How Pairs Interact Over a Multimodal Digital Table

Edward Tse[1,2], Chia Shen[2], Saul Greenberg[1], Clifton Forlines[2]
[1]University of Calgary, [2]Mitsubishi Electric Research Laboratories
[1]2500 University Dr. N.W., Calgary, Alberta, Canada, T2N 1N4
[2]201 Broadway, Cambridge, Massachusetts, USA, 02139
[1](403) 210-9502, [2](617) 621-7500
[tsee, saul]@cpsc.ucalgary.ca, [shen, forlines]@merl.com

ABSTRACT

Co-located collaborators often work over physical tabletops using combinations of expressive hand gestures and verbal utterances. This paper provides the first observations of how pairs of people communicated and interacted in a multimodal digital table environment built atop existing single user applications. We contribute to the understanding of these environments in two ways. First, we saw that speech and gesture commands served double duty as both commands to the computer, and as implicit communication to others. Second, in spite of limitations imposed by the underlying single-user application, people were able to work together simultaneously, and they performed interleaving acts: the graceful mixing of inter-person speech and gesture actions as commands to the system. This work contributes to the intricate understanding of multi-user multimodal digital table interaction.

Author Keywords

Digital Tables, Multimodal Interaction, Speech, Gestures

ACM Classification Keywords

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

INTRODUCTION

Previous research explored how people could interact over existing single user applications (e.g., Blizzard's Warcraft III, Maxis's The Sims, and Google Earth) displayed on a digital table that recognized both speech and expressive hand gestures [10,11]. They listed a number of behavioural foundations motivating this multi-user, multimodal interaction. In particular, they hypothesised that one person's speech and hand gestures used to *command* the application also produced *consequential communication* that others could leverage as cues for validation and assistance. While previous ethnographic studies and empirical investigations indicated that consequential communication occurs regularly in real world situations, e.g., where people interact over physical artefacts such as paper maps [1], we do not know if these behavioural benefits accrue to speech and gesture commands directed to a digital system. To answer this question, we performed an observational study investigating how people used two multi-user speech and gesture wrappers built over existing single user applications. As we will see, our analysis verifies and adds detail to the role that speech and gesture commands play as consequential communication.

