

Examining the Redundancy of Multimodal Input

First Author Name

Affiliation

Address

e-mail address

Optional phone number

Second Author Name

Affiliation

Address

e-mail address

Optional phone number

ABSTRACT

Speech and gesture modalities can allow users to interact with complex applications in novel ways. Often users will adapt their multimodal behaviour to cope with increasing levels of domain complexity. These strategies can change how multimodal constructions are planned and executed by users. In the frame of Baddeley's Theory of Working Memory, we present some of the results from an empirical study conducted with users of a multimodal interface, under varying levels of cognitive load. In particular, we examine the multimodal behavioural features were sensitive to cognitive load variations. We report significant decreases in multimodal redundancy (33.6%) and trends of increased multimodal complementarity, as cognitive load increases.

Author Keywords

Cognitive load, multimodal interface, speech, gesture

ACM Classification Keywords

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

INTRODUCTION

Multimodal interfaces expand the communication channel between the system and the user allowing users to express themselves more naturally and interact with complex information with more freedom of expression (Oviatt, 2003) This research aims to provide more consistent, integrated and intuitive human-computer interaction and decrease the users' cognitive load through the multimodal paradigm. One of the many cited advantages of multimodal interfaces is their ability to facilitate effortful complex tasks (Oviatt, 2003) over unimodal interfaces.

Cognitive load refers to the amount of mental effort imposed by a particular task, and has been associated with the limited capacity of working memory (Paas, 2003) In a Traffic Incident Management (TIM) scenario, operators are bombarded with live information that needs to be integrated, synthesised and entered into the computer system. They need to make decisions on the fly as to how to handle many incidents at once, tasks which combine to induce very high levels of cognitive load.

OzCHI'06, November 22-24, 2006, Sydney, Australia.

Copyright the author(s) and CHISIG

Additional copies are available at the ACM Digital Library (<http://portal.acm.org/dl.cfm>) or ordered from the CHISIG secretary (secretary@chisig.org)

OZCHI 2006 Proceedings ISBN: x-xxxxx-xxx-x

Currently, cognitive load is measured by soliciting subjective load ratings after a task is completed (Paas et. al. 2003). Subjects are rarely assessed in real time and the probe-method interrupts the user's task flow. Research has shown that users adapt their multimodal behaviour in complex situations to increase their performance (Oviatt, 2004). Implicit and unobtrusive measurement of day-to-day multimodal cues to determine users' cognitive load in real time can help us to design interfaces that constantly adapt content selection and presentation processes accordingly, in order to ensure optimal user performance.

This paper describes an experiment designed to identify the relationships between combined speech and gesture input structures and users' cognitive load. The two input modalities are very familiar to users and psychologically closely interrelated, both in terms of planning and execution (McNeill, 1992). Specifically, our hypothesis is that variations in redundant and complementary multimodal constructions can reflect cognitive load changes experienced by the user.

We begin with some definitions, illustrating multimodal patterns that may be monitored to detect cognitive load variations based on symptomatic behavioural features. We then present some early results, concluding with a discussion on the impact such methods may have on the design of dynamic, perceptually-effective human computer interaction systems, but highlight the current limitations of the pattern acquisition methodology.

Definitions

We developed a scheme for multimodal interaction annotations, focussing on identifying redundant input from the users. Each task is divided into a number of *turns*. A turn is a semantically complementary group of input *constructions* with a single intention. An example is "Marking an incident on the map". This turn would require two constructions: a SELECT and a TAG. *Multimodal turns* are any completed using more than one different modality.

Each construction (e.g. SELECT, or TAG) comprises one or two semantic *attributes*: an action and/or an object, as in Epps (Epps, 2004). Subjects may provide input for a construction using one or more modalities. In the latter case, the inputs are, by definition, *semantically redundant*, since they will convey the same meaning. For example, a completely redundant SELECT construction may consist of: [pointing gesture at library icon] + spoken: "Select Library". Either input on its own would have achieved the same action, making this construction redundant. Constructions can also be partially redundant, if only the action or the object is doubled up, e.g.:

[pointing gesture at library icon] + spoken: “Select”. Both partially redundant and completely redundant constructions were annotated as *redundant*.

