Integration of Speech and Gesture Inputs during Multimodal Interaction

Julien Epps[1], Sharon Oviatt[2], and Fang Chen[1]

[1]Multimodal User Interaction
National ICT Australia
Eveleigh NSW 1430, Australia
{julien.epps, fang.chen}@nicta.com.au
http://nicta.com.au/

[2]Department of Computer Science and Engineering
Oregon Health and Science University
Beaverton, OR 97006, USA
oviatt@cse.ogi.edu
http://www.cse.ogi.edu/CHCC/

Abstract

Speech and gesture are two types of multimodal inputs that can be used to facilitate more natural human-machine interaction in applications for which the traditional keyboard and mouse input mechanisms are inappropriate, however the possibility of their concurrent use raises the issue of how best to fuse the two inputs. This paper analyses data collected from a speech and manual gesture-based digital photo management application scenario, and from this derives assumptions and fusion thresholds with which future speech/gesture systems can be designed. Gesture input was found to overlap with speech input in nearly all multimodal constructions (95%), and was completely subsumed by speech input in most multimodal constructions (56%), in distinct contrast to previous similar analyses for combined pen and speech input, in which the pen input frequently precedes the speech input.

Keywords

Multimodal interaction, speech and gesture input, fusion, synchronization, segmentation, interaction styles.

INTRODUCTION

Multimodal interfaces, while still lacking the maturity necessary to begin replacing the ubiquitous QWERTY keyboard or mouse, are increasingly finding applications in tasks for which it is either undesirable or impossible to use a keyboard or mouse. Where interaction occurs without touching or wearing any equipment, speech and gesture are two important modalities for which combined use dates back to the seminal work of Bolt (1980). Bolt demonstrated that by fusing complementary information from deictic gestural input, speech input could become much more natural than had previously been considered feasible. His suggested applications for this work were based on manipulation of objects in a geographical context, and much recent multimodal research has been conducted with these applications in mind, for example real estate consultancy (Oviatt *et al.*, 1997), geographical information systems (Cohen *et al.* 1997, Schapira and Sharma 2001, Krahnstoever *et al.*, 2002), and collaborative military mission planning (Cohen *et al.* 1997, Flippo *et al.*, 2003).

Input fusion is an essential multimodal system component that combines the often complementary (Oviatt *et al.*, 1997) semantic information from different input modalities. The first multimodal system to use both semantic and temporal constraints on what constitutes a 'legal' or viable fusion operation was Quickset (Cohen *et al.* 1997), which used empirical information from pen and speech data reported in (Oviatt *et al.*, 1997) to determine fixed temporal constraints, or thresholds. Fusion has previously been achieved using a number of different approaches, however most researchers aim in one form or another to group simultaneous or sequential inputs that occur close to each other in time. More recently, Gupta (2003) outlined an adaptive method for fusion, however no details of what algorithm should be used to modify the thresholds were given. Another recent fusion technique (Flippo *et al.*, 2003) imposes semantic constraints using a parse tree, which is populated by time-stamped inputs and then operated on by resolving agents that deal with ambiguous input combinations.

In the temporal threshold approach to the integration of multimodal inputs (Cohen *et al.* 1997), a brief time window is allowed after the first modality input has completed for subsequent inputs from other modalities to commence. If the beginning of an input from another modality is detected before the end of this time window, the two inputs are assumed to be semantically related. The duration of this time window is of considerable interest since if it is too short, related inputs may be treated separately by the multimodal system, while if it is too long, users may find the delay in system response unacceptably long. Here it is assumed that a pair of sequential inputs treated separately by the multimodal system's fusion component will produce two unimodal interpretations. Temporally overlapping multimodal inputs can be fused simply by waiting for the longer input to complete.

