Managing Extrinsic Costs via Multimodal Natural Interaction Systems

Rebecca Lunsford, Ed Kaiser, Paulo Barthelmess, Xiao Huang
Natural Interaction Systems
10260 Greenburg Road
Suite 400
Portland, OR 97223 USA
rebecca.lunsford; ed.kaiser; paulo.barthelmess; xiao.huang; @naturalinteraction.com

ABSTRACT
Modern day interactions, whether between remote humans or humans and computers, involve extrinsic costs to the participants. Extrinsic costs are activities that, although unrelated to a person's primary task, must be accomplished to complete the primary task. In this paper we provide a framework for discussing certain extrinsic costs by describing those we term over-specification, repetition, and interruption. Natural interaction systems seek to reduce or eliminate these costs by leveraging peoples' innate communication abilities. However, in conceiving these interfaces, it's critical to acknowledge that humans are naturally *multimodal* communicators, using speech, gesture, body position, gaze, writing, etc., to share information and intent. By recognizing and taking advantage of humans' innate ability to communicate multimodally, extrinsic interaction costs can be reduced or eliminated. In this paper we review previous and ongoing work that demonstrates how multimodal interfaces can reduce extrinsic costs.

Author Keywords
Natural interaction system, multimodal interfaces, mutual disambiguation, extrinsic interaction costs

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
In all but the most straightforward interaction between two co-located individuals, people pay extrinsic costs to support their interactions. Whether giving directions over the phone, transcribing meeting notes, or making plane reservations using a speech enabled application, people must engage in activities that are unrelated to the primary task, in order to accomplish that task. In many situations users accept these extrinsic costs as a minor inconvenience. However, in some cases users feel the costs outweigh the benefit received and refuse to use a system [4].

In our view, many of these extrinsic costs are a direct result of non-existent or beleaguered communication channels that do not allow people to communicate using the broad range of modes typical of co-located human-human communication. As a framework for discussing these extrinsic costs and how multimodal interfaces can minimize or reduce these costs, we focus on *over-specification, repetition,* and *interruption*.

Over-specification
Over-specification occurs when one needs to translate information that can be conveyed easily via one or more communication modalities into other, less natural modes. For example, when giving directions on the telephone, one must translate spatial and directional information into complex verbal descriptions such as "take the 295 exit for Taylor's Ferry, and get in the center lane because you're going to have to take a wide left then an immediate right to actually get on Taylor's Ferry". Alternatively, if the parties were co-located and sharing a map, the same instructions might be as simple as "when you exit you need to get in the center lane because you want to end up here", accompanied by tracing the route with a finger. We term this extrinsic cost "over-specification" in that the user must engage in a more cognitively demanding, less natural communication due to the constraints of the communication channel.

Repetition
Repetition is the process of transcribing information into another format, typically digital. Consider the following example. A project manager and team meet to discuss an ongoing software development project and determine that research is needed on software licensing. The project manager then writes "software licensing" on a flipchart, turns to one of the team members, and says "can you handle

CHI 2006, April 22–28, 2006, Montréal, Québec, Canada.

that?" to which the team member responds with a nod while writing a star next to his meeting notes — his way of identifying the task as his. On returning to their offices the team member adds the action item to his digitized to-do list and the project manager adds the action item to the electronic project plan. Although the action item was discussed verbally, labeled on the flip chart and assigned via conversation and note taking, it was also typed into a to-do item and typed into the project plan. We term this extrinsic cost "repetition" in that the same piece of information is processed multiple times.

Interruption

Interruption is defined as "to stop or hinder by breaking in" [13]. For the modern-day information worker, interruption can be viewed as a side-effect of availability. That is, by having open communication channels such as phone, email, IM, and office door, the user is available to be interrupted by any of these devices. Researchers working on minimizing the impact of interruption have already recognized that multimodal sensors will be needed to identify when a person is most interruptible [6]. However, this work doesn't address interruption within an interaction by the interaction partner misinterpreting the users' behavior.