Turn	Const	Modality	Content
Mark an Incident Pure Redund	Select	Gesture	[point to St Mary’s Church]
		Speech	Select St.Mary’s Church
	Tag	Hand_Shape	[scissors=Incident]
		Speech	Incident
Mark an Accident Pure Compl	Select	Speech	“Select Crown Street Library”
	Tag	Hand_Shape	[fist=Accident]
Mark an Event Part Redund	Select	Speech	“Select”
		Gesture	[point to Collingwood School]
	Tag	Hand_Shape	[open_palm=Event]

Table 1: Pure Redundant, Partial Redundant and Pure Complimentary Turns

Some turns are made up of only redundant constructions, hence called *purely redundant*. Other turns do not contain any redundant constructions and are *purely complementary*. The rest are called *partially redundant*. Examples of each can be seen in Table 1.

Given the documented advantages of multimodal input, we hypothesized that the users’ input would be less likely to be redundant as cognitive load increased. In fact, we expected multimodal turns to become more complementary, i.e. users would begin to employ the broader multimodal communication channel more effectively, directing separate semantic inputs via different modalities, with no semantic overlap between constructions. Thus, rates of purely redundant turns would decrease as cognitive load increased, and rates of purely complementary turns would increase.

EXPERIMENT METHOD

Wizard of Oz Set-up

Given our aim to observe users under varying levels of cognitive load while using a multimodal interface, we constructed a Wizard of Oz (WOz) system that would allow the implementation of ‘error-free’ multimodality. Free of bias from the high number of error rates of actual recognisers, WOz can capture a user’s *natural* multimodal behaviour.

The application was projected onto a large wall-sized screen. The subject stood 2m in front of the screen; a video camera was placed on the right hand side of the screen, facing her. A smaller camera was placed in front of her right arm (all users were right-handed) to capture the gesture input.



Figure 1: Subject Interacting with the System

Based on the audiovisual feed from the video camera, the wizard interpreted the subject’s interaction with the system. The wizard could also see the hand’s position as tracked by the recogniser, and act accordingly, e.g. click on buttons or make selections on her own interface; the output would then be rendered on the subject’s graphical user interface, shown in Figure 2.

Application Scenario: Tasks and Modalities

Our aim was to elicit natural speech and gesture interaction under different levels of cognitive load. Inspired by the TIM scenario, the study required subjects to update a geographical map with traffic conditions information, using either natural speech, or manual gesture, or a combination of these.

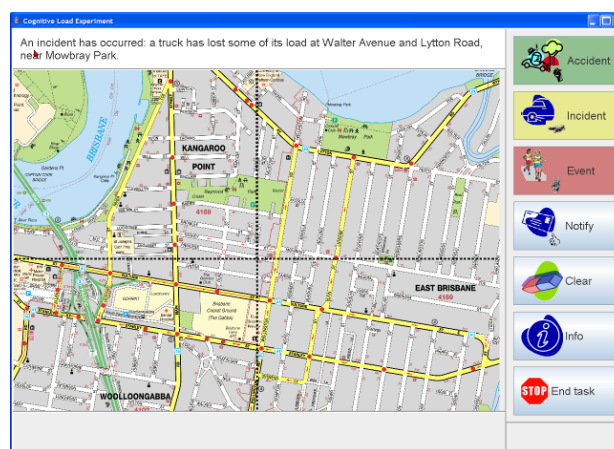


Figure 2: Graphical User Interface

Examples of interaction and system functionality were given during a 30-min training session. Examples are shown in Table 2. Available gestures encompassed:

- Deictic pointing to map locations, items, and function buttons;
- Circling gestures for zoom functions; and
- Predefined hand shapes for item tagging.

The experiment was designed such that all tasks could be completed using any single modality. Each new task was described at the top of the screen and the subjects were allowed freedom of inspection.



System Functionality	Example of Interaction
Zooming in or out of a map	<Point at quadrant>; or "Zoom in to the top right quadrant"
Selecting a location/item of interest	<Point at location>; or "St Mary's Church"
Tagging a location of interest with an 'accident', 'incident' or 'event' marker	<Select location> and: "Incident"; or Scissors shape 
Notifying a recipient (item) of an accident, incident or an event	<Select accident> and "notify"; or fist shape  and <Select recipient>
Starting or ending a task	"End task"; or <Point at End task button>

Table 2: System Functionality and Examples of Inputs

There were four levels of cognitive load, and three tasks were completed for each level. The same map was used for each level to avoid differences in visual complexity. The tasks varied in load through:

- The number of distinct entities in the task description;
- The number of distractors (items not needed for task);
- The minimum number of actions required for the task.