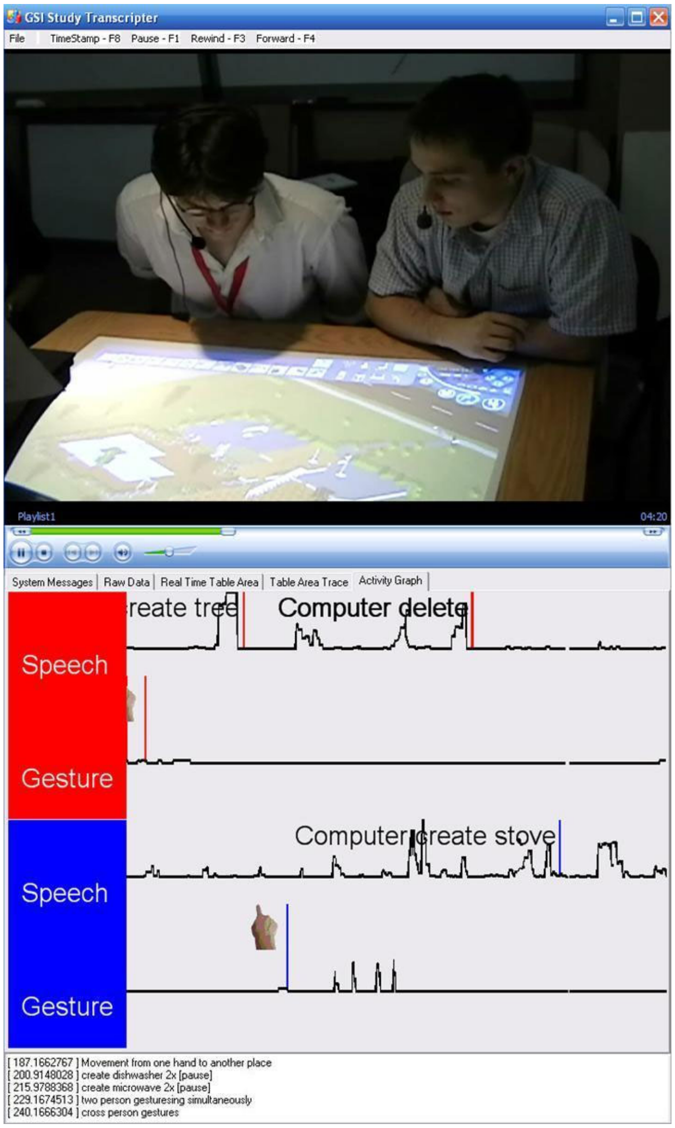
OBSERVATIONAL STUDY DESIGN

We observed 6 computer-proficient participants (3 pairs): 5 males and 1 female, ages 21-30 years. Pairs were seated side by side along the front edge of the digital table displaying an 'upright' single user application (Figure 1, top). Participants interacted with the application using speech via noise cancelling headsets and gestures via a DT107 MERL DiamondTouch table. Speech and gesture input was mapped to GUI commands understood by the application in a manner similar to [10,11]. The speech recognition engine distinguished speech commands from conversational speech by listening for a 'Computer' prefix (e.g. "*Computer*, create phone"). Participants used gestures as commands by directly touching the table surface. Feedback of successful speech and gesture recognition was indicated by the application's visual response and by an audio tone for speech commands. Spoken commands were designed to be easily understood by both the computer and other collaborators (e.g., "fly to [city]"). A printed list of recognizable speech and gesture commands was posted in front of participants, and pairs were encouraged to practice speech and gesture input prior to each trial. Tasks consisted of two scenarios, described below.

Travel Planning: Pairs used Google Earth to plan a European student's three day, all expenses paid, trip to Boston, New York and Chicago. Typical speech commands were "fly to [city]" and "layer [name e.g., roads]", while gestures included using one finger to pan or annotate, two fingers to zoom the camera in and out, and five fingers to tilt the camera. Pairs had to select four or five key places to visit in each city by using the "scratch pad" speech command, circling the area of interest and numbering the attractions in the order they would be visited.

Home Layout: Pairs used The Sims by Maxis to lay out furniture in a bed room, living room, kitchen, and washroom of a newly purchased two story home for a four person family. Typical speech commands were "create

CHI 2007, April 28–May 3, 2007, San Jose, California, USA.
Copyright 2007 ACM 978-1-59593-593-9/07/0004...$5.00.



```
[ 187.1662767 ] Movement from one hand to another place
[ 200.9148028 ] create dishwasher 2x [pause]
[ 215.9788368 ] create microwave 2x [pause]
[ 229.1674513 ] two person gesturesing simultaneously
[ 240.1666304 ] cross person gestures
```

Figure 1. The Study Transcription Application

[object]", "[first/second] floor", "walls [down/up]", while gesture commands included two finger pan, five finger object pick up, and one fist object stamping.

Video and Data Collection

While systems exist that log single person multimodal interactions (e.g., STAMP [3]), we needed a way to capture the interactions of *multiple* people with our multi-user multimodal system. First, we video recorded sessions and took field notes during the experiment. Second, we created a transcription tool that recorded synchronized streams of gesture and speech acts from both participants as seen by the system for playback with recorded video. Figure 1 shows a screen snapshot of our logging tool. The top shows the video. This is synchronized with the middle activity graph: a visualization of both participants' speech and gesture actions and how they were recognized by the system. The bottom pane includes manual transcription notes. For example, Figure 1 is a sequence in time where the left user said "computer create tree" after which the right user specified the location of the tree with a single finger. Other views allow the experimenter to replay recorded gestures over a 2D bird's-eye-view of the table, view aggregated statistics of the experimental session, or simply show the raw data.

