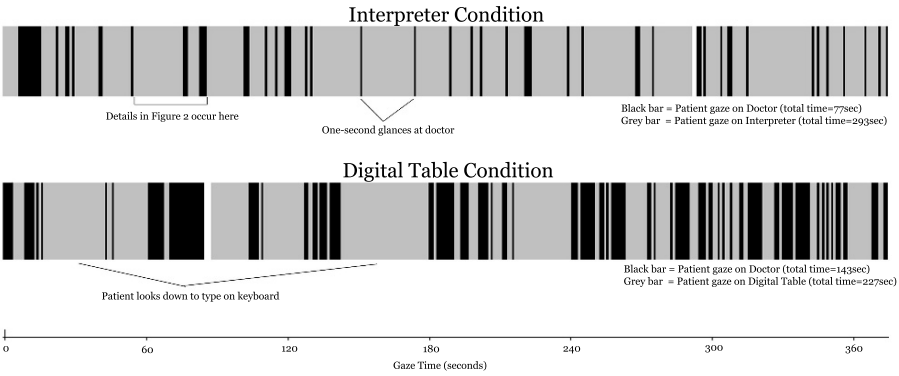


Figure 3. Duration and patterns of eye gaze by the deaf patient during the Interpreter and Digital Table conditions.

Use of Gesture

Communication through the digital table allows the patient to look at the doctor instead of requiring constant focus on the interpreter. Since speech appears in a permanent form on the tabletop display, the urgency of attending to the visual representation of talk is reduced. This allows both the doctor and patient to attend to and exploit multiple modalities of communication. Voice recognition capabilities free the doctor's hands and enable co-occurring gesture-speech in a way that traditional keyboard entry does not afford. Research on synchronized gesture-speech indicates that this activity is often co-expressive and non-redundant, therefore providing interactants with multiple forms of information (McNeill, 1992). Consider another example of interaction in Figures 4. Here, the doctor recommends hand washing techniques to the deaf patient by exploiting multiple modalities of communication including speech, gesture, and eye gaze. First, the patient looks at the doctor as she says "I would recommend." Then the doctor adds her speech to the display and continues "that you wash your hands." Both the doctor and patient look down at the display. Then the patient, likely to demonstrate understanding, holds up his hands and nods his head. The deaf patient's action is an iconic gestural response to the doctor's speech (McNeill, 1992). As he gestures, he shifts his gaze from the tabletop to his hands, likely to call the doctor's attention to his gesture (Gullberg and Holmqvist, 2006; Emmorey et al., 2008).

The patient then looks back at the doctor (Figure 4 middle row, left) as she formulates a recommendation for the patient. She makes a hand rubbing gesture as

