Human-Centered Collaborative Interaction

Paulo Barthelmess, Edward Kaiser, Rebecca Lunsford
David McGee, Philip Cohen, Sharon Oviatt
Natural Interaction Systems
Seattle, WA, USA

ABSTRACT

Recent years have witnessed an increasing shift in interest from single user multimedia/multimodal interfaces towards support for interaction among groups of people working closely together, e.g. during meetings or problem solving sessions. However, the introduction of technology to support collaborative practices has not been devoid of problems. It is not uncommon that technology meant to support collaboration may introduce disruptions and reduce group effectiveness.

Human-centered multimedia and multimodal approaches hold a promise of providing substantially enhanced user experiences by focusing attention on human perceptual and motor capabilities, and on actual user practices. In this paper we examine the problem of providing effective support for collaboration, focusing on the role of human-centered approaches that take advantage of multimodality and multimedia. We show illustrative examples that demonstrate human-centered multimodal and multimedia solutions that provide mechanisms for dealing with the intrinsic complexity of human-human interaction support.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Collaborative Computing; H.5.2 [**User Interfaces**]: User-Centered Design, Theory and Methods, Interaction Styles, Input Devices and Strategies, Evaluation/Methodology, Voice I/O, Natural Language, User Interface Management Systems

General Terms

Design; Human Factors

Keywords

Human-Centered Systems, Multimodal Systems, Multimedia, Design Guidelines.

1. INTRODUCTION

Support for groups of people working together has been increasingly attracting research interest, and for good reason. A considerable amount of resources is spent in meetings and collaborative sessions of various natures. These sessions are at the same time essential for the operation of most organizations as well as expensive in terms of resources, given the accumulated time participants spend in them.

1.1 Challenges of the Domain

Group interaction support brings about new problems. Interface paradigms and approaches used successfully in single-user systems are not readily applicable [13]. Groups working closely together present complex and not well understood dynamics. Social aspects become particularly relevant and may provide the primary motivation for some collaborative interactions.

Priority is naturally focused on the human to human exchanges, at the expense of computational systems. Intrinsic cognitive demands may be very high due to resources spent on articulating public presentations, dealing with problem solving and decision-making, and handling potential social conflicts. Controlling an interface becomes a secondary task, competing for resources against demanding primary group tasks. The technology that should be supporting the group coul very easily become a hindrance.

One of the main sources of problems for collaborative systems is the potential disruption of essential social processes [13]. Changes to the process or to the pace of interaction, e.g. by lengthy interface operations that require the group to break the flow of interaction to accommodate system requirements, may compromise collaboration. Potential technological misfit might force shifts from a primary task to the secondary, interface-related one. As participants have to stop their work to develop strategies to cope with technology problems, e.g. a connection failure, or a convoluted, time-consuming input mechanism, resources are allocated to what corresponds to purely extraneous costs [26].

These factors conspire to make it challenging to introduce group support technology. Not surprisingly, many group applications fail [13, 30]. The usual consequence of system perceived problems is rejection. It is not unusual for groups to circumvent systems and operate through conventional "low-tech" means, particularly for co-located collaboration, but also in remote scenarios (e.g. by falling back to using phone conferencing).

While groups can be very successful at collaborating in the absence of system support, that does not happen without consequences. Collaborative sessions are focal points that have lasting repercussions and impact on future work of the participants. The absence of system support during the group sessions results in the history and potential products of an interaction being lost, except for those parts of it that may survive as disperse, manually provided entries made by participants into a variety of other systems, e.g. calendars, and spreadsheets. Decisions made at meetings, action items assigned and design products are therefore either forgotten or ignored, or require additional individual labor to make it back into

HCM'06, October 27, 2006, Santa Barbara, California, USA.

computational systems [26]. Manual reentry of information is of course costly, error-prone and potentially incomplete.

A challenge is therefore how to make aspects of an interaction visible to systems, so that additional post-interaction support may be provided while avoiding interface costs that may lead to technology induced disruption of work.

1.2 Human-centered collaboration

Human-centered approaches advocate a deeper understanding of human-perceptual and motor capabilities and limitations as a basis for designing more effective systems [18, 41, 39]. The approach furthermore promotes support for the performance of actual user practices [17, 41], and strives to amplify and extend human capabilities [41, 39].

Studies of natural collaborative practices (e.g. [42]) reveal a rich communicative scenario. Participants of collaborative interactions speak, write, sketch and express themselves via gestures, facial expressions and other body motions. Such communication is therefore eminently *multimodal*, as it is performed via a variety of redundant and complementary communication modalities [31]. Such interactions are also commonly *multimedia* [18] in the sense that they are conducted and supported via a wide variety of media and materials ranging from electronic to physical, e.g. paper.

A high synergy can therefore be seen to exist between human-centered approaches and multimodal and multimedia techniques. Human collaboration revolves around rich and complex multimodal and multimedia communication and work performance. Successful solutions can only emerge if a human-centered, detailed understanding of this intrinsic human behavior can be brought to bear.