COMMANDS AS IMPLICIT COMMUNICATION

We recorded and then transcribed a total of 476 minutes of speech and gesture actions from each participant, as recognized by the system at 15 events per second. Over all pairs, we coded 416 commands: 164 speech, 194 gesture and 58 multimodal commands (i.e., where speech and gesture together form the command). Our open coding revealed five categories of speech and gesture commands:

Assistance*:* People invoke commands as actions that directly respond to other people's explicit or implicit requests for help, for example,

R: There's… [zooms in] Fenway Park. Okay, how do you…?
L: Computer scratch pad (successful recognition)

Validation*:* A person's use of a command validates joint understanding and agreement reached in prior conversation,

L: And here [points], maybe a kind of small living room?
R: Computer create couch (successful recognition)

Affirmation*:* A person's command triggers an explicit follow-up agreement about the action or an implicit agreement when both continue with the task at hand,

R: Computer create couch (successful recognition)
L: Yeah [single finger placement in living room] good.

Clarification*:* A person's command is followed up by the other person's indication of confusion or a request for clarifying the meaning of the action,

L: Computer fly to Boston (successful recognition)
R: Huh? What did you just do?

Redundancy*:* A person explicitly mentions the action both in conversation and as a command, i.e., saying the command is redundant,

L: Yeah, let's go, kitchen is basically... oh trash can, computer create trash can (successful recognition)

Assistance, validation and affirmation are all examples of commands that are positively included as conversational elements. Clarification requests indicate that the command did not fit well as a conversational construct, while redundancy is an indication that a person viewed the action as distinct communication and command elements.

Figure 2 shows the average breakdown of the 416 coded speech and gesture command used across both tasks (for this figure, the 58 multimodal commands are split into their speech and gesture components). We coded 264 (64%) as affirmation, 73 (18%) as validation, 68 (16%) as assistance, 11 (2%) as redundancy, and 0 as clarification. When these numbers are considered by task (Figure 2 left vs. right), we saw that the Travel Planning Google Earth task had slightly higher validation and assistance rates than the Home Layout Sims task. We believe this is because many Google Earth commands performed global actions that would affect the entire work area, and for this reason, participants would converse with their partners before issuing the command.

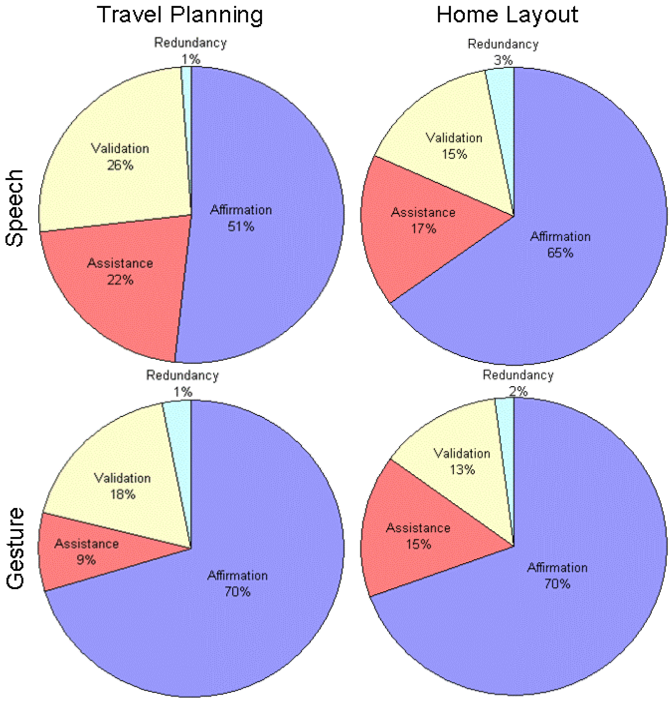


Figure 2. Speech / Gestures as Communicative Categories

Our coding results clearly show that speech and gesture commands directed to the system also served double duty as communication to other collaborators. 98% of our 416 observations were coded as assistance, validation, or affirmation. Only 2% - the clarification and redundancy categories – indicated commands that were not included well within the conversational context. Our own subjective appraisal of pair interactions confirms what these numbers suggest: people integrated speech and gesture commands into their joint conversations and actions.