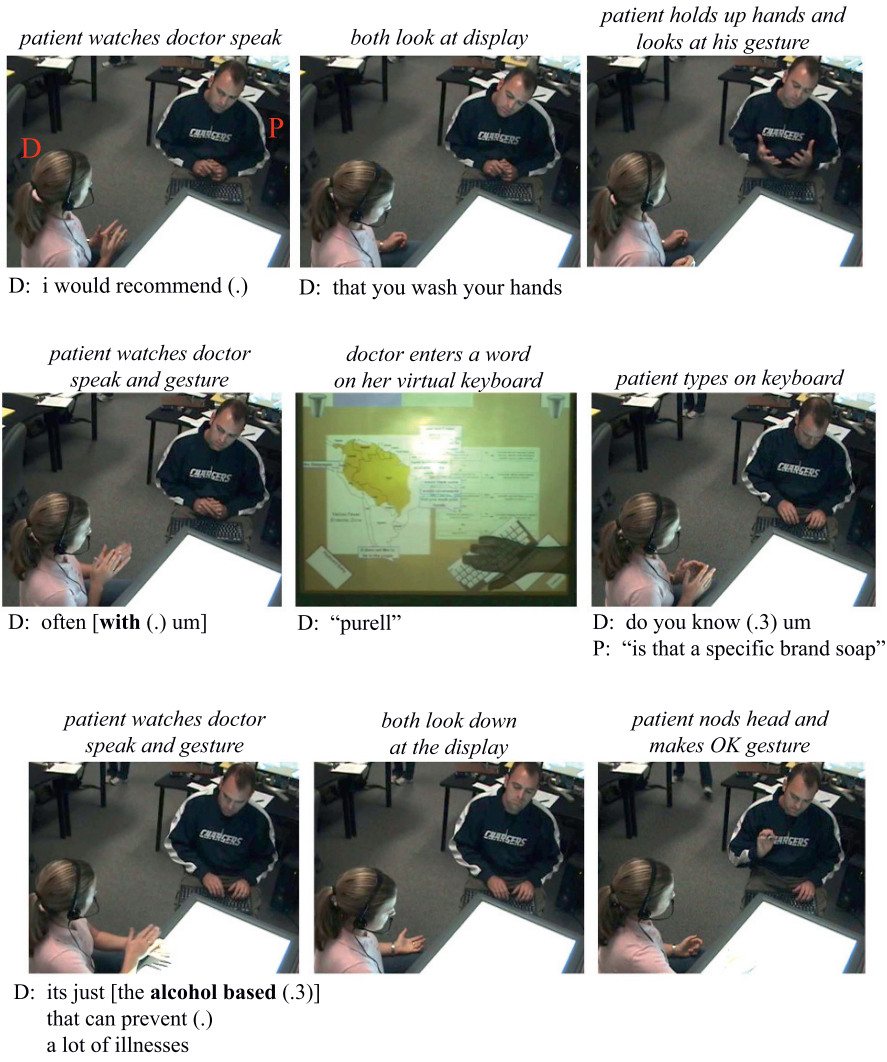


Figure 4. Doctor and patient communicate about hand washing through the digital table.

she says "with um." Then she uses the virtual keyboard to type the word "purell." The patient sees this word and responds by typing "Is that a specific brand soap?" His typing occurs simultaneously with the doctor's speech (middle row, right frame of Figure 4). The doctor's response (see Figure 4 bottom) demonstrates that she attends to the patient's question for clarification. A critical moment in this interaction occurs in the bottom left image of Figure 4. The doctor and patient make eye contact as the doctor performs an iconic hand rubbing gesture timed with the words "alcohol based." Her gesture communicates the method of use for hand san-

itizer, as alcohol-based sanitizers work by evaporating when rubbed into the hands. After this, both look down at the display to see the doctor's speech. Finally, the patient performs an emblematic "ok" gesture while nodding his head to show that he understands the doctor.

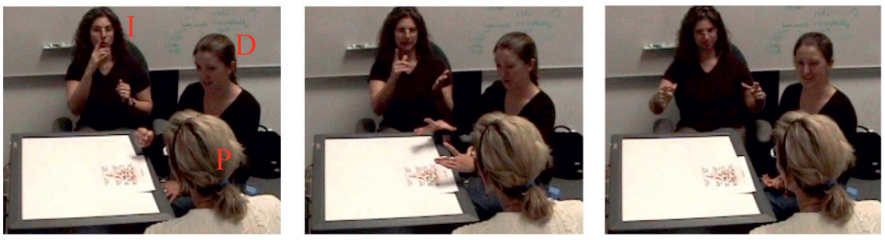
The doctor's carefully timed speech and gesture provide the patient with two pieces of information. First, her speech indicates the specific type of soap. Second, her gesture demonstrates how the soap is used. This information taken together yields a richer communicative form than either channel in isolation. This example demonstrates the importance of freeing the speaker's hands so that she is able to gesture as well as allowing the deaf individual to attend to the speaker's gestures instead of maintaining focus on the interpreter. In this example, and in others, we were struck by the highly coordinated use of speech, gesture, and eye gaze between the doctor and patient. The doctor's rich use of gesture to augment speech occurred often in interaction through the digital table. Similar use of gesture was *not* observed when the interpreter was present.