In this paper, experiments on a simple digital photo management mock application were performed with the objectives of gaining insight into how speech and manual gesture input modalities are used, and deriving information on temporal thresholds for the grouping of related inputs in a multimodal system. The motivation for

embarking upon a new user study to address these objectives is derived mainly from the extreme scarcity of publicly available multimodal corpora, particularly those combining speech and free hand gesture inputs.

METHOD

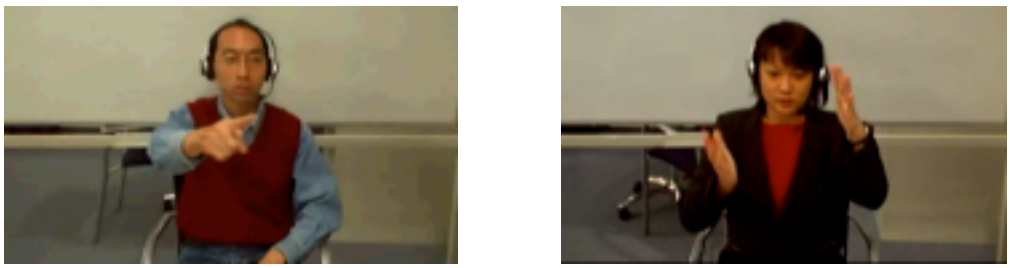
Digital Photo Management Application

Experiments were constructed around a mock digital photo management application, comprising an icon-style layout of slide show folders and photos, and a series of associated commands. A photo management scenario was specifically chosen for this study because the combination of visual icons and the many operations possible on those icons was considered to favour multimodal interaction. A further motivation for the choice was the feasibility of such an application being useful in a physical space such as a family living room, where a traditional PC is an undesirable piece of furniture. A wall-mounted flat screen computing device with a small built-in camera might well be ergonomically and aesthetically acceptable in such a space, as a photo management and display tool, or even ultimately as a more fully blown computing platform (Wilson and Oliver, 2003).

Subjects, Tasks and Procedure

Eighteen subjects completed the experiment, 14 males and 4 females between the ages of 20 and 50. All participants were frequent computer users, however some participants did not have a technology background. All subjects completed the same seven tasks, and the responses of each subject were recorded on video, as seen in the examples of figure 1.

All seven tasks consisted of performing a single action on a photo or slideshow folder: moving a photo into a folder (two tasks), deleting a folder, rotating a photo, zooming in on a photo, renaming a folder and enlarging a photo. These tasks were typical of those likely to be required of a commercial digital photo management application. Physically, each photo or folder in the application was represented on a single colour printed A4 sheet of paper. Participants sat approximately 1.5m from the mock application, and were instructed to perform commands such as moving a photo into a slide show folder and rotating a photo, using speech and/or gesture. The instruction given to subjects was to use what they considered the simplest and most natural means of completing the assigned tasks, with whatever gestures and/or speech they felt were needed.



(a) (b)

Figure 1: Two study participants completing tasks using manual gesture combined with spoken commands “Move that photo to the animal slide folder” (a) and “People picture rotate anticlockwise” (b).

RESULTS AND DISCUSSION

Modality and integration pattern

Of the 126 total constructions, 21% were unimodal speech-only, 16% were unimodal gesture-only, and 63% were multimodal. This is a very similar result to that obtained in a recent study by Oviatt *et al.* (2004), which reported 25.6% speech-only, 12.6% pen-only and 61.8% multimodal constructions during a simulation rather than a hard-copy mock-up as in the present research. Of the 80 multimodal constructions, 95% were delivered simultaneously, with only 5% delivered sequentially. As in Oviatt *et al.* (2004), most subjects showed a strong preference for one modality or the other when interacting multimodally, with 11 out of 18 delivering all their unimodal constructions in their favoured modality. One participant delivered all seven of his responses using speech only. Apart from one subject who was a ‘sequential integrator’ (66% of her multimodal constructions included a lag between signals), the remaining 16 subjects who interacted multimodally were 100% consistent as ‘simultaneous integrators’. This very high overall average consistency (98%) concurs with similar results recently reported for speech and pen inputs (Oviatt *et al.*, 2003).