Further load was achieved in Level 4 by introducing a time limit.

Level	Entities	Actions	Distractors	Time
1	6	3	2	∞
2	10	8	2	∞
3	12	13	4	∞
4	12	13	4	90 sec.

Table 3: Levels of Cognitive Load

Procedure

Twelve (6 females, 6 males, aged 18-49) remunerated, random, native English-speaking participants completed the study. In the actual experiment, subjects were asked to perform a set of tasks under 3 different conditions: gesture-only, using speech-only and using multimodality. Each set consisted of 4 levels, 3 tasks in each. The order of these conditions and the tasks within the levels was randomised to balance order effects. In this paper, we present results for the multimodal condition only.

Biosensor data, namely blood volume pulse (BVP) and skin conductance (GSR), was collected from each subject through sensors placed on their fingertips. Video, position, UI interaction and biosensor data were synchronised and recorded digitally. Users were also debriefed after each task level and were asked to rank the level of load relative to the other levels in that condition.

The video data collected from the subjects was manually annotated. All gestures were classified and tagged with duration; speech commands were orthographically transcribed.

PRELIMINARY RESULTS

Out of 12 subjects, only the data from 9 was usable, since two users had difficulty comprehending the tasks, such that they could not achieve the goals of the task, and one did not finish for external reasons. The data collected for the multimodal condition for Levels 1, 2 and 4, was annotated for 6 users. In total, 1119 modal inputs were annotated, forming 394 turns and 644 constructions. However, smaller numbers were used for the analysis of individual levels. We carried out the statistical analysis only on 5 of these 6 users for reasons outline in the discussion section. To measure the perceived level of cognitive load, users ranked the tasks in increasing levels of difficulty along a 9-point Likert scale, the average difficulty score for Levels 1, 2 and 4 across these 6 users was 2.2, 4.2 and 5 respectively.

Multimodal Redundancy and Complementarity

For each user, we classified the multimodal turns into three groups: purely redundant, purely complementary and partially redundant turns. Figure 3 shows the mean percentage and range of purely redundant turns across users, for each level, over all multimodal turns.

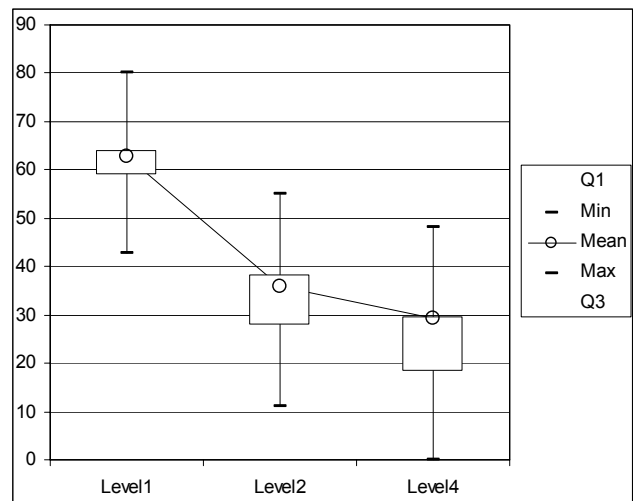


Figure 3: Proportion of Purely Redundant Turns by Level

We observed a steady decrease in redundancy as task difficulty increased. An ANOVA test between-users, across levels, shows there are significant differences between the means ($F = 3.88$ ($df=2$); $p < 0.05$). Subsequent t-tests show significant differences, 27.16% between Level 1 (62.91%) and Level 2 (35.74%) ($p = 0.03$, < 0.05 , two-tailed) and 33.61% between Level 1 and Level 4 (29.29%) ($p = 0.01$, < 0.05 , two-tailed).

By the same token, we expected the rate of purely complementary constructions to increase. In Level 1, the average percentage of purely complementary turns was 12.86%, increasing to 45.53% and 36.02% in Levels 2 and 4 respectively. Though not significantly different, there is a clear trend across users of an increased use of complementary multimodal constructions in higher load

tasks when comparing Level 1 and 2, and 1 and 4, corroborating with decreasing redundancy. There is also an increase in partially redundant constructions between Levels 1 and 4, with averages of 24.24%, and 34.69% respectively, which can be interpreted as representative of the shift from purely redundant to partially redundant to purely complementary as can be seen in Figure 4.

