When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns

Sharon Oviatt Rachel Coulston Rebecca Lunsford
Center for Human Computer Communication
Department of Computer Science
Oregon Health & Science University
20000 NW Walker Road
Beaverton, OR 97006, USA
+1 503 748 1342
oviatt, rachel, rebeccal@cse.ogi.edu

ABSTRACT

Mobile usage patterns often entail high and fluctuating levels of difficulty as well as dual tasking. One major theme explored in this research is whether a flexible multimodal interface supports users in managing cognitive load. Findings from this study reveal that multimodal interface users spontaneously respond to dynamic changes in their own cognitive load by shifting to multimodal communication as load increases with task difficulty and communicative complexity. Given a flexible multimodal interface, users' ratio of multimodal (versus unimodal) interaction increased substantially from 18.6% when referring to established dialogue context to 77.1% when required to establish a new context, a +315% relative increase. Likewise, the ratio of users' multimodal interaction increased significantly as the tasks became more difficult, from 59.2% during low difficulty tasks, to 65.5% at moderate difficulty, 68.2% at high and 75.0% at very high difficulty, an overall relative increase of +27%. Analysis of users' task-critical errors and response latencies across task difficulty levels increased systematically and significantly as well, corroborating the manipulation of cognitive processing load. The adaptations seen in this study reflect users' efforts to self-manage limitations on working memory when task complexity increases. This is accomplished by distributing communicative information across multiple modalities, which is compatible with a cognitive load theory of multimodal interaction. The long-term goal of this research is the development of an empirical foundation for proactively guiding flexible and adaptive multimodal system design.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *user-centered design, theory and methods, interaction styles, input devices and strategies, evaluation/methodology, voice I/O, natural language, prototyping.*

General Terms

Performance, Design, Reliability, Human Factors

ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.

Keywords

Multimodal interaction, speech and pen input, unimodal interaction, dialogue context, task difficulty, cognitive load, human performance, system adaptation, multimodal integration, individual differences.

1. INTRODUCTION

Multimodal interfaces are recognized to be inherently flexible, and to provide an especially ideal interface for accommodating both the changing demands encountered during mobile use and also the large individual differences present in the population – a clear requirement for universal access [9, 18, 21, 25]. These interfaces can be designed to support simultaneous use of input modes, to permit switching among modes to take advantage of the modality best suited for a task, environment, or user capabilities, or to "translate" information from one mode to another in order to expand accessibility for users with selective limitations. Compared with traditional keyboard and mouse interfaces, multimodal interfaces are better able to support rich expressiveness using familiar communication modalities, and they can be particularly well suited for mobile use – as in the case of speech and pen multimodal interfaces now being developed for small handhelds. Furthermore, multimodal interfaces that fuse speech and pen input have been demonstrated to substantially reduce speech recognition errors, and also to stabilize system reliability during mobile use in noisy field environments [18]. Since it is important that any mobile interface serving field tasks be flexible and able to minimize demands on users' attention, one major theme explored in the present paper is whether a flexible multimodal interface may be well suited for assisting users in *self-managing their cognitive load* and improving overall performance as the complexity of field tasks and related communications increase.

Educational Literature on Cognitive Load

When learning new intellectual tasks, the cognitive effort required on the part of learners can fluctuate dramatically and occasionally exceed a user's ability. Over the past decade, *cognitive load theory* has maintained that in the process of learning and developing expertise it is easier to acquire new schemas and automate them if instructional methods can minimize demands on a user's working memory, thereby reducing their cognitive load [5, 12, 22, 24, 27]. Advocates of this theory typically assess the "extraneous complexity" associated with instructional methods or related interface design separately from the "intrinsic complexity" related to a student's main learning task, which is done by

comparing performance indices of cognitive load as students use different methods during their tasks. In a series of related education experiments, it was revealed that a dual-mode presentation format involving diagrams and audiotapes supported expansion of working memory and problem solution in geometry tasks better than a single visual mode [12]. Furthermore, greater performance advantages for this effect have been demonstrated for more difficult instructional materials, compared with simpler ones [27]. Essentially, it was shown that an integrated multimodal presentation format can expand the capacity of working memory in a manner that expedites classroom instruction. This basic finding on the advantages of multimodal presentation format for students' tutorial performance has been replicated for different tasks, dependent measures, and presentation materials, including computer-based multimedia animations [11, 27].

Cognitive and Linguistic Theory Related to Cognitive Load

There currently is no coherent theoretical framework that accounts for multimodal interaction patterns, which would be invaluable for proactively guiding the design of future multimodal interfaces to be optimally compatible with human capabilities and limitations. However, empirical results have been accumulating rapidly on many aspects of *multisensory perception* [4], especially in areas such as audio-visual processing. Compared with unimodal perception, advantages in perceptual discrimination have been documented for audio-visual multimodal stimuli in recent experiments [4, 26]. In addition, information presented via audio-visual means has been demonstrated to yield an intelligibility advantage during spoken communication [6], as well as in the type of learning, retention, and transfer of learning tasks mentioned in the last section. From the viewpoint of advancing multimodal interface design, however, there has been a gap in our knowledge of how this multisensory perceptual research may relate to users' *multimodal production* during human-computer interaction.