Types of gesture used

For tasks that involved moving a photo into a photo slideshow folder, nearly all responses exhibited a straightforward gesture: pointing at the photo, then moving their arm to point at the slideshow folder. 44% of

constructions that included gesture comprised a straightforward pointing gesture, while 25% used a more complex gesture. Commonly used complex gestures included rotating the hand(s) for a 'rotate photo' task, or opening/closing the hand(s) for a 'zoom' task. Twenty-nine percent of all multimodal constructions used a spoken deictic expression that referred to the accompanying gesture (i.e. the speech included "this", "that" or "there"). In all cases the deictic expression was redundant due to the simple nature of the tasks, however for a more complex task (such as selection of multiple objects by gesture and speech: "select that and that and that"), this information would become more significant.

Use of redundant vs. complementary inputs

Of the 80 multimodal constructions, 35% contained purely complementary information, while 30% contained purely redundant information. An example of a complementary input would be pointing to a photo and saying 'delete'. Purely redundant here means any multimodal construction for which the separate modalities taken individually would each contain sufficient information to successfully complete the task, e.g. saying 'rotate the group photo' while pointing to it and rotating the hands. Of the multimodal constructions, 18% contained redundant information only about the object referred to in the task, while 16% contained redundant information only about the action required by the task.

Every task completed using gesture only began by specifying the object, and then indicating the action to be performed. Of tasks completed by speech only, 73% began by specifying the action, followed by the object on which it was to be performed. The most likely reason for this is that in spoken English, the verb usually precedes the object. Subjects using both speech and gesture, whether redundant or complementary, tended to specify the action first if they spoke first, and the object first if they used gesture first. Subjects who used only the minimal complementary commands (e.g. a pointing gesture together with a spoken 'delete' command) tended to begin by specifying the object.

Temporal relationships of input modes

All but four of the multimodal constructions were found to be temporally overlapping (95%), and in 81% of these constructions the longer input completely overlapped the shorter, suggesting that for this task, subjects usually found it natural to use simultaneous speech and gesture. These contrast with results of similar analyses for speech and pen (Oviatt *et al*., 1997, Oviatt *et al.*, 2003) where a simultaneous integration pattern has represented a lower percentage of users' multimodal input in different studies (68-91%). The sub-classifications of overlap for simultaneous integrations are shown in Table 1, where the temporal accuracy is 0.1 seconds. The main difference of note between Table 1 and the similar sub-classifications for speech and pen data reported in Table 4 of (Oviatt *et al*., 1997) is that it was most common in these data for gesture to be completely subsumed by speech (56% in Table 1 vs. 19% in speech and pen data). In contrast, subjects using pen and speech preferred to draw something first and then refer to it by speech (57% in speech and pen data vs. 32% in Table 1). The speech and pen integration data also exhibited more co-timing of signal onsets and offsets than is seen in Table 1.

Table 1. All logically possible temporal overlap patterns between speech and manual gesture input for simultaneous integrations, sub-classified by temporal precedence of input mode.

Speech precedes (60%)		Gesture precedes (32%)		Neither mode precedes (8%)	
G ——— S ————	(6%)	G ———— S ———	(9%)	G ———— S ————	(0%)
G ——— S ————	(46%)	G ———— S ———	(12%)	G ———— S ———	(4%)
G ——— S ———	(8%)	G ——— S ———	(11%)	G ——— S ————	(4%)

It is instructive to examine the four instances of temporally disjoint multimodal inputs. In one instance, the subject pointed at the nominated photo, withdrew his hand slightly, but did not fully withdraw his hand until he had finished speaking. In another, the subject appeared to be pausing after his gesture to refer to the study instructions before giving the spoken command. In the two remaining instances, the intermodal lags between gesture and speech inputs were around 0.3 and 0.8 seconds. This suggests that a threshold of around one second should be sufficient for the fusion of normal multimodal speech and gesture inputs for tasks of the kind studied in this paper. Alternatively, the sequential time window could be discarded altogether, and only overlapping signals processed, with an anticipated false segmentation rate of around 1.6%.