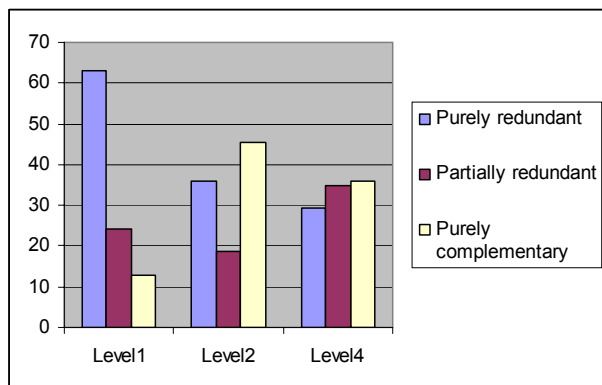


Figure 4: Averages of Purely Redundant, Purely Complementary, and Partially Redundant Turns

Overall, we explain the lack of difference between Level 2 and Level 4 in the above features as a symptom of the fact that users' average subjective rating of these difficulty levels was very similar in both cases (4.2 and 5), but at least 2 points higher than Level 1 (2). This was probably due to increased familiarity with the system by the time the 4th level was attempted, highlighting the semantic distinction between task complexity and cognitive load. A more complex task does not necessarily imply a higher experienced load.

DISCUSSION

To begin, we did not use the data for one of the subjects, who exhibited a rate of 0% redundant constructions in Levels 1 and 2. The subject's difficulty rating was the same for both levels. We rationalise that measuring the level of redundant input in such a user would not be useful in collecting evidence for fluctuations in cognitive load. This suggests that for our approach to be viable in automatic cognitive load estimation, a compound measure would be necessary as not all features may be suitable indicators for all users. The sensitivity of rates of complementarity or redundancy to variations in cognitive load needs to be established. Robustness may be achieved by weighting each multimodal index specifically for each user, to ensure the more reliable indicators can influence the combined reading more strongly.

However, the overall findings we present in this paper are indicative of changes in multimodal input behaviour by users as cognitive load increases. The reduced level of redundancy, combined with the increased level of complementarity can be interpreted within the frame of Working Memory Theory. Baddeley's modal Model of Working Memory suggests there are areas of working memory that are dedicated *exclusively* for modal use, e.g. the visuo-spatial sketchpad, and the phonological loop (Baddeley, 1992). During interaction at high levels of

load, a strategy that can be employed by users when planning constructions, is to maximise the usage of *modal* working memory. This could be achieved by channelling the required semantic chunks to different modalities, with the least amount of replication possible. This approach would result in increased purely complementary constructions and a reduction in purely redundant constructions as cognitive load increases. The results of this study give initial evidence for this behavioural symptom of cognitive load management employed by users.

CONCLUSIONS

Assessing a user's cognitive load through their multimodal behaviour requires identifying a number of indices that reliably reflect load fluctuations. A major advantage of this approach is that cognitive load can be determined implicitly by monitoring variations of specific multimodal features during day to day tasks. The feasibility of using rates of redundancy or even complementarity in multimodal input as an index of cognitive load is supported by the results of our study. A significant decrease was observed in the number of purely redundant turns from 62.91% to 29.9% of all multimodal turns, as cognitive load increases. Trends of increasing purely complementary and partially redundant turns were also observed.

The ability to implicitly measure the perceived level of cognitive load through changes in multimodal behaviour could play a crucial role in human computer interaction design, as applications could adapt the output flow and presentation to the current load of the user without intrusive probes. This would allow continuously optimal delivery of content, in a very user-centric way.

In spite of these encouraging initial results, there are still major scientific challenges to be addressed. Further annotation of our experiment data will help refine the methodology and findings. The implementation of automated processing of redundant and complementary turns data also represents an important technical challenge, though customised unification algorithms from a multimodal input fusion module are envisaged.

REFERENCES

- Baddeley, A., Working Memory. Science, 1992. 255: 556-559.
- Epps, J., Oviatt, S., and Chen, F., Integration of Speech and Gesture Inputs during Multimodal Interaction, in *Proc Aust. Int. Conf. on CHI* (2004).
- McNeil, D. (1992) Hand and Mind: What Gestures Reveal about Thought. Chicago: University Chicago Press.
- Oviatt, S., Coulston, R., and Lunsford, R., When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. In *Proc. Int. Conf. on Multimodal Interfaces* (2004).
- Paas, F., et. al., Cognitive load measurement as a means to advance cognitive load theory. Educational Psychologist (2003), 38, 63-71.