Grudin [14] has observed incremental phases of interface development, starting with the original hardware-based ones, to ones based on programming languages, to interactive systems used by end-users. A continued evolution would lead to systems based on deeper cognitive foundations, moving towards interfaces that extend into the social and work environments [14]. In a metaphorical sense, computer interfaces would "extend beyond the keyboard and display surface [...] into the mind of the user" [14, p.264] and the environment. Advanced human-centered multimodal and multimedia systems take steps forward in this evolution.

Of particular interest to the discussion presented in this paper is the transformed role of systems, from passive, command-oriented to pro-active and observant. While the former traditional systems require users to provide them every minute, detailed input, the latter "reach out", taking responsibility for analyzing natural human communication and understanding work patterns to autonomously extract the information they require to be of service to their users without introducing additional burdens.

Traditional systems require users to provide information via interfaces structured to facilitate interpretation, usually by constraining user input to unambiguous choices represented by menu items or button widgets. This type of interfaces is convenient from a system's perspective, but fall short when tasks (such as collaborative ones) do not fit into the limited button pushing / passive paradigm.

We are therefore interested in systems that perform recognition of natural communication, such as speech, handwriting, sketching and natural gestures. The interest is on investigating how modalities and media can be exploited to construct systems that take proactive responsibility for identifying user intentions and automatically extract input from unencumbered human-human communication that is observed by the system, with the goal of achieving the seamless support to natural practices prescribed by the human-centered approach.

1.3 Paper organization

In this paper we discuss how multimedia and multimodality can provide human-centered solutions to the challenges of supporting collaboration:

- We start by discussing the role of media in collaboration, and show how the integration of tangible materials - particularly digital paper - can have a positive impact on user performance (Section 2).
- We then proceed to show in Section 3 how multimodality can be leveraged to implicitly track intentions and recover semantics from natural human-to-human communication.
- Finally, in Section 4 we introduce interface paradigms that aim at reducing the cognitive impact of interface operation by exploring autonomous, pro-active observation of user behavior and automatic extraction of information.

We illustrate the approaches with results obtained by the group over the last few years.

2. THE ROLE OF MULTIMEDIA

In this section we examine the role multiple media play in supporting user practices and providing interfaces that promote high task performance. Here we use multimedia as defined by Jaimes [18] - as "a combination of digital, analog, spatial, and sensory inputs and outputs".

We discuss here the importance and impact of supporting a variety of materials, in particular those tangible materials that are commonly used in actual practices, such as paper.

We start by discussing the limitations of conventional interfaces in supporting collaboration (Section 2.1). Section 2.2 discusses the use of digital paper, an enhanced tangible material that can be used to bridge conventional and digital media. The cognitive impact of different media is then examined in Section 2.3. We show results that indicate that the media itself affects performance particularly of more challenged individuals.

2.1 Limitations of Conventional Interfaces

The use of conventional interfaces to support group work has proved problematic. Except perhaps for very small two-person groups, conventional single user interfaces do not provide the required affordances for group input and output. Keyboard and mouse are designed primarily for single user operation; computer displays are meant to be viewed primarily by a single person at a time.

The naïve solution based on providing each participant with her own single-user interface is unnatural in most co-located settings. Screens partially occlude other participants and prevent full access to gestures and facial expressions that might provide deeper, direct insights into the intended meaning of communication, which is the reason why people gather to collaborate. The focus switches to the computational devices at the expense of the overall human-human interaction. This is particularly harmful in interactions that require a high degree of collaboration, such as design or problem solving sessions.

Conventional interfaces also fall short with respect to operational requirements. A study comparing a conventional interface for map-based military planning task against a multimodal one – Quickset [7] found that the multimodal interface provided an almost four-fold increase in speed compared to the conventional one [9]. While this experiment involved single users, the lesson is nevertheless applicable to collaborative scenarios, in which the speed of operation is of fundamental importance, given that lengthy operational delays may endanger the interaction by disrupting its pace.

To overcome these limitations, group-friendly devices and interfaces have been developed, e.g. Stanford's Interactive Mural [15], DynaWall [11], and HoloWall [37]. While these solutions provide interesting functionality, they are limited in the sense that they are restricted to instrumented meeting rooms. Other limitations have to do for instance with the number of simultaneous users that can write simultaneously on these boards, usually restricted to only one at a time (an exception is DiamondTouch [10], which allows for multiple simultaneous users to operate on a tabletop configuration).

We have been exploring a lighter-weight, mobile alternative that fits well within current user work practices - digital paper. We discuss this in further detail in the next section.

2.2 Multimodal Interaction with Paper

Paper, as one of the oldest communication and collaboration devices presents many advantages. It is cheap, light weight, and high-definition. Paper is also robust – it requires no power (is always on), and will still work even if a hole is punched through its middle or if it gets crumpled or wet. The same cannot be said of electronic equipment [8]. Not surprisingly, paper is ubiquitous. Its use, contrary to predictions is on the rise [38].