In a follow-up interview the doctor said that she intentionally tried not to gesture when the interpreter was present. She went on to explain that she did not want to compete with the interpreter for the patient's visual attention. In addition, interaction without the interpreter allowed the patient to frequently look at the doctor during communication, as is shown in Figure 3. This was a common pattern in the data. Having a larger percentage of the deaf patient's visual attention may have encouraged the doctor to elaborate her explanations with gesture (although this hypothesis needs to be examined with additional studies). Our analysis suggests that the multimodal tabletop system allows the doctor and patient to attend closely to each other's use of speech, gesture, and eye gaze as mechanisms for mediating communication. This also enables the doctor and patient to better monitor and exploit multiple modalities of communication such as co-occurring gesture-speech.

Monitoring Multiple Modalities of Communication

One challenge for deaf individuals involves monitoring multiple sources of visual information during conversation. Noticing and attending to co-occurring gesture-speech is a particularly challenging process when communication is mediated by an interpreter. Interpreter-mediated communication requires the deaf individual to notice co-occurring gesture-speech by the speaker and then put the speaker's gestures in context of the interpreter's gestural interpretation. Professionally trained interpreters are highly skilled, but they only occasionally replicate a speaker's gestures. Furthermore, through interviews with professional interpreters we found that their formal training does not specify when, if ever, they should replicate gestures made by the speaker. Overall, there were limited speech-gesture acts by the doctor in the interpreter condition, but this behavior did happen occasionally. Figure 5 is an example of the doctor using co-occurring gesture-speech. Here, she makes a fist like gesture (left) and then a two-handed iconic gesture (middle) to clarify portion size. Timing of speech and gesture is an issue, as the doctor completes each gesture

before the interpreter begins signing her speech. In this example, the interpreter did in fact recreate the doctor's gestures in context of her sign language interpretation but often the interpreter may not recreate the speaker's gesture, meaning that for at least a portion of communication the deaf individual must notice and attend to the speaker's gesture on their own. Even in cases in which the interpreter does recreate the gesture, it may not be formed or timed in exactly the same way as the original, thus creating interpretation challenges. In contrast, communication through the digital table provides opportunity for the deaf individual to look directly at the speaker's gestures, and as Figure 4 illustrates, gestures played an important role in establishing a shared understanding.



D: [**cooked pasta should only** be the size of your fist] (.4)

D: [not the **big bowls** (.)]

D: that are served in restaurants

Figure 5. Doctor uses gesture with her speech. Interpreter relays speech and gesture information..

Persistent, Collaboratively Produced Representations

Unlike other assistive technologies that enable communication between deaf and hearing individuals, the shared tabletop display provides a central focal point and space for establishing common ground (Clark and Brennan, 1991). The horizontal tabletop surface provides a space through which the doctor and patient cooperatively create, view, and manipulate representations of conversation. The shared conversation space allows the doctor and patient to gesture around and point to previous speech, thereby anchoring their gestures to objects (physical and virtual) in the environment (Clark, 2003). Referencing interface objects most often occurs through situated, context-specific pointing gestures (Goodwin, 2003). Both hearing and deaf participants used deixis to reference the material and symbolic world in front of them. With the interpreter, there is no record or explicit external representation of discourse. Consider Figure 6 (top row) where the doctor annotates a food pyramid diagram. Here, the doctor uses pointing gestures on a food pyramid diagram as she explains a balanced diet. The deaf patient must attend to both the interpreter's interpretation of speech as well as the doctor's pointing gesture occurring with her speech.

In this example, the doctor uses her speech and pointing gestures to walk the patient through parts of a food pyramid. Each time she points to a section of the