In addition, relevant empirical work and a *cognitive resource theory* have been developed by Wickens and colleagues [28], the theme of which is resource competition between modalities during tasks. This theory essentially states that there can be competition between modalities like audition and vision during tasks, such that the human attention and processing required during both input and output result in better performance if information is distributed across modalities [28]. In related theoretical work, Baddeley has presented a *theory of working memory* which maintains that short-term or working memory consists of multiple independent processors associated with different modes [2]. According to this theory, a visual-spatial 'sketch pad' maintains visual materials such as pictures and diagrams in one area of working memory, while a separate phonological loop stores auditory-verbal information. Although these two processors are believed to be coordinated by a central executive, in terms of lower-level modality processing they are viewed as functioning largely independently, which is what enables the effective size of working memory to expand when people use multiple modalities during tasks. Both Wickens' and Baddeley's theories have strongly influenced the educational research summarized earlier on cognitive load, and they likewise provide useful perspectives for guiding future interface design research.

Within linguistics, there is substantial evidence that linguistic expressions referring to old versus new information are distinctly different categories, often referred to as *given and new information* in the literature [23]. It has long been clear that it is easier to refer to an entity already established in dialogue than to introduce a new entity, and that speakers avoid the superfluous effort of rearticulating an established referent [7]. However, it is only recently that theories have been developed to establish the relation between cognitive load and resolution of noun-phrase anaphora with and without discourse context [1]. In his *informational load hypothesis*, Almor [1] claims that noun phrase anaphoric processing is an optimizing process during which speakers trade off the cost of activating semantic information (e.g., when they identify an antecedent or add new information) against the cognitive load or verbal working memory limits incurred. According to this view, when speakers establish reference to a highly accessible entity (i.e., one in attentional focus), it only requires a low cost referring expression such as a pronoun. However, when they establish reference to a new or less accessible one, it requires a referring expression with a higher cost such as a full noun phrase. The informational load hypothesis asserts, like the Gricean maxim of quantity [7], that speakers will make a dialogue contribution only as informative as is minimally required. That is, they will adopt the least linguistically complex noun phrase needed to achieve their communicative goals, thereby attempting to manage any constraints imposed on their working memory [2]. One theme explored in the present research is whether users of a multimodal dialogue system also will choose to communicate in a manner that decreases their cognitive load during human-computer interaction. In particular, the present study examines whether users are more likely to communicate multimodally when establishing new dialogue context than when following up on an established one.

Interface Literature on Cognitive Load

As described earlier, one critical objective of all mobile interface design is the need to manage multitasking, interruption, attentional distraction, fluctuations in the difficulty of natural field tasks and situations, and resulting cognitive overload [8, 18]. This is essential because mobile users need to focus on complex primary field tasks that can vary substantially in difficulty and also involve dual tasking between the field task and secondary tasks involved in controlling an interface. These forcing functions for mobile interface design are further compounded when the user group (e.g., seniors) has working memory and other performance limitations that place them at greater risk for task failure or even physical accidents while using mobile systems [25].

Within speech interface research, several linguistic indices associated with task performance have been revealed to change as users' cognitive load increases. In particular, disfluencies are known to be a sensitive predictor of planning demands and cognitive load during human-computer interaction [14]. In previous work involving spatial tasks in which participants and their tasks were controlled, users' disfluency rate was shown to be higher on utterance constituents containing locative information compared with non-locative content [15]. The following utterance, which a user spoke to a real estate map application, illustrates this finding (locative constituent italicized): "Show me homes for sale under $200,000 *west, uh, no east of May Lake*." During mobile interactions with simulated systems, disfluencies, intersentential pausing, fragmented sentences, and slower speech

rate all have been found to increase when users are subjected to time pressure or navigational obstacles that increase their processing load [13]. Adaptive interface design is viewed as one direction for managing sources of increased cognitive load in future interfaces [10, 21].

Another major direction for managing users' cognitive load is multimodal interfaces. Within this area, considerable empirical work has been conducted specifically on human-computer interaction and performance during the use of speech and pen-based multimodal interfaces [19] and audio-visual speech and lip movement ones [3]. Past work indicates that people prefer to interact multimodally when given a choice, especially in spatial tasks [15]. They also can complete spatial map tasks with 50% fewer disfluencies, briefer and simpler linguistic constructions, and 36% fewer task-critical errors when given a multimodal interface in which they can select what content to speak versus write [15]. This kind of flexibility permits people to use the input mode they believe is most accurate and efficient for conveying particular content, which can improve their own and the system's accuracy [17]. Based on all of the literatures summarized, there are reasons to believe that a multimodal interface may be effective at minimizing users' cognitive load and supporting their performance, especially when confronted with complex tasks (e.g., spatial) or mobile dual-tasking challenges. To explore this possible relation, research is needed to investigate when users select to interact multimodally, and whether such a shift in their communication pattern is associated with task, dialogue, and situational complexities that co-vary with their performance level.