We have been pursuing multimodal solutions that are based on the prevalent use of paper in the workplace. These solutions (such as Rasa, NIS Map and NIS Chart) build upon existing paper-based practices such as those used by the military and in health care institutions [8]. The overall objective is to support existing physical workplace routines and languages, allowing users to operate in their customary fashion while simultaneously updating a digital version of the information [8]. As an example, take Rasa [28], a system that supports military officers working on command and control tasks. Officers maintain awareness of an evolving operation by annotating a large-scale wall-mounted paper map with Post-it$^{\text{TM}}$ notes onto which military symbols are hand-drawn, e.g. representing a platoon. As reports are received, existing notes might be moved, or new ones might be introduced. Rasa adds an observant computational layer to the unaltered procedure. Sketches on the Post-it notes are captured and interpreted, along with spoken utterance further characterizing the symbol, e.g. the identity of the platoon whose symbol was just sketched. The interpreted information is propagated to other system while users perform their customary manual process. More importantly, Rasa is robust to failures. In case of disruption of the computational systems (emulated in a study [29]), operation can proceed uninterrupted. Once the system resumes, the last known situation is projected against the map, and users can quickly re-synchronize.

Our multimodal paper applications use Anoto Functionality [3]. This technology employs a pen with an embedded camera and processor, which is capable of identifying the absolute position of tip by examining tiny black dots placed on a grid, forming a pattern. The decoded information is stored in memory by the pen, and may also be transmitted in real-time via a Bluetooth connection. The pattern can be printed on regular paper; the pen produces ink like any conventional pen. Anoto-based digital stylus and paper interfaces, which span the physical and digital worlds, also are a promising interface for knowledge-gathering tasks in which users need to combine, cross-reference, and personalize information from different sources with pen-based annotations.

2.3 Cognitive advantages of paper

Very recent work compared non-computational work practice for mathematics education (i.e., paper and pencil work practice) with different interface alternatives (i.e., an Anoto-based digital stylus and paper interface (DP), a pen tablet interface (PT), and a graphical tablet interface (GT)). The results revealed that the same *students completing the same geometry problem solving activities* experienced greater load as interfaces departed more from existing work practice (GT $>$ PT $>$ DP), which was evident in a constellation of performance indices showing degradation in speed, attentional focus, meta-cognitive control, correctness of problem solutions, memory, and fluency [32]. In addition, lower-performing students experienced elevated cognitive load, with the more challenging interfaces (GT, PT) disrupting their performance disproportionately more than high performers. More specifically, students were significantly faster and more attentive to the math when using the digital stylus and paper interface (i.e., the *tangible paper-based interface*), compared with either of the tablet interfaces. Low-performing students also remembered math information better after using the digital stylus and paper interface than either of the tablet interfaces. On the other hand, the two *pen-based interfaces* (DP, PT) supported better planning and meta-cognitive control of students' work (i.e., indicated by advance diagramming and high-level math comments) than the graphical tablet interface (GT). Also, low-performing students solved more problems correctly with the pen-based interfaces, and high-performing students expressed themselves more fluently with them (i.e., using symbolic, diagrammatic, linguistic, and numeric representational systems).

The enhanced performance of the digital stylus and paper interface can be associated with the fact that this interface was the most similar to students' existing hardcopy pencil and paper work practice, a finding also consistent with the predictions of Cognitive Load Theory [40]. In particular, it incorporated pen input rather than a keyboard and mouse, and also the familiar and tangible paper medium. In comparison, the pen tablet interface included the familiarity of a pen but not the paper medium, and the graphical interface least resembled students' existing work practice. Within the math domain, both the digital stylus and pen tablet interfaces also support a broad range of expressive input in different representational systems, including linguistic, numeric, symbolic, and diagrammatic. Such pen interfaces are particularly compatible with complex problem solving in domains like mathematics, which requires input fluency in all four representational systems and flexible translation among them (e.g., from word problems, to diagrams, to algebraic formulas). In contrast, whereas graphical interfaces provide good support for linguistic and numeric content, symbolic and diagrammatic input is poorly supported or not at all. Other attractive characteristics of the pen-based interfaces include their suitability for collaboration, mobility, and "bridging" of formal, informal, and mobile contexts.

3. EXPLOITING MULTIMODAL COMMUNICATION

As discussed in the Introduction, systems that take advantage of naturally occurring multimodal communication e.g. by interpreting users' speech, handwriting, sketching and gestures have a potential for better supporting users in challenging collaborative tasks. Studies (e.g. [33, 12]) point to the flexibility and economy afforded by multimodal language, particularly when the cognitive demands are high [33]. Such advantages may be derived from the underlying human cognitive architecture, which would be intrinsically multimodal [4, 36].

Natural multimodal language reveals a variety of semantic cues that can be exploited as mechanisms for automatic detection of intentions and meaning without requiring direct user intervention. In this section we examine some of these techniques.

Section 3.1 introduces evidence that in some collaborative situations people will present information redundantly, e.g. by handwriting and speaking the same terms. We demonstrate how this behavior can be leveraged to provide robust cross-modal recognition, including the recovery of out-of-vocabulary terms and abbreviations. In Section 3.2 we examine the role of speech amplitude, semantic content and gaze as indicators of whether a speaker is addressing a computer or a fellow participant.

3.1 Recovering meaning from redundant language

In some human-human interactions information is presented redundantly across multiple modes, for example via handwriting and speech. To describe this phenomenon we use the term *multimodal redundancy*, which we define to mean the simultaneous or sequential delivery of the same semantic information in more than one mode.