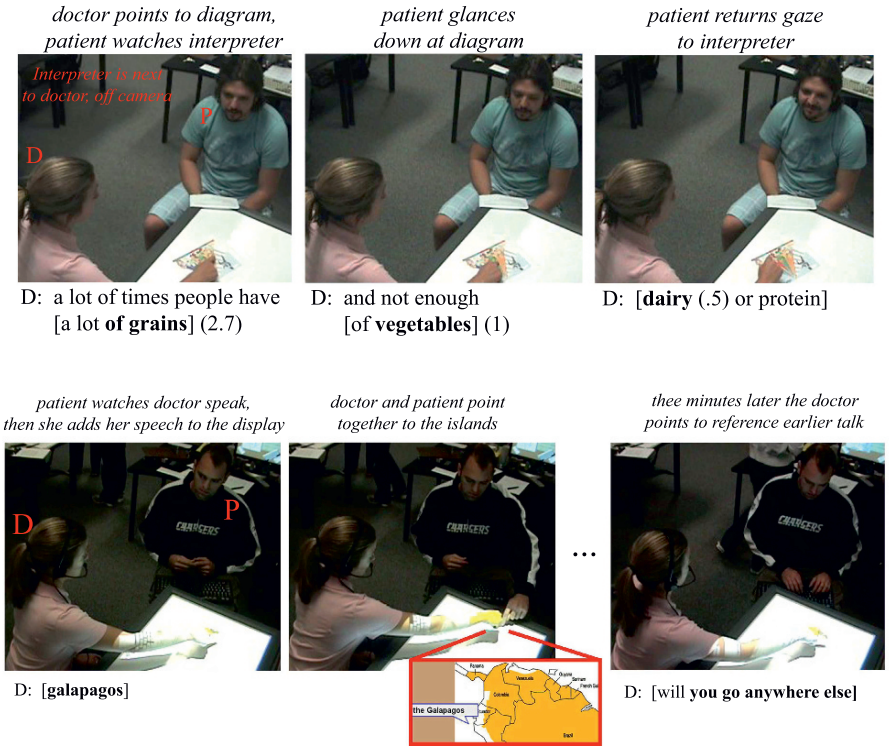


Figure 6. Top: Doctor points to parts of a diagram as she speaks. Patient monitors interpreter and doctor's pointing gestures. Bottom: Using the digital table, the Doctor labels the Galapagos Islands on the map and then points to the speech bubble three minutes later.

diagram, she shifts her gaze to the table, likely an attempt to draw her listener's attention to the diagram. Several minutes later the doctor references this diagram again to summarize her recommendation about a well-balanced diet, but the conversation and gestures she made to the patient are now only a memory.

The digital table stores collaboratively created representations of speech and allows users to rearrange and manipulate speech bubbles. Images in the top row of Figure 6 illustrate challenges with pointing to parts of a diagram while speaking; the digital table uniquely supports this form of interaction. We observed an interesting form of pointing that occurred through the strategic placement of speech bubbles. The tangible and persistent nature of speech bubbles affords certain interactions by serving as manipulatable cognitive artifacts (Hutchins, 1995). A speech bubble gains meaning beyond its literal text depending on how it is situated, or anchored, with respect to other parts of the activity. The canonical shape of speech bubbles, specifically the tail, allows the doctor and patient to use the objects as a pointing mechanism. That is, participants strategically placed speech bubbles around the display so that the tail of the speech bubble touched a relevant part of the background

or another speech bubble. Figure 6 (bottom center frame) provides an example of this behavior. In this interaction the doctor uses a speech bubble to label and reference part of a map. The patient mentions that he is traveling to the Galapagos Islands. The doctor says "Galapagos" as she points, and the patient points along with her to clarify the location. Subsequently, the doctor moves the "Galapagos" speech bubble to label the islands on the map. Then she uses this action to show that the islands are outside the Yellow Fever endemic zone (bottom center frame of Figure 6) and explain that the patient will not need the Yellow Fever vaccine. Conversation continues, and the topic changes. Approximately three minutes later the doctor comes back to "the Galapagos" speech bubble. She points to the speech bubble while asking, "will you go anywhere else?"

The persistent nature of speech along with the shared context of the tabletop display affords referencing both new and previously created external representations of speech. The persistent nature of speech also allows participants to review their entire conversation. Both the doctor and patients looked back over their previous conversation at some point during the activity. In a post-session interview, the doctor said, "It was good to look back at what I had covered with that particular patient," and explained that, "[The digital table] would be helpful because it is not uncommon in medicine to have very similar conversations with different patients throughout the day."