Goals of the Current Study

The general goals of the present research were to model and predict users' multimodal interaction patterns as a function of the task, dialogue, and user state. In particular, this study examines the impact of changes in users' cognitive load on their likelihood of interacting multimodally with a system, rather than unimodally. It explores whether users given a flexible multimodal interface increase their ratio of multimodal interaction under circumstances in which cognitive processing load becomes more elevated.

First, from a dialogue processing viewpoint, users' likelihood of producing a multimodal construction was compared in matched pairs of adjacent tasks in which they needed to either establish a new dialogue context or follow up on a previously established one. Given the increased cognitive demands inherent in establishing dialogue context, it was hypothesized that users would interact multimodally more frequently when setting a new context, but switch to unimodal interaction when referencing an established entity.

Secondly, people's ratio of multimodal versus unimodal constructions also was assessed during four task difficulty levels ranging from low to very high, with the prediction that multimodal interaction would increase across these difficulty levels. To provide corroboration that the four task difficulty levels and users' associated load were in fact increasing progressively, behavioral data also were evaluated on users' task-critical errors and response latencies while completing tasks at each difficulty level.

Finally, users' ratio of multimodal interaction was compared across the first versus second half of their session. In terms of task order, it was hypothesized that multimodal interaction would begin to decrease as users gradually became more familiar with a system and its tasks over the course of a session. Additional goals of this study involved examining individual differences in users' modality preferences and multimodal integration patterns, including whether they would adopt either a predominantly *simultaneous* or *sequential integration pattern* with high consistency as reported in past research [16, 20, 29]. Previous work has revealed systematic changes in users' temporal synchronization of modalities during multimodal interaction as they encounter system errors and more difficult tasks, which has been referred to as *multimodal hypertiming* [20]. However, past research has not investigated whether users' basic predilection to interact unimodally or multimodally is influenced by such factors. The long-term objective of this research is the development of an empirical foundation of information for proactively guiding the design of more flexible and adaptable multimodal interfaces, especially ones capable of optimal performance in natural usage contexts and while mobile.

2. METHODS

2.1 Participants

There were ten adult participants, between 19 and 50 years of age, five male and five female. All were native speakers of English and paid volunteers.

2.2 Application Scenario

Participants were asked to act as volunteer assistants, helping to coordinate emergency resources during a major flood in a municipal area. They were given a simulated map-based multimodal interface that displayed text prompts, which represented instructions from headquarters for them to accomplish. Participants could use speech input, pen input, or a combination of both to deliver instructions to the map system. Their tasks varied and could involve obtaining information (e.g., "Find out how many sandbags are at Couch School Warehouse"), placing items on the map (e.g., "Place a barge in the river southwest of OMSI"), creating or closing routes (e.g., "Make a jeep route to evacuate tourists from Ross Island Bridge"), and controlling the map display (e.g., "Scroll north on the map").

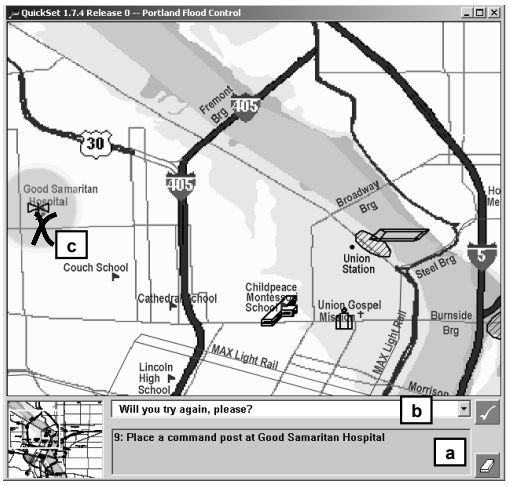

Figure 1. User interface

Figure 1 illustrates the interface. In this example, the message from headquarters was "Place a command post at Good Samaritan Hospital," shown in area (a). To complete this task, the participant said "We're gonna need a command post at Good Sam," while drawing an "X" at the desired location shown on the map (c). In this case, a simulated system error was received as feedback in area (b), "Will you try again, please?" Eventually, the participant would reenter their input to the system, which would result in a text confirmation instead being delivered to area (b), along with iconic feedback in the appropriate location on the map.

The tasks that users worked on represented four levels of difficulty: low, moderate, high and very high. Low difficulty tasks required the user to articulate just one piece of directional information (e.g., north, west) or one location (e.g., Cathedral School). Moderate difficulty tasks contained two such pieces of directional or location information, whereas high difficulty tasks contained three pieces, and very high difficulty tasks contained a total of four such pieces of information. Table 1 shows sample tasks from each of these task difficulty levels.