One example of this type of interruption are automated features (such as in Microsoft's XP operating system) which alter the interface state based on a user's actions, for example spontaneously adding a clipboard pane if the user selects a second item to copy without pasting the first selection. However, in doing this, the application has not only "stopped or hindered" but has effectively forced an added task onto an already busy user - that of figuring out how to return the interface to a state where they can comfortably continue with their task. For the purposes of this paper, we term this extrinsic cost "interruption", in that the user's primary task must be postponed while the secondary task is concluded.

Background - Multimodality in Evolving Interfaces

We believe that as computational systems become more and more perceptually capable of recognition at all levels — (1) symbolic/syntactic input levels (e.g. speech, handwriting and gesture recognition), (2) semantic understanding of both single and integrated multiple input modes (e.g. parsing and semantic modeling), and (3) pragmatic contextualization and tracking of interaction structure — they will benefit from multimodal integration. We believe that reality-based interfaces will be fundamentally multimodal interfaces, or what we call natural interaction systems.

Multimodal systems support and leverage the fact that people communicate in various modes. On one hand multimodal inputs carry complementary information. For example, people are adept at using fine information in one input mode to disambiguate coarse information in another input mode. In a classroom full of desk/chair groups pointing to a distant pair and saying, "Move that chair," is an effectively understood command. However, saying only, "Move that chair," without pointing, nor pointing at the distant group while saying only, "Move that," are not effectively understood commands — they both require further specification. This mutual disambiguation, using information from one uncertain input mode to disambiguate complementary information in another uncertain input mode, is a powerful combinatory tool for dealing with the inherent ambiguity in individual recognizer interpretations. Ranked lists of semantic interpretations can be combined and filtered during mutual disambiguation. Choosing the best scoring combinations using mutual disambiguation reduces semantic error rates by between 40%-67% [14] [8].

On the other hand, aside from carrying complementary information, multiple input modes at times carry semantically redundant information. In human-human, computer mediated interactions like distance lecture delivery or whiteboard-centered business meetings the communications (particularly those of a "presenter") are often semantically redundant. One recent study of tablet-PC delivered distance-learning lectures found that 100% of randomly selected handwriting instances were accompanied by semantically redundant speech. Our own analysis of a spontaneous white board planning session (recorded from a quarterly project planning meeting of 15-20 participants) also found that whiteboard and flip-chart handwriting were accompanied by semantically redundant speech 98% of the time.

Thus we argue that evolving ubiquitous, tangible and perceptive interfaces need to leverage people's natural use of multimodal communication. Multimodality is what users expect in natural human-human interactions. System integration of multimodal inputs using mutual disambiguation has been shown time and again to increase overall system recognition levels. In the area of human-computer interactions the primary benefit of mutual disambiguation is due to complementarity of inputs. Whereas in human-human, computer-mediated interactions multimodal semantic redundancy is common, and leveraging its occurrence can be the basis for perceptual interfaces that learn better how to observe, recognize and understand multiple input streams as they are used.

ELIMINATING OVER-SPECIFICATION

An important source of extrinsic cost can be associated with the way technological constraints force users to adapt their communicative behavior and consequently their work practices. Conventional interfaces require users to express their intentions by means of textual input and widget selections; that is at the same time both very limited and too different from usual communication modes people use.

As discussed above, people communicate through multiple channels simultaneously, in such a way that these channels of information complement or reinforce each other,

reducing ambiguity. In contrast, conventional interfaces impose a strictly sequential operation performed by manipulating physical devices such as keyboards and mice.

While conventional interface paradigms may be appropriate for a potentially large class of domains, they are in general less acceptable when users are dealing with more demanding domains e.g. when visual spatial components are involved, or in multiparty situations. In these situations, interface limitations lead to use of circuitous language (e.g. [2, 10]), as users attempt to compensate for technological limitations by exploiting in unnatural ways remaining available channels. In addition, there is commonly a disruption of the natural flow of work, resulting from the need to perform actions whose sole purpose is to steer the technology, e.g. by requiring users to explicitly control turn-taking and pointer control through a series of widget selections while on distributed meeting.

In contrast to conventional interfaces, systems that are able to analyze multiple modalities, as produced in the course of work performance, support unchanged work and communicative practices while introducing minimal additional system-related costs. In the next paragraphs we highlight some of the previous work of the group that illustrates this quality.