Discussion

Our analysis highlights differences in interaction between a deaf and hearing individual when communication is mediated by a multimodal tabletop display as compared to a human sign language interpreter. These differences reveal several tradeoffs. Although speech recognition technology can not yet provide the richness and accuracy associated with translation by a competent interpreter, it does allows the doctor to exploit gesture for communicative purposes without fearing that she might distract the deaf individual from the interpreter. One example is the hand washing iconic display coinciding with speech depicted in Figure 4. In addition, transcribed speech-to-text allows the doctor and patient to have a shared record of conversation. This provides new artifacts in the environment, enabling pointing and other gestures (Roth and Lawless, 2002). Removing the time-critical visual demands of interpreter-mediated communication allows the deaf individual to focus more on the doctor while she is speaking. In turn, this helps the patient attend to the doctor's speech-gesture acts and enables the doctor to better gauge patient understanding through increased eye contact. Speed of communication is another important tradeoff issue. Current speech recognition is no match for a skilled interpreter. When using the speech recognition system, the doctor must speak slowly and carefully in order to help ensure accurate recognition. Time is also taken in selecting the appropriate alternative from the output of the recognition system and in correcting it when required. But necessitating slower dialogue on the part of the doctor is not an entirely negative outcome. Considering that English is a second language for many

deaf individuals, slowing the doctor's speech could in fact be a positive cognitive consequence of communicating through the tabletop display.

The SSI system technology has the potential to benefit multiple user groups and enable new cooperative computing applications. Shared displays, especially tabletop displays, are beneficial for a variety of group work tasks. Since inception of our project and the idea to visually represent conversation on a tabletop display, members of the Deaf community have mentioned numerous contexts in which this could be useful. Of these, the most frequently identified are counseling or therapy sessions, banking and financial services, meetings with an architect or interior designer, ASL and lip reading education, classroom group work, and even retail environments. Beyond the Deaf community, the cognitive affordances of SSI have implications for individuals with moderate hearing loss as well as unimpaired hearing users. The challenge of medical conversations is certainly not restricted to the Deaf community. Because of associated stress and other factors, it is easy to forget details to tell the doctor and even easier to forget specific instructions given during consultation. The affordances of SSI such as preloading questions for the doctor and referencing a transcript of a previous conversation extend to all populations. Similarly the ability to archive and subsequently revisit past multimodal conversations and collaborations has interesting potential to augment interaction.

The concepts behind SSI also have specific implications for user populations with other language-related communication barriers. For example, representing speech on a shared display has pedagogical benefits for language learning. Consider a case in which speech bubbles store textual and auditory information from a native speaking teacher and a student learning a second language. Here, both textual and auditory representations can be accessed in a shared collaborative context. The availabilty of visual and spatial representations of language also stand to benefit individuals with linguistic processing disabilities such as Aphasia or Apraxia. Language could take on a variety of representations including textual, auditory, and pictorial forms. For these individuals and other populations, a shared, co-located workspace has considerable promise to help in establishing common ground and assisting communication.

Conclusions and Future Work

Analysis of multimodal human interaction data is primarily used in ethnographic approaches to understanding everyday activity (e.g., Goodwin, 1981; Heath, 1986), but there is a growing interest in using multimodal analysis to understand the role of gesture occurring in experimental cooperative work settings (Bekker et al., 1995; Kraut et al., 2003; Kirk et al., 2005). We suggest that multimodal analysis can aid laboratory evaluations of tabletop technology as well as other cooperative work technologies in the following ways: (1) analysis of eye gaze provides a metric for understanding how people coordinate visual attention, (2) evaluation of gesture types and frequency of use provides a way to measure differences in interaction between experimental conditions, and (3) the interplay between speech, gesture, and

eye gaze can reveal cognitive and social consequences of new interactive media that would be difficult to detect with other methods.

