Usability of MultiPoint

Anoop K. Sinha
CS Division, EECS Department
University of California at Berkeley
Berkeley, CA 94720-1776 USA
+1 510 642 3437
aks@cs.berkeley.edu • http://guir.berkeley.edu

ABSTRACT

MultiPoint is a speech and pen user interface for building presentations, implemented as an add-on to Microsoft PowerPoint^{TM}. We compared users' satisfaction and performance between building presentations with MultiPoint and with PowerPoint. We also compared participants' performance and reactions between using Wizard of Oz (WOz) speech recognition and computer speech recognition.

In aggregate, most of the participants ranked the naturalness of MultiPoint speech commands high regardless of whether they were using WOz recognition or computer speech recognition. Six participants with WOz recognition completed tasks in about the same time and with about the same number of errors and with the same satisfaction, as they did using PowerPoint alone. Six participants with computer speech recognition took twice the time and committed four times as many errors and had significantly lower satisfaction, as they did using PowerPoint.

INTRODUCTION

Pen and speech interfaces are likely to become more important in the future with increasing use of tablet computers and other handheld devices. In an attempt to expand on the tasks where multimodal input might be more appropriate [COHEN, OVIATT], we undertook the task of building a multimodal interface for building presentations, a common desktop task, and compared it to a regular graphical user interface with keyboard and mouse.

Our interviews with professionals in industry led us to the conclusion that building presentations is a task potentially well-suited to a multimodal interface on a tablet computer. Out interviewees often draw the first draft of their slides; we designed MultiPoint as a tool for creating the first draft of a set of slides on a pen computer.

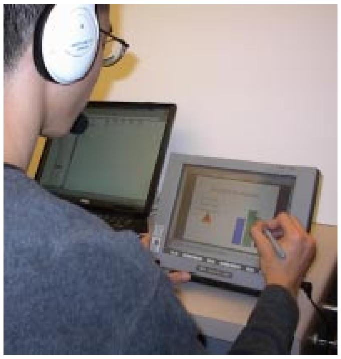


Figure 2. A participant using MultiPoint on a tablet computer

MULTIPOINT IMPLEMENTATION

Multipoint is implemented as an add-on to Microsoft PowerPoint and uses SRI's Open Agent Architecture (OAA) [MORAN]. It communicates to PowerPoint via Visual Basic for Applications (VBA). OAA facilitates synchronous and asynchronous communication managed by a central facilitator among different recognition agents in the system. This provides the facilities for pen and speech recognition in MultiPoint.

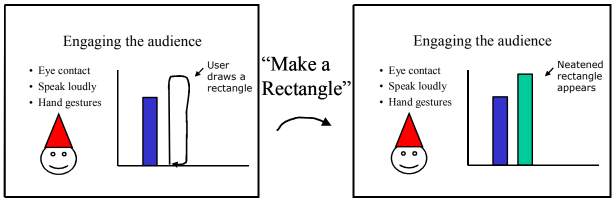


Figure 1. The MultiPoint command "make a rectangle" converts a sketched rectangle into neatened form

MultiPoint includes a subset of PowerPoint functionality; there are approximately 60 multimodal commands that can be used to draw and dictate slides. A user starts with the freeform drawing tool and draws a word or a shape on the slide. He simultaneously speaks speech commands to act on the freeform object.

As an example, the participant draws a rectangular shape, says, "make a rectangle," and the sketch is converted into recognized form (Figure 1). MultiPoint confirms the command audibly by repeating the recognized command via text to speech.

Speech commands in MultiPoint are particularly useful for

1. guiding recognition of sketched items
2. changing properties of existing objects
3. setting animations of existing objects

Some example commands used in the user study include:

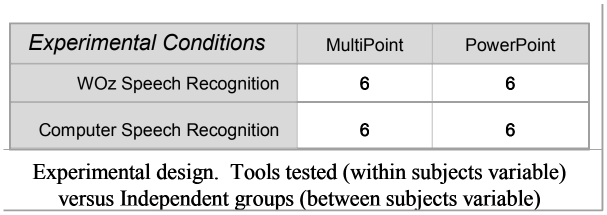
- "add title"
- "add bullet"
- "make a rectangle"
- "make a triangle"
- "color shape red"
- "fly from right"
- "delete"
- "undo"

By explicitly switching into "dictation mode" a participant can also dictate text. Text can also be inputted via a handwriting recognizer agent.