It has been shown, in the context of multimodal map-based and form-filling tasks, that speech and handwriting co-occur redundantly for only between 1%-5% of interactions [35, 34, 16]. On the other hand, in the educational-technology literature on human-human, computer-mediated interactions, like the presentation of distance-learning lectures, as much as 15% of all pen interactions were found to be handwriting [2], and a follow-on study to that work found that 100% of randomly sampled instances of handwritten lecture text were accompanied by semantically redundant speech [1]. Thus, when humans believe they are directly addressing a computer the current evidence is that they use multiple modes for presenting their input in a complementary rather than redundant fashion, but in contexts where the computer is a mediator or observer of natural multiparty interactions then redundancy in human-human multimodal presentation does occur.

3.1.1 Empirical evidence

To further document the occurrence of multimodal redundancy, we have analyzed two additional data collections. For the first, we recorded a spontaneous whiteboard and flip-chart brainstorming session, which occurred during a two-day project-planning meeting with 20 participants and lasted for 1.5-hours. We annotated all visible handwriting events for pen-down/pen-up timing, and transcribed all speech related to the handwriting. There were 41 recoverable handwriting events, and 40 of them (98%) were accompanied by semantically redundant speech.

For these 40 instances of multimodal redundancy, we determined the percent of handwriting and speech matches in each of the categories shown in Table 1. For exact matches (76% of instances) the handwritten text was repeated verbatim in the speech. For approximate matches (15% of instances) the handwritten text differed from the associated spoken phrases by extra words, tense differences, or order differences. For abbreviation exact matches (5% of instances) the handwriting included an abbreviation whose semantic meaning was spoken verbatim. For almost exact matches (2% of instances, not shown in Table 1) the speech differed only in number from the handwriting. There was, as mentioned above, one non-match, which occurred due to an inaudible speech elision.

These results closely parallel the results found for speech and handwriting redundancy in distance lecture delivery. That study found 100% multimodal redundancy with a 74% exact match [1], while we have found 98% multimodal redundancy with a 76% exact match. It is also interesting to note that while for the lecture data the speech associated with the handwriting came always from the presenter (who was also the handwriter), that was not the case for our planning session data. Here only 52% of the speech accompanying the handwriting was spoken by the handwriter. The other 48% came from seven of the other meeting participants. The percent of contributions from each of those seven roughly matched their positions in the organizational hierarchy underlying the meeting; so, the project manager's contributions were greatest (14%) followed by those of the team leads (10%) and then of the project engineers (5%).

3.1.2 Applications

We have been exploring this phenomenon to develop a proof-of-concept system that can leverage such richly informative events to dynamically learn new words and dynamically adapt the systems underlying recognition vocabulary.

SHACER (our Speech and HAndwriting reCognizER) implicitly tracks natural multimodal communication while presenters or lecturers handwrite terms and redundantly say them. That redundant multimodal event is both a natural part of multiparty human-human interactions, and a rich seed-event for adaptive system understanding.

There are several ways in which a system, capable of perceiving such redundant multimodal events can better understand the interactions it is tracking:

1. Combining information from the individual recognition streams (e.g. handwriting and speech recognition) can yield better recognition for known terms and serve as a trigger event for dynamically learning new vocabulary and enrolling it in system dictionaries [22, 24, 23, 19].

2. Our initial analysis of picture annotation data [6] indicates that redundant multimodal terms occur more frequently than unimodal terms: in this situation an observant system could better track topic changes over time, which in turn could support improved topic or key-word based searching over archived interaction records.

3. A system can recognize participants' deictic 3D gestures towards visually persistent handwriting on a shared space, and unobtrusively enrich the remote presentation of that pointing event by adding an appropriate remotely perceivable representation as shown in Figure 4 [5]. With such support for distributed pointing presentation the remote participants can avoid the difficult cognitive task of trying to determine the referent under discussion when they cannot perceive the deictic pointing event that identifies it. An observant system, having tracked a deictic event and recognized its spatially located handwritten referent can further ease the remote participants cognitive load by presenting the learned semantics of the handwritten term (as a hover label) in the case where it is an abbreviation.

Table 2 shows test results for multimodal recognition across handwriting and speech for Gantt chart labels in both a development test set series and held-out test set series. Two recognition conditions are shown: (1) ink-only (which corresponds to recognition results like those shown in the middle column of Figure 1 - for the ink shown in the left column of Figure 1), and (2) speech and ink combined (which corresponds to recognition results like those shown in the right column of Figure 1 that combines handwriting recognition with speech recognition - and by virtue of that combination also adds semantic labels to the abbreviations).

These results illustrate the efficacy of integrated recognition within a cumulative-observant multimodal interface. By focusing the

Match Type	%	Handwriting	Speech
Exact	76%	**Goals**	*... ahhhh set up* **goals ...**
Approximate	15%	**head nodding**	*Notice I indicated my assent by* **nodding my head.**
Abbrv. Exact	5%	**Information Q's**	*...will be* **informative questions.**

Table 1: Primary categories of *multimodal redundancy* across handwriting and speech, during a 1.5 hours spontaneous whiteboard and flip-chart brainstorming session. (not shown: Almost exact - 2%, and Non-matching - 2%.