Multimodal analysis, however, is tedious and extremely time-consuming. When analysis is so difficult, few analyses can be done and datasets are severely underutilized. Researchers come to have a large investment in the chosen data segments. Since each analysis may appear as an isolated case study, it can be difficult to know how common the observed phenomena may be. Larger patterns and contradictory cases can easily go unnoticed. Well-known human confirmation biases can affect the quality of the science when each analysis requires so much effort. The analyses presented in this paper, for example, resulted from a year-long iterative process of analysis of video and audio data to understand how differing communication media shapes interaction. This form of detailed analysis plays an increasingly central role in our ongoing investigation of tabletop display systems. One way our research group is addressing the difficulties of such analysis is by exploring techniques to assist video analysis. Applying computer vision techniques make it possible to tag video frames with certain characteristics of interest such as movement of hands or arms. We are currently evaluating computer vision methods for object recognition, face detection, and head pose estimation. For example, SIFT (Scale Invariant Feature Transform) (Low, 2004) is one popular and useful technique we are exploring. We see tremendous potential for computer vision techniques to assist video analysis for the types of data we report here and are exploring this as part of our ongoing work.

In this paper we have examined communication about medical topics through *Shared Speech Interface*, a multimodal tabletop display designed to assist communication between a hearing and deaf individual by converting speech-to-text and representing dialogue history on a shared interactive display surface. Specifically, we compared communication mediated by a multimodal tabletop display and by a human sign language interpreter. Results indicate that the multimodal tabletop display (1) allows the deaf patient to watch the doctor when she is speaking, (2) encourages the doctor to exploit multimodal communication such as co-occurring gesture-speech, and (3) provides shared access to persistent, collaboratively produced representations of conversation. Finally, we discuss extensions of our system and practical aspects of conducting a multimodal analysis of tabletop interaction.

Acknowledgments

Research is supported by a NSF Graduate Research Fellowship, NSF Grant 0729013, and a Chancellor's Interdisciplinary Grant. We thank our study participants, faculty and staff from UCSD Medical School, Whitney Friedman, and MERL for donating a DiamondTouch table.

References

Argyle, M. and M. Cook (1976): *Gaze and mutual gaze*. Cambridge University Press.

Baker, C. (1977): 'Regulators and turn-taking in American Sign Language discourse'. In: *On the other hand: New perspectives on American Sign Language*. New York, pp. 215–236, Academic Press.

Baker, C. and C. Padden (1978): 'Focusing on the nonmanual components of American Sign Language'. In: *Understanding Language through Sign Language Research*. New York, pp. 27–57, Academic Press.

Beattie, G. and H. Shovelton (1999): 'Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation'. *Semiotica*, vol. 123, no. 1-2, pp. 1.

Bekker, M. M., J. S. Olson, and G. M. Olson (1995): 'Analysis of gestures in face-to-face design teams provides guidance for how to use groupware in design'. In: *Proceedings of conference on Designing Interactive Systems (DIS)*. pp. 157–166.

Cassell, J., D. McNeill, and K.-E. McCullough (1999): 'Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information'. *Pragmatics cognition*, vol. 7, pp. 1.

Cavender, A., R. E. Ladner, and E. A. Riskin (2006): 'MobileASL: intelligibility of sign language video as constrained by mobile phone technology'. In: *Proceedings of conference on Computers and Accessibility (ASSETS)*. pp. 71–78.

Clark, H. (2003): 'Pointing and Placing'. In: S. Kita (ed.): *Pointing: Where Language, Culture, and Cognition Meet*. Mihwah, NJ, pp. 243–268, Lawrence Erlbaum Associates.

Clark, H. and S. Brennan (1991): 'Grounding in Communication'. In: L. Resnick, J. Levine, and S. Teasley (eds.): *Perspectives on Socially Shared Cognition*. Washington, APA Books.

Dietz, P. and D. Leigh (2001): 'DiamondTouch: a multi-user touch technology'. In: *Proceedings of symposium on User Interface Software and Technology (UIST)*. pp. 219–226.

Duncan, S. (1972): 'Some signals and rules for taking turns in conversations'. *Journal of personality and social psychology*, vol. 23, no. 2, pp. 283.