Table 1. Examples of task difficulty levels, with spatial-location lexical content in italics

Difficulty	Message from Headquarters
Low	Situate a volunteer area near ***Marquam Bridge***
Moderate	Send a barge from ***Morrison Bridge barge area*** to ***Burnside Bridge dock***
High	Draw a sandbag wall along ***east riverfront*** from ***OMSI*** to ***Morrison Bridge***
Very High	Place a maintenance shop near the ***intersection of I-405*** and ***Hwy 3***0 just ***east*** of ***Good Samaritan***

2.3 Procedure

During the training session, each participant was given 15 tasks to familiarize her with the system's coverage, capabilities, and input alternatives. An experimenter was present to give instructions, answer questions, and offer feedback and help. The training tasks were divided into three groups of five. During the first five tasks, the participant was told to express herself naturally using speech input with an open-microphone implementation. For the next five tasks, she was given an electronic stylus and instructed to only use pen input. During the final five training tasks, she was invited to use either speech input, pen input, or both modes in any way she wished to communicate information to the map system. Upon finishing the orientation and training, participants were told, "For the rest of your session, you are free to use either or both input modes in any way you like." Following training, the experimenter left the room and each participant completed their session independently, which involved 93 tasks. Upon completion, participants were debriefed on the purpose of the study. Until that point, all participants believed they were interacting with a fully-functional computer system. The entire experiment lasted about an hour per participant.

2.4 Simulation Technique

The data collection process used in this study was a *dual-wizard* high-fidelity semi-automatic simulation technique similar to that described in previous work [20]. Users' input was logged in real time during data collection on a command-by-command basis by the Input Wizard who, in this study, labeled each as having been delivered: (1) unimodally with either speech or pen, or (2) multimodally and either simultaneously or sequentially integrated. Basically, the Input Wizard's function was to record the users' input modality and pattern in real time throughout the session by observing a video feed of the session. This information then was routed to both a data log and to the Output Wizard's system. In a nearby room, the Output Wizard monitored the content of the user's input and responded with appropriate feedback sent directly to the user's display. The Output Wizard used an optimized interface with preloaded information to expedite speed of responding. In addition, although the system required input from both wizards, rapid responding also was expedited by the fact that no explicit coordination was necessary between the wizards using this dual-wizard technique.

In this simulation study, the random error generator delivered a 20% rate of errors distributed across the 93 tasks. When triggered, this mechanism occasionally overrode the system's response (transmitted by the Output Wizard), and instead responded with a failure-to-understand system message, such as "Will you try again, please?" Any error messages were delivered in the system feedback area (Figure 1, area b), which was highlighted in red.

2.5 Research Design and Data Capture

The experimental design involved analysis of within-subject data from each of the ten users while engaged in their 93 tasks. The main independent variables included: (1) Task difficulty level (Low, moderate, high, very high), with 12 exemplars of each of the four difficulty levels distributed throughout the session to control for order, (2) Dialogue context (No context, context established), with 7 matched pairs of immediately adjacent tasks distributed approximately evenly between the initial and final half of the session to control for order effects. During the first task of each pair, the user established a new dialogue context (e.g., "*Place a headquarters at OHSU*"), whereas the second utterance of the pair was delivered as a follow-up (e.g., "*Find out what resources are available at OHSU headquarters*"), and (3) Task Order (Initial, final), which involved a median split of the session.

2.6 Dependent Measures and Coding

Information about the type of input mode and modality integration pattern, as well as performance, was coded using SVHS video editing equipment. Users' utterances were transcribed verbatim, and data were analyzed using a suite of tools designed and developed in-house for the purpose. These tools enable coders to both annotate and summarize user performance. Coders could view precise selections of the audio-visual record of the participant and interface during the session, take frame-accurate (0.03 second) measurements and link the transcript to the video record of the session. After labeling data according to the parameters described below, they could selectively filter and view subsets of the coded data. The following provides a description of the dependent measures and their scoring.

2.6.1 Modality and Integration Pattern

For each task, the utterance issued by the participant was coded as either unimodally or multimodally delivered. If unimodal, input

was scored as either involving speech input or pen input. When multimodal, the integration pattern was coded as either a simultaneous one (i.e., speech and pen input at least partially overlapped in time), or a sequential one (i.e., one input mode delivered before the other, with a lag between modes and no temporal overlap). For individual conditions, the percentage of multimodal constructions was summarized for each participant and condition.

2.6.2 Performance Measures

The following performance measures were assessed on a subset of 48 tasks that were matched on task difficulty level (12 low, 12 moderate, 12 high and 12 very high difficulty). Human performance measures were analyzed to determine whether task difficulty had an impact on task-critical performance errors and on response latency to plan and execute a task.