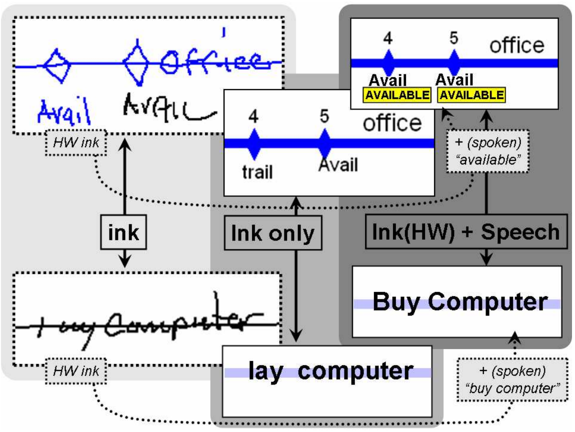

Figure 1: Sample of collected ink (left column), ink-only recognition results (middle column), and recognition results from combining information from both handwriting and speech recognition, with semantic abbreviation labels shown in the upper right.

Meeting Series	Development Test Set		Held-Out Test Set	
Chart Label Type	non abbrev	abbrev/ sem	non abbrev	abbrev/ sem
Total	27	24/24	49	16/16
Ink Only	18	12/0	27	6/0
Speech/Ink	24	20/20	27	6/6
Relative Error reduction	66%	75%	0%	37.5%

Table 2: Summary results on development test set and held-out test set. Abbreviated labels are considered correctly recognized only when both abbreviation spelling and semantics (abbrev/sem) are correct.

system's perceptive capabilities on human communicative modes (e.g. handwriting and speech), which naturally occur redundantly, we can apply cross-stream correlation at the recognizer level to improve the systems understanding, as evidenced by the significant 37.5% relative reduction in abbreviation labeling error seen in the held-out test set [20].

3.2 Detecting intended addressee

In contexts in which the computer system is an active assistant, expected to recognize and respond to user's spoken instructions immediately, there remains the problem of determining whether a speaker is addressing the computer assistant or another human participant. Towards this end, there currently is considerable interest in developing new open microphone engagement techniques that can perform robustly in noisy multi-party environments. State-of-the-art open microphone engagement systems aim to eliminate the need for explicit engagement by processing user's implicit cues of intended addressee, resulting in reduced cognitive load.

In terms of multi-party computer-assisted interactions, much recent research has been predicated on the premise that people will interact with a computer much as with a subordinate human peer, directing their gaze to the computer when addressing it while employing command-like speech content and prosody [43, 25]. However, in pursuing this course of work, researchers found that gaze was not a reliable indicator of computer-directed speech, primarily because users often looked at the computer while addressing their peers. Additionally, recent work conducted in our lab revealed that speakers do not direct significantly more command-like language to a computer than to a peer during multiparty interaction. In this study comparing spoken instructions matched on illocutionary force, participants issued 37.4% of their instructions as commands when addressing the computer, compared with 35.5% when addressing a human peer. As illustrated in Figure 2, participants' likelihood of using commands when addressing the computer was instead highly individual and strongly correlated with their use of commands when addressing their human peers. In fact, 81% of the variability in speakers' ratio of command-style instructions when addressing the computer is predictable simply by knowing their ratio of command-style instructions directed to humans. In effect, the social constraints of an interpersonal setting may dissuade speakers from abruptly changing to curter command-style language when addressing the computer.

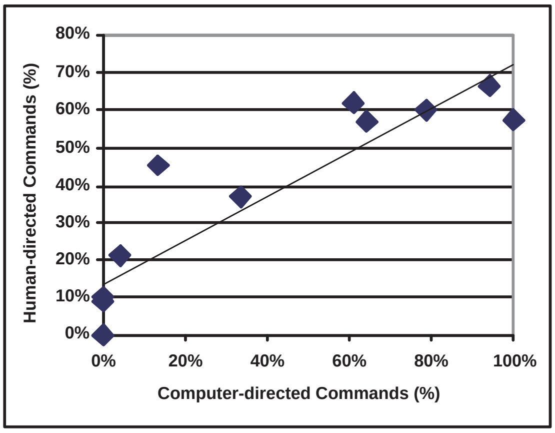

Figure 2: Linear regression of participants' ratio of commands when addressing computer versus human.

Additional analyses from the above noted study revealed that as a substitute for the gaze and speech style cues anticipated by previous researchers, speakers instead use a substantial amplitude change to differentiate human- versus computer-directed instructions, increasing their average amplitude by 2.4 dB when instructing the computer. When using a lexical marking such as "computer" to identify a computer-addressed instruction or a peer's name to identify a human-addressed instruction, the difference in speaker's computer versus human-directed amplitude was only 1.72 dB, implying that people use increased amplitude as an alternative strategy to lexical marking to clearly signal intended addressee [27]. These differences in amplitude were replicated for adjacent human-human and human-computer utterances pairs in which the amplitude of human- and computer-directed instructions was compared to the amplitude of immediately preceding conversational utterances. As shown in Figure 3, speakers increased their amplitude 4.63 dB when the subsequent instruction was addressed to a computer, but increased amplitude only 1.97 dB when addressing instructions to a human, a substantial 135% difference.