Duncan, S. (1974): 'On the Structure of Speaker-Auditor Interaction During Speaking Turns.'. *Language in society*, vol. 3, no. 2, pp. 161.

Duncan, S. and D. W. Fiske (1977): *Face-to-face interaction: Research, methods and theory*. Hilldale, NJ: Lawrence Erlbaum Associates.

Emmorey, K., R. Thompson, and R. Colvin (2008): 'Eye Gaze During Comprehension of American Sign Language by Native and Beginning Signers'. *Journal of Deaf Studies and Deaf Education*.

Engberg-Pedersen, E. (2003): 'From Pointing to Reference and Predication: Pointing Signs, Eyegaze, and Head and Body Orientation in Danish Sign Language'. In: S. Kita (ed.): *Pointing: Where Language, Culture, and Cognition Meet*. Lawrence Erlbaum Associates.

Goldin-Meadow, S. (2003): *Hearing Gesture: How our Hands Help Us Think*. Harvard University Press.

Goodwin, C. (1981): *Conversational Organization: Interaction Between Speakers and Hearers*. New York: Academic Press.

Goodwin, C. (2000): 'Practices of Seeing, Visual Analysis: An Ethnomethodological Approach'. In: *Handbook of Visual Analysis*. London, pp. 157–182, Sage.

Goodwin, C. (2003): 'Pointing as Situated Practice'. In: S. Kita (ed.): *Pointing: Where Language, Culture, and Cognition Meet*. Lawrence Erlbaum Associates.

Goodwin, C. and M. Goodwin (1996): 'Formulating Planes: Seeing as Situated Activity'. In: *Cognition and Communication at Work*. pp. 61–95, Cambridge University Press.

Gullberg, M. and K. Holmqvist (2006): 'What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video'. *Pragmatics cognition*, vol. 14, no. 1, pp. 53.

Heath, C. (1986): *Body movement and speech in medical interaction*. Cambridge University Press.

Hindmarsh, J. and C. Heath (2000): 'Embodied reference: A study of deixis in workplace interaction'. *Journal of Pragmatics*, vol. 32, no. 12, pp. 1855.

Hollan, J. D., E. Hutchins, and D. Kirsh (2000): 'Distributed cognition: toward a new foundation for human-computer interaction research'. *ACM transactions on computer-human interaction*, vol. 7, no. 2, pp. 174–196.

Hutchins, E. (1995): *Cognition in the Wild*. Cambridge, MA: MIT Press.

Kendon, A. (1967): 'Some functions of gaze-direction in social interaction.'. *Acta Psychologica*, vol. 26, no. 1, pp. 22–63.

Kendon, A. (1980): 'Gesticulation and Speech: Two Aspects of the Process of Utterance'. In: *The Relationship of Verbal and Nonverbal Communication*. p. 388, Walter de Gruyter.

Kendon, A. and C. Muller (2001): 'Introducing: GESTURE'. *Gesture*, vol. 1, no. 1, pp. 1.

Kirk, D., A. Crabtree, and T. Rodden (2005): 'Ways of the hands'. In: *Proceedings of European Conference on Computer Supported Cooperative Work (ECSCW*. pp. 1–21.

Kita, S. (2003): *Pointing: where language, culture, and cognition meet*. Lawrence Erlbaum Associates.

Kleinke, C. (1986): 'Gaze and eye contact: A research review'. *Psychological Bulletin*, vol. 100, no. 1, pp. 78–100.

Kraut, R. E., S. R. Fussell, and J. Siegel (2003): 'Visual information as a conversational resource in collaborative physical tasks'. *Hum.-Comput. Interact.*, vol. 18, no. 1, pp. 13–49.

Liddell, S. K. and M. Metzger (1998): 'Gesture in sign language discourse'. *Journal of Pragmatics*, vol. 30, no. 6, pp. 657 – 697.

Low, D. G. (2004): 'Distinctive image features from scale-invariant keypoints'. *International Journal of Computer Vision*, vol. 60, pp. 91–110.