Human Performance Errors

On each task, the total number of task-critical human performance errors (HPEs) was scored during user input to the system whenever the participant specified an incorrect location, direction, or name for a location, or if the task content was completely in error. For example, given the task "Close Morrison Bridge at the East end," a participant might mistakenly speak the incorrect bridge name, as in "Close *Marquam* Bridge East." She might also speak correctly but make a wrong gesture, as in the task "Put a sandbag wall along east riverfront from Morrison Bridge to Steel Bridge," when a participant drew her sandbag wall on the *west* bank of the river, this time saying just "Sandbag wall." HPEs were coded in part to validate that task difficulty levels as experienced by participants in this study were in fact increasing in difficulty systematically between the low and very high levels as participants were required to keep track of additional spatial information. All tasks for each participant were coded for the total number of such errors, which then was converted to a ratio of errors out of the total tasks completed at each difficulty level.

Response Latency to Initiate Task

Response latencies were measured as the total time interval between receiving a task instruction in area (a) of the interface (Figure 1) and when the participant first initiated speech or pen input to the system to complete the same task.

2.6.3 Reliability

Each task was independently scored by a second scorer, who carefully double-checked all of the real-time unimodal and multimodal judgments and multimodal integration patterns by hand from video records in order to verify their accuracy. Second scoring on 20% of the performance data also indicated that 86% of scored HPEs matched between raters. Response latency measures were coded using the same procedure and coders as in other recent studies, for which 80% of all durations have matched to within 0.1 seconds.

3. RESULTS

Data on approximately 1,100 constructions were available to be scored in total, although comparisons between specific conditions often were based on a smaller subset of matched pairs. Users' input involved relatively conversational rather than command-style language ("Okay, close the highway west of Burnside Bridge"). On average, the utterance length for users' spoken or multimodal constructions in this domain ranged between 1.9 and 14.4 words, with a mean of 6.8 words.

3.1 Modality and Integration Pattern

As shown in Figure 2, the system involved 25.6% speech-only input and 12.6% pen-only input, although 61.8% or the majority of all input was delivered multimodally. Of these multimodal constructions, 85.6% were delivered simultaneously and 14.4% sequentially.

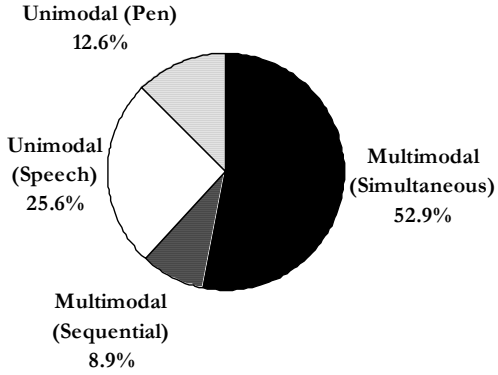

Figure 2. Percentage of unimodal (pen only and speech only) and multimodal (sequentially integrated and simultaneously integrated) constructions across entire corpus

With respect to individual differences, most participants showed a strong preference to use one modality or the other when interacting unimodally, with six out of ten delivering *all of their unimodal constructions* in the favored modality. For five of these six participants, their favored modality was speech, whereas one participant consistently favored pen.

Table 2. Multimodal integration patterns for individuals

Participant	SIM	SEQ
1	**87%**	13%
2	**100%**	0%
3	**90%**	10%
4	**97%**	3%
5	**99%**	1%
6	**98%**	2%
7	5%	**95%**
8	**72%**	28%
9	**97%**	3%
10	0%	**100%**
Overall Average Consistency		**93.5%**

In terms of individual differences in integration pattern, Table 2 illustrates that two out of ten participants were predominantly sequential integrators (i.e., 60% or more of their integration patterns involved a lag between signals), whereas the other eight participants were predominantly simultaneous integrators. As also shown in Table 2, whatever an individual participant's dominant multimodal integration pattern was, it was delivered 93.5% consistently on average, with all integration patterns ranging between 72-100% consistent.

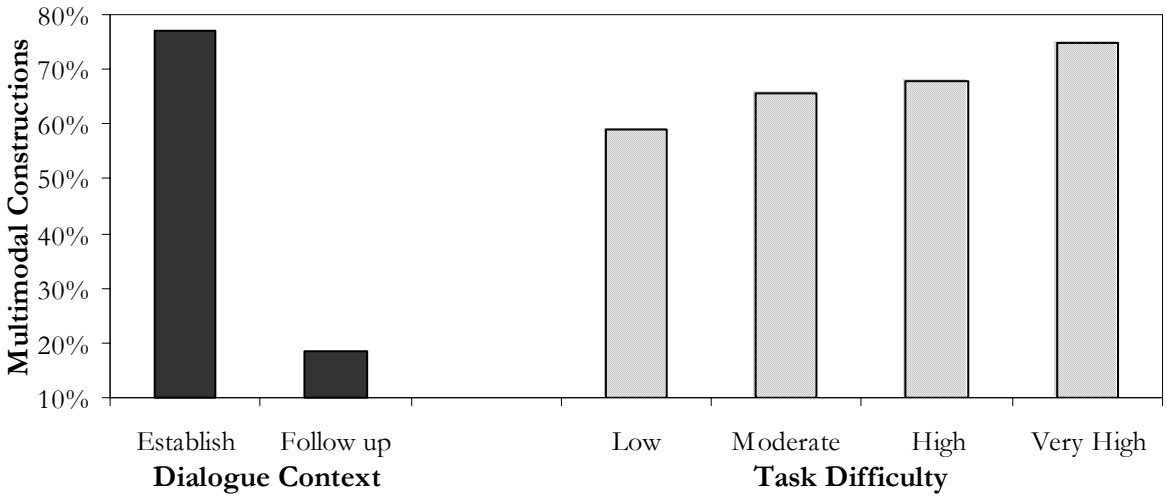

Figure 3. Percentage of multimodal constructions as a function of establishing dialogue context and increasing task difficulty