To effectively allay the cognitive impact associated with explicit microphone engagement, future implicit engagement systems will need to support, and potentially promote, those cues people spontaneously use to differentiate human- from computer-directed speech, such as the substantial amplitude shifts revealed in our empirical work.

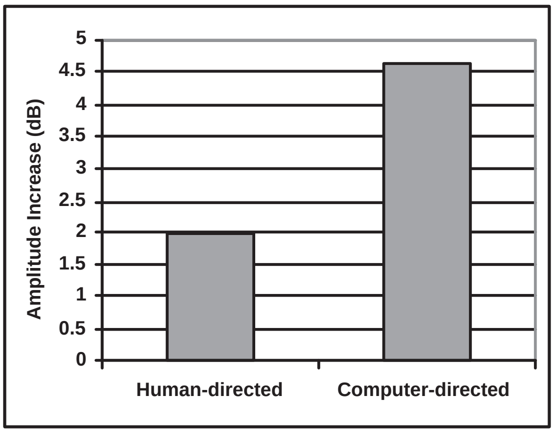

Figure 3: Average amplitude increase on adjacent utterance pairs when an instruction was addressed to a human versus computer.

4. LOW-DEMAND INTERFACES

Paradigms derived from single-user interfaces are not readily applicable to group situations. The former assumes that a user will be willing to work through an interface by activating elements structured e.g. within menus and toolbars. The main “dialog” takes place therefore between a user and a computer; the user is responsible for taking the initiative of providing detailed and minute guidance. In group situations, on the other hand, such a detailed handling of interface elements would be clearly unacceptable. The bulk of the users’ effort must by necessity be channeled into handling the primary human-human communication. The usually high cognitive load associated with the interaction itself leaves little extra room for interface manipulations, particularly if they are cumbersome.

These considerations led us to explore alternative paradigms in which the responsibility for obtaining input and presenting results is shifted from users to the system. The resulting systems - that we call *cumulative-observant* - operate mostly in the background and make their presence know to users very infrequently based on perceived opportunities [21]. Instead of supporting a sequence of user-command/system-display turns, a cumulative interface gathers continuous input across a structured interaction, like the multiparty construction of a Gantt schedule chart (Figure 4).

Various perceptive sensors are employed, some of which are ambient and unobtrusive (like an ink-capture whiteboard or a stereo-vision based gesture tracker) while others require user awareness (like close-talking microphones) [22]. These sensors collect information second-hand; that is, they observe and overhear what is naturally occurring between the people involved.

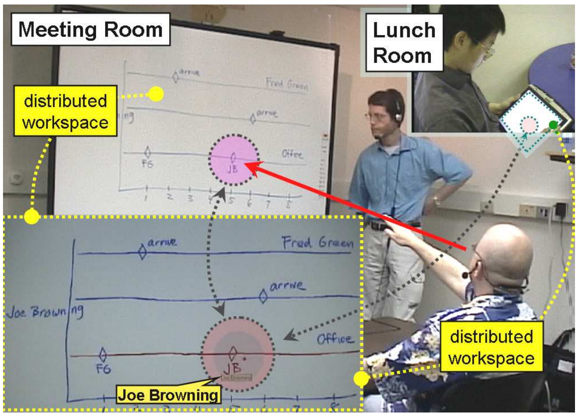

Figure 4: Multiparty distributed construction of a Gantt schedule chart. Dynamically learned abbreviation semantics (e.g. JB is an abbreviation of Joe Browning) along with distributed pointing recognition, for better informing a remote participant (upper right) of meeting information.

The participants do not explicitly interact with the computational interface. Instead they interact with each other. However, they are nonetheless aware that an ambient recognition system is processing their interaction and producing in the background artifacts the users would normally have to build manually after the interaction.

An observant interface also has available the evolving context of the structured interaction itself (e.g. a Gantt chart). This added context provides extra spatial and temporal constraints which help to improve recognition. In our Distributed Charter system [5], an interactive whiteboard collects pen ink and a stereo camera (centered above the whiteboard) tracks gestures, like Figure 4’s seated user’s deictic point at a milestone while speaking about it. The focus area (Figure 4, blue circle) generated by the 3D stereo body-tracking of the pointing gesture is annotated on the remote user’s shared view of the Gantt chart by a transparent hover label (e.g. Figure 4, lower left) which displays learned semantics, *Joe Browning*, associated with its abbreviation *JB*. Without this focus area and without the semantic hover label the remote user cannot easily resolve potentially ambiguous references made by remote participants.

Cumulative-observant systems such as the Distributed Charter system [5] briefly described above attempts to introduce minimal disruption while supporting users along two dimensions: 1) acting as proxies for remote participants and 2) keeping a structured record of the interaction for post-meeting consumption:

1. As a remote participant proxy, a system tracks events of interest that remote participants wouldn’t be able to access (e.g. the natural pointing gestures performed at a remote site), and pro-actively initiates a display mechanism to promote awareness.