Luff, P., C. Heath, H. Kuzuoka, K. Yamazaki, and J. Yamashita (2006): 'Handling documents and discriminating objects in hybrid spaces'. In: *Proceedings of the conference on Human Factors in Computing Systems (CHI)*. pp. 561–570.

McNeill, D. (1985): 'So You Think Gestures Are Nonverbal?'. *Psychological review*, vol. 92, no. 3, pp. 350.

McNeill, D. (1992): *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Miller, D., K. Gyllstrom, D. Stotts, and J. Culp (2007): 'Semi-transparent video interfaces to assist deaf persons in meetings'. In: *ACM-SE 45: Proceedings of the 45th annual southeast regional conference*. New York, NY, USA, pp. 501–506, ACM.

Morris, M. R. (2006): 'Supporting Effective Interaction with Tabletop Groupware'. Ph.D. thesis, Stanford University, Stanford, CA.

NIDCD (2008): 'National Institute on Deafness and Other Communication Disorders'. http://www.nidcd.nih.gov.

Piper, A. M. and J. D. Hollan (2008): 'Supporting medical conversations between deaf and hearing individuals with tabletop displays'. In: *Proceedings of the conference on Computer-Supported Cooperative Work (CSCW)*. pp. 147–156.

Rogers, Y., W. Hazlewood, E. Blevis, and Y.-K. Lim (2004): 'Finger talk: collaborative decision-making using talk and fingertip interaction around a tabletop display'. In: *CHI '04: CHI '04 extended abstracts on Human factors in computing systems*. New York, NY, USA, pp. 1271–1274, ACM.

Rogers, Y. and S. Lindley (2004): 'Collaborating around vertical and horizontal large interactive displays: which way is best?'. *Interacting with Computers*, vol. 16, no. 6, pp. 1133 – 1152.

Roth, W.-M. (2001): 'Gestures: Their Role in Teaching and Learning'. *Review of Educational Research*, vol. 71, no. 3, pp. 365–392.

Roth, W.-M. and D. V. Lawless (2002): 'When up is down and down is up: Body orientation, proximity, and gestures as resources'. *Language in Society*, vol. 31, no. 01, pp. 1–28.

Schull, J. (2006): 'An extensible, scalable browser-based architecture for synchronous and asynchronous communication and collaboration systems for deaf and hearing individuals'. In: *Proceedings of the conference on Computers and Accessibility (ASSETS)*. pp. 285–286.

Shen, C., F. D. Vernier, C. Forlines, and M. Ringel (2004): 'DiamondSpin: an extensible toolkit for around-the-table interaction'. In: *Proceedings of the conference on Human Factors in Computing Systems*. pp. 167–174.

Siple, P. (1978): 'Visual Constraints for Sign Language Communication'. *Sign Language Studies*.

Tang, J. C. (1991): 'Findings from observational studies of collaborative work'. *International Journal of Man-Machine Studies*, vol. 34, no. 2, pp. 143 – 160. Special Issue: Computer-supported Cooperative Work and Groupware. Part 1.

Tang, J. C. and S. L. Minneman (1990): 'VideoDraw: a video interface for collaborative drawing'. In: *Proceedings of the conference on Human Factors in Computing Systems (CHI)*. pp. 313–320.

Tse, E. (2007): 'Multimodal Co-Located Interaction'. Ph.D. thesis, The University of Calgary, Calgary, Alberta, Canada.

Tse, E., C. Shen, S. Greenberg, and C. Forlines (2007): 'How pairs interact over a multimodal digital table'. In: *Proceedings of the conference on Human Factors in Computing Systems (CHI)*. pp. 215–218.

Turoff, M. (1975): 'Computerized conferencing for the deaf and handicapped'. *SIGCAPH Comput. Phys. Handicap.*, no. 16, pp. 4–11.

Zemel, A., T. Koschmann, C. Lebaron, and P. Feltovich (2008): "'What are We Missing?" Usability's Indexical Ground'. *Computer Supported Cooperative Work*, vol. 17, no. 1, pp. 63–85.