To explain our results, Clark [2] describes how speech acts can be broken up into two tracks: track one describes the business of the conversation and, track two describes the efforts made to improve communication. With commands, track one becomes the act of issuing a command to the computer, while track two serves a communication role to other collaborators. We deliberately crafted speech commands so they were both machine and human recognizable (e.g., fly to Boston *vs.* reposition 135436). Our results suggest that pairs' used speech commands as dual purpose speech acts that fit into both tracks.

Similarly, consequential communication happens when one monitors the bodies of other's moving around the work surface [5, 6]. For example, as one person moves her hand in a grasping posture towards an object, others can infer where her hand is heading and what she likely plans to do. In our system, gesture commands are designed so that they provide consequential communication to others when used. For example, using five fingers to pick up a digital couch also produces awareness to collaborators around the table.

SIMULTANEOUS ACTIVITY AND INTERLEAVING ACTS

We now consider how people interact in this multimodal tabletop setting. We were particularly interested in whether the single user nature of the underlying application (i.e., where multi-user input is multiplexed into a single input stream) forced a situation in which people predominantly worked sequentially (e.g., by gross turn taking), or whether they were able to converse and interact simultaneously over this surface.

Simultaneous activity

First, we used our logger to mark each person's gesture and speech actions as either on or off: speech is on when it is above a volume threshold, while a gesture is on whenever the logger detects a finger or hand posture placed on the table. Thus for any instant in time we can determine if a simultaneous speech and gesture act is occurring. We then examined those times when at least one person was speaking and/or gesturing (53% of the time). For about 14% of this 53%, we found that the other person was also speaking and/or gesturing at the same time. i.e., they were interacting concurrently with each other. This number actually underestimates simultaneous activity, as it only includes those gestures which are direct touches to the table. In actual practice, we saw many gestures occur immediately above and around the table, as well as nodding and many other forms of body language. We observed (both during the experiment and from a review of the video recordings) that participants were highly engaged in each other's task and actions; it was rare to find a participant idling. They were involved both in how they attended to each other, and in the interleaving of their speech and gestures when talking about what they were doing. This supports other people's findings of simultaneous interaction over tables [7,9].

Interleaving acts

Next, we examined how people worked together during those episodes in which we saw at least one person direct speech and gesture commands to the application. Here, we analyzed our video transcriptions using an open coding method (e.g., [7]) to look for different styles of interleaving actions. Our analysis revealed that even though the underlying application could not recognize simultaneous activity, people managed to cooperate through *interleaving acts:* a graceful mixing of people's speech and gesture actions in the construction of commands. We saw four different interleaving interactions that can be described along the dimensions of coupling [8] and the input modality used.

Tightly Coupled, Inter-Modal. This category occurs when one person issues the speech component of a command and the other issues the gesture component. For example, the following interaction separates one's decision of creating a chair from the specification of the location for it.

L: Computer create chair (successful recognition)
R: [points to location to tell the system where the chair is to be created]

Tightly Coupled, Intra-Modal. One person discusses or gestures over what should be done while the other person performs the command on the system. These interleaving acts were primarily used for two purposes.

First, people used them to support coaching, validation and assistance. By suggesting what command should be performed next, participants are implicitly seeking validation of their suggestion from their partners. Second, we saw this mode used for cooperative error correction. In particular, when a person was having problems getting the system to recognize a particular speech or gesture command as valid input, the other person would often provide support by issuing the same command on their behalf.

To digress momentarily, cooperative error correction within this mode is extremely important: it provides an additional level of robustness to multimodal systems. Previous empirical studies described how multimodal systems can add robustness; each modality provides a check for erroneous recognition [4]. For example, a "create stove" speech command would be ignored by our system if no location-indicating gesture followed. Cross person error correction adds further robustness over this system correction. To illustrate, we noted 84 speech recognition errors in our transcriptions where the system failed to correctly recognize a speech command. Of these, partners stepped in ~1/3 of the time to correct another's error. Most participants would start by trying to reissue the command themselves. Two or more failed speech recognition attempts might be seen as an implicit request for assistance according to Clark's [2] description of track two efforts to improve communication, and repair conversation.