3.1.1 Multimodality and Dialogue Context

Data on approximately 140 constructions, or 70 matched pairs, were available for analysis on the likelihood of interacting multimodally as a function of establishment of dialogue context. As illustrated in Figure 3, the percentage of constructions delivered multimodally was 77.1% when a new dialogue context was being actively established, but dropped to 18.6% during follow-up constructions after the user already had established context. This was a significant decrease by *a priori* paired t test, $t = 6.40$ (df = 9), $p < .001$, one-tailed. Furthermore, this difference represented a +315% relative increase (i.e., over 4-fold) in the likelihood that users will communicate multimodally to a system when they need to establish dialogue context.

3.1.2 Multimodality and Task Difficulty

Data on approximately 480 constructions were available for analysis on the likelihood of interacting multimodally as a function of task difficulty level. As shown in Figure 3, the percentage of tasks completed multimodally increased from 59.2% in the low difficulty tasks, to 65.5% in the moderately difficult ones, to 68.2% during high difficulty, and 75.0% during very high difficulty. These changes represented a significant difference by *a priori* paired t-test between the low/moderate difficulty levels and the high/very levels, $t = 2.22$ (df = 9), $p < .03$, one-tailed. The net change in the likelihood of users communicating to the system multimodally across the full range between the lowest and highest task difficulty levels represented a +27% relative increase.

3.1.3 Multimodality and Task Order

When constructions for the session were divided according to the first versus second half, the likelihood that users communicated multimodally to the system decreased from 64.0% to 59.9% between these initial and final phases as they became more familiar with the system and tasks, although this change did not represent a significant difference by *a priori* paired t-test, $t = 1.06$ (df = 9), N.S.

3.2 Performance Measures

Data on 480 constructions were available for analysis of task-critical performance errors as a function of task difficulty level.

3.2.1 Human Performance Errors

During the course of their session, all participants made task-critical performance errors. As illustrated in Figure 4, the average ratio of such errors per total tasks increased from .008 during the lowest difficulty tasks, to .033 during moderately difficult tasks, to .192 during high difficulty tasks, and then remaining constant at .192 when tasks became very difficult.

These changes in task-critical performance errors represented a significant increase by Wilcoxon Signed Ranks test between the low/moderate difficulty levels and the high/very levels, $T+ = 55$ ($N = 10$), $p < .001$, one-tailed. The net change in users' task-critical performance errors between the lowest and highest task difficulty levels represented a +2300% relative increase (i.e., over 24-fold) in errors.

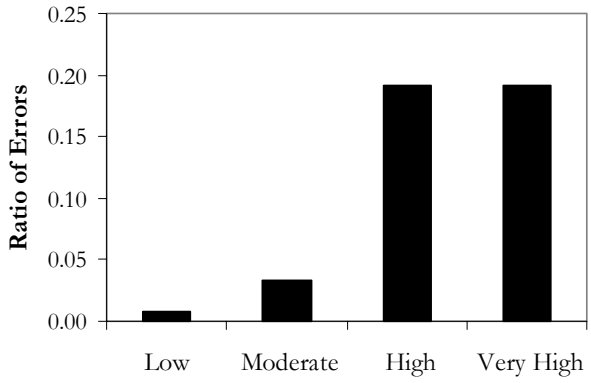

Figure 4. Task-critical human performance errors per 100 constructions as a function of task difficulty level

3.2.2 Response Latency

Response latency averaged 9.3 seconds, and ranged between 1.6 and 50.4 seconds. As shown in Figure 5, response latency

between receiving an instruction and initiating a task increased from an average of 5.5 seconds during the low difficulty tasks, to 6.0 seconds in the moderately difficult ones, 11.2 seconds during high and 14.5 seconds during very high task difficulty.

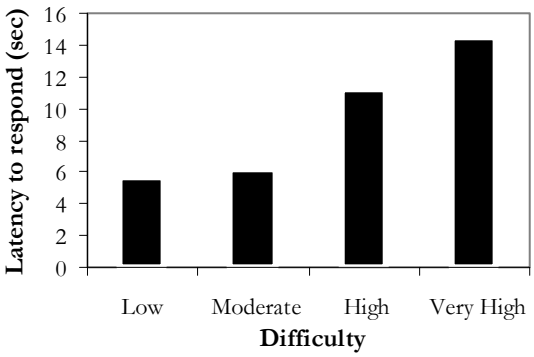

Figure 5. Average latency to initiate a task

This difference was significant for response latencies (log transformed) between the low and moderately difficult task levels, *a priori* paired t-test, $t = 2.22$ (df = 9), $p < .03$, one-tailed. It also represented a significant difference between the moderate to high levels, *a priori* paired t-test, $t = 16.06$ (df = 9), $p < .001$, one-tailed, and between the high to very high difficulty tasks, *a priori* paired t-test, $t = 3.12$ (df = 9), $p < .01$, one-tailed.