Work within our group has explored the use of speech and pen input as means to support a natural way for users to perform a variety of map-based tasks. NISMap [4], derived from the Quickset system [3] allows for users to annotate maps with military symbols by sketching conventional symbols and speaking. Earlier versions of NISMap required the use of interactive surfaces such as touch sensitive screens; more recently, NIS Map has been extended to support drawing over regular paper maps [4]. This transition to support conventional tangible materials makes NIS Map use virtually indistinguishable from the non-automated, paper-based operation users are already accustomed to, significantly lowering extrinsic costs associated with its use.

More recently, we explored multimodal perception and integration to support remote collaboration. The Distributed Charter system [1] transparently handles the propagation of project schedule sketches produced by small teams working at distributed sites. In this system, the multimodal contributions made by distributed participants, as they sketch on instrumented boards or tablet computers and talk about a schedule, are integrated by the system into a coherent semantic interpretation of the interaction. The system thus keeps the representation of the schedule consistent across sites, independently of device and resolution used. In addition, a stereo-vision tracker and recognizer is used to perceive deictic gestures made towards the interactive boards at each site and transparently propagates the gestures, making participants aware of naturally occurring pointing gestures made by users at remote sites. Once more, the goal of this system is to take advantage of naturally occurring group behavior to drive system mediated awareness services without requiring explicit user actions to drive the system.

REDUCING REPETITION

By leveraging existing work practices (allowing users to draw units on post-it notes and place them on the map) our group's earlier Rasa system reduced the repetition necessary to digitally record a complex battle-scene or planning map [12]. Instead of hand copying the large map, or taking pictures and typing in information to digitize it, the underlying digitizing technology simultaneously and unobtrusively recorded and recognized the inputs. Since perceptual technologies have improved, even the modest intrusiveness of these recording methods is no longer necessary. Reliable digitizing ink technology (e.g. Anoto[1]) is based on a special pen containing a camera, which operates on paper printed with a (almost invisible) dot pattern. With the dot patterns, the camera in the pen determines the relative location of the ink coordinates with respect to the paper. These coordinates can then be transmitted to a computer for storage and further processing. Rasa's digitizing tablets can be and have been replaced by digital paper and pens [4]. The map itself is now printed on digital paper making the touch sensitive whiteboard and the post-it notes optional.

Using digital paper and speech recording devices can have wide applicability. For example, repetition costs in human-computer interaction are prevalent in note taking and form-filling. To minimize these costs, we are developing a new collaborative note-taking facility called NISNotes. Regular handwritten notes taken on digital paper can be directly transferred in real time to current digital note taking systems like Microsoft OneNote. This digital ink can be combined with speech to produce more accurate recognition transcripts. That means less copying of notes, less correction of transcripts, etc. By incorporating our MOOVR/SHACER architecture (Multimodal Out-Of-Vocabulary Recognition using a Speech and HAndwriting reCognizER) [7] we intend to evolve versions of NISNotes that learns and improves over time as their vocabulary and understanding increase.

We are already able to show the advantages of systems that learn and improve over time in Charter, our automated interface to MS Project scheduling. By combining sketch, speech, gesture and task-contextual information Charter supports the dynamic learning of new vocabulary like abbreviation semantics (e.g. that "JB" stands for "Joe Browning"). This learning also occurs in the context of existing work practices — the creation of a whiteboard chart in a collaborative meeting. Eventually when Charter is combined with NISNotes, we believe that the collaborative creation and manipulation of complex project charts will

[1] http://www.anoto.com

become even better as more sources of information all contribute to mutually disambiguate each other. Repetition can be eliminated because to-do lists and project plans need not be re-typed later. Also, having a digital memory of meetings facilitates easier review and sharing of summaries and project planning views as well as supporting more informative distributed interfaces as described above in the section on Eliminating Over-Specification.

Another example of reduced extrinsic costs is related to Physicians who are accustomed to using paper forms. With only a paper record, opportunities are missed for improving both individual care and institutional efficiency. To avoid this issue in current practice, clinicians are often expected to engage in repetitive data entry [5].