Loosely Coupled, Inter-Modal. One person issues the next speech command while the other is finishing their gesture, i.e., they overlap command sequences, which the system then queues to the underlying single user application. This allowed pairs to efficiently issue overlapping multimodal commands without having to wait for the other person to finish their action. We noticed that each participant peripherally monitored the workspace to find an appropriate place to insert their next command; they rarely overlapped commands in ways that resulted in system confusion.

Loosely Coupled, Intra-Modal. One person issues a speech or gesture command within a conversation to assert informal floor control of not only the application, but of the conversational direction. For example in the travel planning task, people would often assert control of the map to signal that it was their turn to speak or to advance the discussion in a new direction. The other person would follow this lead.

In summary, we were pleasantly surprised that people were able to converse and communicate using simultaneous speech and gesture commands, much as they do in real world interactions when working over work surfaces. Similarly, people were able to do fine-grained mixing of their actions, conversation, and commands using what we called interleaving acts.

CONCLUSION

Of course, there is much left to do. Our study is small; larger studies are needed to confirm our numbers and to investigate additional details. Another obvious next step is to build multimodal tables running true multi-user applications, and to see if their use differs from what we saw here. Limitations aside, this paper contributes to the understanding of multi-user multimodal interactive systems. We saw that speech and gesture commands directed to the computer also serve double duty as implicit communication to others. We saw that people's simultaneous interactions were not inhibited by the underlying single-user application. Similarly, we saw that people were able to compose sequential actions through interleaving acts: the graceful mixing of both participant's speech and gesture actions as commands were being constructed. All these are positive. They suggest that people can use multi-user multimodal tabletops - even when limited by single user application constraints – in much the same way as they work over visual work surfaces.

REFERENCES

1. Cohen, P.R., Coulston, R. and Krout, K., (2002), Multimodal interaction during multiparty dialogues: Initial results. Proc. IEEE ICMI 2002, 448-452.
2. Clark, H. *Using language.* Cambridge Univ. Press, 1996
3. Clow, J. and Oviatt, S. (1998) STAMP: A suite of tools for analyzing multimodal system processing, *Proc. Int. Conf. Spoken Language Processing*, 1998.
4. Oviatt, S. L. Ten myths of multimodal interaction, *Comm. ACM*, 42(11), 1999, 74-81.
5. Pinelle, D., Gutwin, C. and Greenberg, S. Task analysis for groupware usability evaluation: Modeling shared-workspace tasks with the mechanics of collaboration. *ACM TOCHI*, 10(4), 2003, 281-311.
6. Segal, L. Effects of checklist interface on non-verbal crew communications, NASA Ames Research Center, Contractor Report 177639. 1994
7. Scott, S.D., Carpendale, M.S.T, & Inkpen, K.M. (2004). Territoriality in Collaborative Tabletop Workspaces. *Proc. CSCW 2004*, ACM Press, 294-303.
8. Tang, A., Tory, M., Po, B., Neumann, P., and Carpendale, M. S. T. (2006). Collaborative Coupling over Tabletop Displays. *Proc. CHI 2006*, ACM Press, 1181-1190.
9. Tang, J. (1991) Findings from Observational Studies of Collaborative Work, *IJHCS*, 34(2), 143-160.
10. Tse, E., Shen, C., Greenberg, S. and Forlines, C. (2006) Enabling Interaction with Single User Applications through Speech and Gestures on a Multi-User Tabletop. *Proc. AVI 2006*, ACM Press, 336-343.
11. Tse, E., Greenberg, S., Shen, C. (2006) GSI Demo: Multiuser Gesture / Speech Interaction over Digital Tables by Wrapping Single User Applications, *Proc. ICMI 2006*, ACM Press.