4. DISCUSSION

The present findings indicate that users of a multimodal interface spontaneously respond to dynamic changes in their own cognitive load by shifting to multimodal communication as load increases with task difficulty and communicative complexity. Given a flexible multimodal interface, users' ratio of multimodal interaction (versus unimodal) increased dramatically from 18.6% when following up on an established dialogue context, to 77.1% when they had to establish an entirely new dialogue context— a +315% relative increase that is illustrated in Figure 3. Likewise, the ratio of users' multimodal interaction increased significantly as tasks became more difficult, from 59.2% during low difficulty tasks, to 65.5% at moderate difficulty, 68.2% at high, and 75.0% at very high difficulty—an overall relative increase of +27%. These adaptations in multimodal interaction levels reflect users' effort to self-manage limitations in their working memory as discourse-level demands and task complexity increased. They accomplished this by distributing communicative information across multiple modalities in a manner compatible with a cognitive load theory of multimodal interaction. This interpretation is consistent with Wickens' cognitive resource theory [28] and Baddeley's theory of working memory [2], as well as the growing literatures within education [24], linguistics [1], and multisensory perception [4].

The striking findings on dialogue context involved adjacent pairs of tasks, with no other dialogue turns intervening to interject time delay. In addition, when establishing new dialogue context a noun phrase incorporating spatial information expressed the creation of a new entity (e.g., "Make an evacuation route from Peniel Missions to Washington High School shelter" [draws route]), whereas during the adjacent follow-up task a briefer deictic or noun phrase expression replaced the more complex antecedent noun phrase (e.g., "Now evacuate the people out of Peniel Missions along *that route*."). These two factors (i.e., the difference in complexity of noun phrase expressions, and the adjacency of dialogue turns) probably played a major role in generating the very large contrast observed in users' likelihood of communicating multimodally as a function of dialogue state. Almor's work [1] on information load during processing different types of anaphoric reference suggests that there may in fact be finer-grained variations in users' likelihood of communicating multimodally, which could be modeled as a function of more subtle differences in referring expressions, attentional focus, and the temporal distance between dialogue contributions. At any rate, future extensions of the present research should assess the generality of the present findings with a larger and more diverse range of dialogue contributions.

In terms of performance, the incidence of task-critical errors (e.g., placing an object at the wrong location) increased systematically with task difficulty, as did the duration of users' response latencies between receiving a task and initiating their input to the system. From a methodological viewpoint, this establishes the validity of the present study's task difficulty levels, and it provides behavioral corroboration of the additional cognitive load that users were experiencing. It also is clear that people's ability to perform accurately was extremely sensitive to increases in location and directional information in this spatial domain.

With respect to preferences and individual differences, 61.8% of all user interactions with the system involved multimodal communication, confirming past findings of a strong preference for multimodal interaction [16]. When interacting multimodally, all participants were classifiable as either predominantly simultaneous or sequential integrators, as shown in Table 2. Their consistency level in maintaining this dominant pattern averaged 93.5%, which replicates previous reports [16, 20, 29]. When users did interact unimodally, they were twice as likely to prefer speech over pen input, and their dominant modality preference also tended to remain highly consistent throughout the session.

Future multimodal systems will need to distinguish between instances when users are and are not communicating multimodally, so that accurate decisions can be made about when parallel input streams should be interpreted jointly versus individually. Although users like being able to interact multimodally, they don't always do so when given free choice. Their natural communication patterns involve mixing unimodal and multimodal expressions, with the overall ratio of multimodal constructions predictable based on dialogue context and task difficulty. The long-term goal of this research is the development of an empirical foundation for proactively guiding flexible and adaptive multimodal system design. Future work should explore people's likelihood of interacting multimodally while using other modalities such as speech and manual gesturing, and during tasks other than visual-spatial ones.

5. ACKNOWLEDGMENTS

Thanks to Benfang Xiao and Matt Wesson for acting as wizards during testing, and Benfang Xiao and Josh Flanders for assistance with data collection, scoring, and second scoring. This research was supported by DARPA Contract No. NBCHD030010 and NSF Grant No. IIS-0117868. Any opinions, findings or conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency, or the Department of the Interior.

6. REFERENCES

[1] Almor, A., Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, 1999. **106**: 748-765.

[2] Baddeley, A., Working Memory. *Science*, 1992. **255**: 556-559.

[3] Benoit, C., J.-C. Martin, C. Pelachaud, L. Schomaker, & B. Suhm, Audio-visual and multimodal speech-based systems, *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*, R. Moore, ed. 2000, Kluwer Academic Publishers: Boston, MA. 102-203.