METHOD

In the user study, 12 paid student participants were asked to copy four slides using MultiPoint and PowerPoint (Figure 4). For MultiPoint, each used a Fujitsu Stylistic 4000 pen computer and a headset microphone (Figure 2). For PowerPoint each used a laptop computer and an external mouse. Participants were also allowed to create a freeform slide about their jobs after the copy tasks.

Each participant was trained on the MultiPoint commands and the PowerPoint operations that they could use to draw the slides. The copy tasks were presented in randomized order. One-half (six) performed their tasks with WOz speech recognition, and the other half (six) of the participants performed the tasks with computer speech recognition from IBM ViaVoice™.



Experimental Conditions	MultiPoint	PowerPoint
WOz Speech Recognition	6	6
Computer Speech Recognition	6	6

Experimental design. Tools tested (within subjects variable) versus Independent groups (between subjects variable)

Figure 3. A total of 12 participants were tested across two speech recognition conditions and two tools.

The User Interface community typically operationalizes "performance" as the time, steps, and errors for completion of certain tasks. Each task was timed, and the number of steps and errors were counted. Steps included "adding a square," "changing the color of a triangle," and were counted at the same granularity in both MultiPoint and PowerPoint. Errors included mistakes, undos, and speech misrecognitions.

Participants were surveyed about their PowerPoint experience before the study and about their evaluation of the system after the study.

RESULTS

Time, steps, and errors were the three metrics most important to us in the comparison of MultiPoint and PowerPoint. Satisfaction values from the survey were also important.

Within Subject Analysis

A set of error bar plots with 95% confidence intervals (Figure 5, Figure 6, Figure 7) show the mean value of these metrics for each of the four tasks in the two groups of participants, those who ran MultiPoint with Wizard of Oz speech recognition and those who ran MultiPoint with PowerPoint speech recognition. The Wizard of Oz speech recognition cases generally overlap with PowerPoint, suggesting a lack of statistically significant differences between the two. The Computer speech recognition cases exceed the PowerPoint values by some amount for some of the tasks.

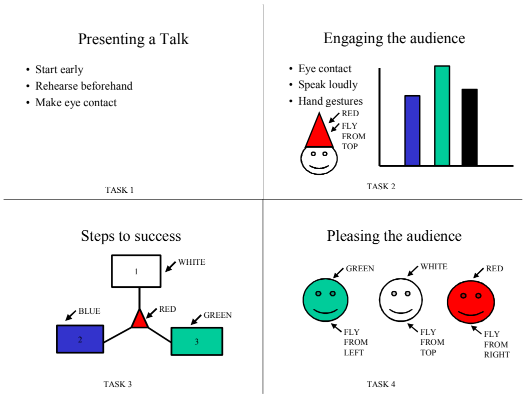
Rather than compare each task individually, we totaled values and performed analysis on the sum of the time, sum of the steps, and sum of the errors for each of participants. In what follows, we consider the MultiPoint and PowerPoint performance Wizard of Oz speech recognition group and the Computer Speech recognition group separately. This allows us to do a within subjects analysis of the performance and satisfaction metrics that were collected.

In the WOz speech recognition tests, there were no statistically significant differences in the total time (*MD*=21, *SD*=210), *t*(5)=0.243, *p*=0.818 (Figure 8) or the total steps for completion (*MD*=-2.2, *SD*=31), *t*(5)=-0.173, *p*=0.870 (Figure 9) or the total errors (*MD*=-8.2, *SD*=12), *t*(5)=-1.70, *p*=0.149 (Figure 10) between MultiPoint and PowerPoint. (There were no speech recognition errors in the WOz SR condition.) This result means that MultiPoint performs as well as PowerPoint, even for participants experienced with PowerPoint and new to MultiPoint.