2. In the latter case, the system may output searchable or browsable, audio-visual playback of the meeting, or in Charter’s case, provide a direct semantic interpretation of an artifact built during the interaction - the MS Project object representing a whiteboard Gantt schedule [22]. The dynamically learned semantic understandings are built up in the back-

ground by piggy-backing on the rich natural communication between the participants.

Cumulative-observant systems impose a very low overhead – users are required to work with/through instrumented devices such as microphones and an interactive white board. However, the main is that the interaction among participants remains mostly unchanged. For this to happen it is necessary for system manifestations to be carefully crafted, particularly in face of misrecognitions (see [21] for a description of techniques that address that).

5. SUMMARY

Support for collaboration introduces challenges related to the intrinsic complexity of interactions among humans, and to the vulnerability of the associated social processes to technology related changes. While collaborating, people employ, when given the opportunity, rich multimodal language over a variety of media. This allows them to communicate in an economical way, making better use of cognitive resources to handle task complexity.

Human-centered approaches are concerned with understanding and working within human capabilities and limitations, with the goal of supporting actual practices. The approach is therefore well matched to the challenges of providing support to groups of people collaborating. Understanding and exploiting human communicative capabilities and limitations is essential for obtaining multimodal and multimedia systems that work in practice.

We have discussed the importance of supporting multiple media, including tangible, paper-based artifacts, which are commonly found in actual practice. We described approaches and solutions for providing robust support to groups of people using a combination of media and modalities; we showed the impact of media choices on performance, particularly of more challenged participants.

We have also described the analysis of multimodal phenomena that can be leveraged to provide insights into intentions and meaning. We identified *multimodal redundancy* as a natural phenomenon that occurs in group situations; we described how systems can take advantage of it to provide robust cross-modal disambiguation and learning of out-of-vocabulary terms and of abbreviations. We also discussed multimodal features indicative of intended addressee (human or computer) during computer-assisted collaborative sessions, and identified *amplitude* as a robust predictor of addressee.

Finally, we discussed the need for specific interface paradigms that would fit the specificities of collaborative interactions, which typically revolve primarily around the human-human communication. We identified *cumulative-observant* as a class of systems that operate by "overhearing" the primary human-human communication, from which required information is autonomously extracted. Such systems attempt to reduce technology induced disruptions, allowing users to concentrate on their primary tasks rather than on explicitly driving an interface.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOINBC).

6. REFERENCES

[1] R. Anderson, C. Hoyer, C. Prince, J. Su, F. Videon, and S. Wolfman. Speech, ink and slides: The interaction of content channels. In *ACM Multimedia*, 2004.

[2] R. J. Anderson, R. Anderson, C. Hoyer, and S. A. Wolfman. A study of digital ink in lecture presentation. In *CHI 2004: The 2004 Conference on Human Factors in Computing Systems*, Vienna, Austria, 2004.

[3] Anoto Corporation. Anoto technology - how does it work? `http://www.anotofunctionality.com/cldoc/aof3.htm`, May 2006.

[4] A. Baddeley. Working memory. *Science*, 255:556–559, 1992.

[5] P. Barthelmess, E. Kaiser, X. Huang, and D. Demirdjian. Distributed pointing for multimodal collaboration over sketched diagrams. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, New York, NY, USA, October 2005. ACM Press.

[6] P. Barthelmess, E. Kaiser, X. Huang, D. McGee, and P. Cohen. Collaborative multimodal photo annotation over digital paper. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*. ACM Press, 2006.

[7] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. Quickset: Multimodal interaction for distributed applications. In *Proceedings of the Fifth ACM International Multimedia Conference*, 1997.

[8] P. R. Cohen and D. R. McGee. Tangible multimodal interfaces for safety-critical applications. *Communications of the ACM*, 47(1):41–46, 2004.

[9] P. R. Cohen, D. R. McGee, and J. Clow. The efficiency of multimodal interaction for a map-based task. In *Proceedings of the sixth conference on Applied natural language processing*, pages 331–338, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[10] P. Dietz and D. Leigh. Diamondtouch: a multi-user touch technology. In *UIST '01: Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 219–226, New York, NY, USA, 2001. ACM Press.

[11] J. Geissler. Shuffle, throw or take it! working efficiently with an interactive wall. In *CHI '98: CHI 98 conference summary on Human factors in computing systems*, pages 265–266, New York, NY, USA, 1998. ACM Press.

[12] D. Gergle, R. Kraut, and S. Fussel. Language efficiency and visual technology: Minimizing collaborative effort with visual information. *Journal of Language and Social Psychology*, 23(4):491–517, 2004.

[13] J. Grudin. Why cscw applications fail: problems in the design and evaluation of organization of organizational interfaces. In *CSCW '88: Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, pages 85–93, New York, NY, USA, 1988. ACM Press.

[14] J. Grudin. The computer reaches out: The historical continuity of interface design. In *Human Factors in Computer Systems (CHI)*, pages 261–268, April 1990.

[15] F. Guimbretiére, M. Stone, and T. Winograd. Fluid interaction with high-resolution wall-size displays. In *UIST '01: Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 21–30, New York, NY, USA, 2001. ACM Press.