[4] Calvert, G., C. Spence, & B.E. Stein, eds. *The handbook of multisensory processing*. 2004, MIT Press: Cambridge, MA.

[5] Chandler, P. & J. Sweller, Cognitive load theory and the format of instruction. *Cognition and Instruction*, 1991. **8**: 293-332.

[6] Grant, K.W. & S. Greenberg. Speech intelligibility derived from asynchronous processing of auditory-visual information. *Workshop on Audio-Visual Speech Processing (AVSP-2001)*. 2001. Scheelsminde, Denmark

[7] Grice, H.P., Logic and conversation, *Syntax and Semantics: Speech Acts*, J.L. Morgan, ed. 1975, Acad Press: NY. 41-58.

[8] Hinckley, K., Pierce J., E. Horvitz, & M. Sinclair, Foreground and background interaction with sensor-enhanced mobile devices. *ACM Transactions on Computer Human Interaction, in press (Special Issue on Sensor-Based Interaction)*.

[9] Jacko, J., L. Barnard, T. Kongnakorn, K. Moloney, P. Edwards, V. Emery, & F. Sainfort. Isolating the effects of visual impairment: Exploring the effect of AMD on the utility of multimodal feedback. *Conf. on Human Factors in Comp. Systems: CHI '04*. 2004. NY, NY: ACM Press

[10] Jameson, A., Adaptive interfaces and agents, *The human-computer interaction handbook*, A. Sears, ed. 2003, Lawrence Erlbaum Associates: Mahwah NJ. 305-330.

[11] Mayer, R.E. & R. Moreno, A split-attention effect in multimedia learning: evidence for dual processing systems in working memory. *Journal of Educational Psychology*, 1998. **90**(2): 312-320.

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

[13] Müller, C., B. Großmann-Hutter, A. Jameson, R. Rummer, & F. Wittig. Recognizing time pressure and cognitive load on the basis of speech: an experimental study. *User Modeling*. 2001: Springer

[14] Oviatt, S.L., Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 1995. **9**: 19-35.

[15] Oviatt, S.L., Multimodal interactive maps: Designing for human performance. *Human Computer Interaction*, 1997. **12**(1-2): 93-129.

[16] Oviatt, S.L., Ten myths of multimodal interaction. *Communications of the ACM*, 1999. **42**(11): 74-81.

[17] Oviatt, S.L. Mutual disambiguation of recognition errors in a multimodal architecture. *ACM SIGCHI Conf. on Human Factors in Comp. Sys. (CHI'99)*. 1999. Pittsburgh, PA: ACM Press: 576-583.

[18] Oviatt, S.L. Multimodal system processing in mobile environments. *13th ACM Symp. on User Interface Software Tech. (UIST'2000)*. 2000. New York: ACM Press: 21-30.

[19] Oviatt, S.L., P.R. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T.G. Holzman, T. Winograd, J. Landay, J. Larson, & D. Ferro, Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions. *Human Computer Interaction*, 2000. **15**(4): 263-322.

[20] Oviatt, S.L., R. Coulston, S. Tomko, B. Xiao, R. Lunsford, M. Wesson, & L. Carmichael. Toward a theory of organized multimodal integration patterns during human-computer interaction. *Internat. Conf. on Multimodal Interfaces*. 2003. Vancouver, B.C.: ACM Press: 44-51.

[21] Oviatt, S.L., T. Darrell, & M. Flickner, Multimodal Interfaces that flex, adapt, and persist, *Comm. of the ACM*. 2004. 30-33

[22] Penney, C.G., Modality effects and the structure of short-term verbal memory. *Memory and Cognition*, 1989. **17**: 398-422.

[23] Prince, E., Toward a taxonomy of given-new information, *Radical Pragmatics*, P. Cole, ed. 1986, Academic: NY. 223-255.

[24] Sweller, J., Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 1988. **12**: 257-285.

[25] Technology for adaptive aging: Reports and papers. 2003, Nat. Acad. of Sci. Workshop: Nat. Acad. Press. http://www.nap.edu/books/0309091160/html/

[26] Teder-Sälejärvi, W.A., J.J. McDonald, F. Di Russo, & S.A. Hillyard, An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cognitive Brain Research*, 2002. **14**: 106-114.

[27] Tindall-Ford, S., P. Chandler, & J. Sweller, When two sensory modes are better than one. *Journal of Experimental Psychology: Applied*, 1997. **3**(3): 257-287.

[28] Wickens, C., Sandry, D., and Vidulich, M., Compatibility and resource competition between modalities of input, central processing, and output. *Human Factors*, 1983. **25**(2): 227-248.

[29] Xiao, B., R. Lunsford, R. Coulston, M. Wesson, & S.L. Oviatt. Modeling multimodal integration patterns and performance in seniors: Toward adaptive processing of individual differences. *Internat. Conf. on Multimodal Interfaces*. 2003. Vancouver, BC: ACM Press: 265-272.