Limitations of this study include the single application scenario with its relatively few tasks, and the fact that the photos could easily be distinguished using spoken language (i.e. they could be named uniquely). The constrained nature of the tasks did not produce any ambiguous constructions (e.g. pointing at a slide show folder and saying 'rotate'), although these could be anticipated to occur occasionally in a fully-blown multimodal application.

CONCLUSION

This paper has analysed data collected from a speech and manual gesture-based digital photo management application scenario, and found that for this application, the majority or 63% of tasks were completed using multimodal rather than unimodal input. Speech and gesture were used with approximately equal frequency over all 18 subjects, indicating that both modalities were equally appropriate for the application and that users found advantage in combining them. The majority or 95% of combined speech and gesture inputs were observed to overlap in time (more so than for combined speech and pen inputs), indicating that the temporal fusion of combined speech and gesture input may be relatively simpler to process than speech and pen. The very high average consistency (98%) of simultaneous and sequential integration patterns for individuals found in this work adds to a growing body of empirical research that points to the need for user-adaptive temporal thresholds for multimodal fusion, both to accommodate the individual preferences of different users and to exploit the very high consistency in their integration pattern over time (Oviatt *et al.*, 2003).

Given the difference between speech and gesture and speech and pen integration patterns, future work could explore adaptation of temporal thresholds during multimodal processing to the specific modes used. An examination of integration patterns for more complex speech and gesture tasks is also envisaged.

ACKNOWLEDGMENTS

The authors would like to thank the other NICTA MMUI team members for helpful discussions, and numerous volunteers for participating in the data collection.

REFERENCES

Bolt, R. A. (1980) “Put-that-there”: Voice and gesture at the graphic interface, *Proceedings of Computer Graphics (SIGGRAPH’80)*, 14 (3), 262-270.

Cohen, P.R., Johnston, M., McGee, D.R., Oviatt, S.L., Pittman, J., Smith, I., Chen, L., and Clow, J. (1997) QuickSet: Multimodal interaction for distributed applications, *Proceedings of the Fifth International Multimedia Conference (Multimedia '97)*, ACM Press: Seattle, WA, November, 31-40.

Flippo, F., Krebs, A., and Marsic, I. (2003) A Framework for Rapid Development of Multimodal Interfaces, *Proceedings of the International Conference on Multimodal Interfaces*, 109-116.

Gupta, A. (2003) An adaptive approach to collecting multimodal input, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Krahnstoever, N., Kettebekov, S., Yeasin, M., and Sharma, R. (2002) A Real-Time Framework for Natural Multimodal Interaction with Large Screen Displays, *Proceedings of the International Conference on Multimodal Interfaces*, 119-124.

Oviatt, S., DeAngeli, A., and Kuhn, K. (1997) Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 415-422.

Oviatt, S., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M., and Carmichael, L. (2003) Toward a Theory of Organised Multimodal Integration Patterns during Human-Computer Interaction, in *Proceedings of the Fifth International Conference on Multimodal Interfaces*, 44-51.

Oviatt, S., Coulston, R., and Lunsford, R. (2004) When Do We Interact Multimodally ? Cognitive Load and Multimodal Communication Patterns, to appear in *Proceedings of the Sixth International Conference on Multimodal Interfaces*.

Schapira, E., and Sharma, R. (2001) Experimental evaluation of vision and speech based multimodal interfaces, *Proceedings of the 2001 Workshop on Perceptive User Interfaces*, 1-9.

Wilson, A., and Oliver, N., (2003) GWindows: Robust Stereo Vision for Gesture-Based Control of Windows, *Proceedings of the International Conference on Multimodal Interfaces*, 211-218.

COPYRIGHT