[16] A. Gupta and T. Anastasakos. Integration patterns during multimodal interaction. In *International Conference on Spoken Language Processing - INTERSPEECH*, pages 2293–2296, 2004.

[17] R. R. Hoffman, A. Roesler, and B. M. Moon. What is design in the context of human-centered computing? *IEEE Intelligent Systems*, 19(4):89–95, 2004.

[18] A. Jaimes. Human-centered multimedia: Culture, deployment, and access. *IEEE MultiMedia*, 13(1):12–19, 2006.

[19] E. Kaiser. Shacer: a speech and handwriting recognizer. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, New York, NY, USA, October 2005. ACM Press. Workshop on Multimodal, Multiparty Meeting Processing.

[20] E. Kaiser. Using redundant speech and handwriting for learning new vocabulary and understanding abbreviations. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*. ACM Press, 2006.

[21] E. Kaiser and P. Barthelmess. Edge-splitting in a cumulative multimodal system, for a no-wait temporal threshold on information fusion combined with an under-specified display. In *Proceedings Interspeech 2006 - ICSLP)*, 2006.

[22] E. Kaiser, D. Demirdjian, A. Gruenstein, X. Li, J. Niekrasz, M. Wesson, and S. Kumar. Demo: A multimodal learning interface for sketch, speak and point creation of a schedule chart. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 329–330, New York, NY, USA, October 2004. ACM Press.

[23] E. C. Kaiser. Dynamic new vocabulary enrollment through handwriting and speech in a multimodal scheduling application. In *Making Pen-Based Interaction Intelligent and Natural, Papers from the 2004 AAAI Symposium*, pages 85–91, Arlington, VA., USA,, 2004. Technical Report FS-04-06.

[24] E. C. Kaiser. Multimodal new vocabulary recognition through speech and handwriting in a whiteboard scheduling application. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 51–58, 2005.

[25] M. Katzenmaier, R. Stiefelhagen, and T. Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 144–151, New York, NY, USA, October 2004. ACM Press.

[26] R. Lunsford, E. Kaiser, P. Barthelmess, and X. Huang. Managing extrinsic costs via multimodal natural interaction systems. In *CHI'06 Workshop: What is the Next Generation of Human-Computer Interaction?*, 2006. `www.eecs.tufts.edu/\~jacob/workshop/papers/lunsford.pdf`.

[27] R. Lunsford and S. Oviatt. Toward open-microphone engagement for multiparty field interactions. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, New York, NY, USA, October 2006. ACM Press.

[28] D. McGee and P. Cohen. Creating tangible interfaces by augmenting physical objects with multimodal language. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI 2001)*, 2001.

[29] D. R. McGee, P. R. Cohen, R. M. Wesson, and S. Horman. Comparing paper and tangible, multimodal tools. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 407–414, New York, NY, USA, 2002. ACM Press.

[30] W. J. Orlikowski. Learning from notes: organizational issues in groupware implementation. In *CSCW '92: Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pages 362–369, New York, NY, USA, 1992. ACM Press.

[31] S. Oviatt. Multimodal interfaces. In J. Jacko and A. Sears, editors, *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, chapter 14, pages 286–304. Lawrence Erlbaum Assoc., Mahwah, NJ, 2003.

[32] S. Oviatt, A. Arthur, and J. Cohen. Quiet interfaces that help students think. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology (UIST'06)*, 2006.

[33] S. Oviatt, R. Coulston, and R. Lunsford. When do we interact multimodally?: cognitive load and multimodal communication patterns. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 129–136, New York, NY, USA, October 2004. ACM Press.

[34] S. Oviatt, A. DeAngeli, and K. Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of Conference on Human Factors in Computing Systems: CHI '97*, New York:, 1997. ACM Press.

[35] S. Oviatt and E. Olsen. Integration themes in multimodal human-computer interaction. In *International ConferenceonSpoken Language Processing (ICSLP '94)*, pages 551–554, 1994.

[36] A. Paivio. *Mental representations: A dual coding approach*, volume 9 of *Oxford Physchology Series*. Oxford University Press, 1990.

[37] J. Rekimoto and N.Matshushita. Toward a human and object sensitive interactive display. In *Perceptual User Interfaces (PUI)*, 1997.

[38] A. J. Sellen and R. H. Harper. *The Myth of the Paperless Office*. MIT Press, Cambridge, MA, USA, 2003.

[39] M. G. Shafto and R. R. Hoffman. Guest editors' introduction: Human-centered computing at nasa. *IEEE Intelligent Systems*, 17(5):10–13, 2002.

[40] J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science: A Multidisciplinary Journal*, 12(2):257–285, 1988.

[41] N. Talbert. Toward human-centered systems. *IEEE Comput. Graph. Appl.*, 17(4):21–28, 1997.

[42] J. C. Tang. Findings from observational studies of collaborative work. *Int. J. Man-Mach. Stud.*, 34(2):143–160, 1991.

[43] K. van Turnhout, J. Terken, I. Bakx, and B. Eggen. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 175–182, New York, NY, USA, October 2005. ACM Press.