Using the Anoto technology, we have built NISChart [4], a digital paper-based charting application where handwriting and speaking are the primary input modalities. A physician can enter values, text, check marks, and so on into the hospital's standard forms, printed on Anoto paper. After finishing writing, the physician can put the digital pen into a cradle, from where the digital ink is sent to the computer. The application recognizes the texts and symbols written on the form. Both the digital ink and transcripts are saved in a relational database. Thus, we have the physical paper with real ink as the primary record. Moreover, the ink is digitized, analyzed, recognized and saved as reusable records without retyping. In other words, by eliminating repetition costs, now data entry is "as easy as writing" with a multimodal natural interaction system.

REDUCING INTERRUPTION

Audio-visual open-microphone speech-based interfaces often assume the user is addressing the system if he or she is speaking while facing the system. However, in recent work it's been shown that users often look at a system while addressing a peer [9], or while engaging in self-directed speech relevant to completing the task at hand [11]. In attempting to respond to these non-system directed utterances, systems implement a state-change that the user must then reverse.

In recent empirical work, we found that users engaging in self-directed speech while interacting with a speech-enabled application will consistently use significantly lower amplitude as compared to speech addressed to the system [11]. By leveraging this additional mode already available in the speech channel, future applications will be able to reduce interruption by attempting interaction only when the user is, in fact, addressing the system. Ongoing work will explore whether amplitude, in concert with other natural behaviors, can differentiate self-, system, and other human directed speech in a multiparty environment leading to more natural human-human-computer interactions.

CONCLUSION

In describing the extrinsic costs of over-specification, repetition and interruption inherent in modern-day interactions, we provided an initial framework within which to discuss and evaluate those costs. In addition, by detailing ways in which multimodal applications can and do reduce these costs, we demonstrate ample evidence to support our belief that the goal of natural interaction systems should be to reduce the extrinsic costs inherent in modern-day interactions by recognizing and utilizing natural human interaction modes and supporting effective current working practices.

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).

REFERENCES

[1] Barthelmess, P., E. Kaiser, X. Huang, & D. Demirdjian. Distributed pointing for multimodal collaboration over sketched diagrams. ICMI'05: ACM Press: 10-17.

[2] Bly, S.A. A use of drawing surfaces in different collaborative settings. CSCW 1988: ACM Press: 250-256.

[3] Cohen, P.R., M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, & J. Clow. QuickSet: multimodal interaction for distributed applications. ACM Multimedia 1997: ACM Press: 31-40.

[4] Cohen, P.R. & D.R. McGee, Tangible multimodal interfaces for safety-critical applications. Commun. ACM, 2004. 47(1): 41-46.

[5] The Computer-based Patient Record: An Essential Technology for Health Care. 2 ed. Institute of Medicine. 1997, Washington. D.C.: National Academy Press.

[6] Horvitz, E., P. Koch, & J. Apacible. BusyBody: creating and fielding personalized models of the cost of interruption. CSCW 2004: ACM Press: 507-510.

[7] Kaiser, E. SHACER: a Speech and Handwriting Recognizer, Workshop on Multimodal, Multiparty Meeting Processing. ICMI'05

[8] Kaiser, E. & A. Olwal. Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality. ICMI'03

[9] Katzenmaier, M., R. Stiefelhagen, & T. Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. ICMI'04: ACM Press: 144-151.

[10] Luff, P., C. Heath, H. Kuzuoka, J. Hindmarch, K. Yamakazi, & S. Oyama, Fractured Ecologies: Creating Environments for Collaboration. Human-Computer Interaction, 2003. 18(1&2): 51-84.

[11] Lunsford, R., S. Oviatt, & R. Coulston. Audio-visual cues distinguishing self- from system-directed speech in younger and older adults. ICMI'05: ACM Press: 167-174.

[12] McGee, D.R., P.R. Cohen, R.M. Wesson, & S. Horman. Comparing paper and tangible, multimodal tools. CHI 2002: ACM Press: 407-414.

[13] Merriam Webster Online. http://www.m-w.com/

[14] Oviatt, S. Mutual Disambiguation of Recognition Errors in a Multimodal Architecture. CHI '99