For computer speech recognition, there were statistically significant differences in total time, total steps, and total errors. MultiPoint users took 215% of the time (*M*=409, *SD*=163), *t*(5)=6.12, *p*=0.0017 (Figure 8), 148% of the steps (*M*=33, *SD*=13), *t*(5)=6.11, *p*=0.0017 (Figure 9), and 427% of the errors (*M*=27, *SD*=13), *t*(5)=5.11, *p*=0.0037 (Figure 10), versus PowerPoint. From observation during the test, it was clear that most of the additional MultiPoint steps versus PowerPoint steps were from speech recognition errors and correcting those errors. Reducing the error rate might make performance more similar to WOz recognition.

For WOz speech recognition participants, there were no statistically significant differences between MultiPoint and PowerPoint in participant rankings, scaled 1-5, of ease of use (*MD*=-0.33, *SD*=0.82), *t*(5)=-1, *p*=0.36 (Figure 11), quickness (*MD*=-0.33, *SD*=1.63), *t*(5)=-0.5, *p*=0.638 (Figure 12), and naturalness (*MD*=0.17, *SD*=1.33), *t*(5)=0.307, *p*=0.771 (Figure 13). All three beginners to PowerPoint ranked MultiPoint higher for ease of use, noting specifically its simplicity and pleasantness.

Participants using the computer speech recognizer were



Presenting a Talk

- Start early
- Rehearse beforehand
- Make eye contact

Engaging the audience

- Eye contact
- Speak loudly
- Hand gestures

Steps to success

Pleasing the audience

Figure 4. Participants' slide copy tasks.

uniformly disappointed with the speech recognition performance, with a statistically significant ranking of the speech recognizer versus those who ranked the WOz recognition (*MD*=1.83, *SE*=0.31), *t*(10)=5.97, *p*=0.000138 (Figure 14). Those who used the computer speech recognizer perceived themselves as making many errors versus PowerPoint (*MD*=1.67, *SD*=0.816), *t*(5)=5, *p*=0.00410 (Figure 15) and waiting long times for commands to execute, though there was no statistically significant difference in the time per command versus PowerPoint (*MD*=0.167, *SD*=0.753), *t*(5)=0.542, *p*=0.611 (Figure 16). There was approximately one speech recognition error per four commands for computer speech recognition participants.

Participants in the Computer Speech Recognition category had to make significant adjustments as they were using the computer system. They needed to speak each command in a full breath into the speech recognizer; they needed to properly segment their commands with silence in between; they could not make extraneous side comments without the speech recognizer picking them up. The speech recognizer clearly interfered with the flow of completing tasks. Participants did not like repeating commands, and did not like waiting for the recognition result and confirmation of the command.

Error correction rarely needed to be used in the WOz case, in which there were no speech recognition errors. WOz recognition interfered much less with the flow of the task completion.

The frustration level between the participants in these two groups was extremely different. Computer speech recognition participants displayed a wide set of nonverbal and verbal cues of frustration. One participant worked very fast, talking in intolerant tones. Another sighed and raised eyebrows in frustration at the system. Many participants made side comments, which themselves sometimes interfered with the performance of the speech recognizer.

Between Subject Analysis

From the design of this experiment, it is also possible for us to perform a set of between subjects analysis between the Wizard of Oz speech recognition group and the Computer speech recognition group.

The data suggests that there are no statistically significant results between WOz speech recognition performance and Computer speech recognition performance for total time (MD=-101, SE=122), t(10)=-0.832, p=0.424 and total steps (MD=-19.5, SE=10.3), t(10)=-1.90, p=0.0868. It does suggest that there are differences for errors (MD=-27.7 7, SE=6.10), t(10)=-4.53, p=0.00109. Clearly the performance of the computer speech recognizer is the major contributor to the differences in errors.

Further analysis reveals that there were actually statistically significant differences between the PowerPoint performance in the two groups for total time (MD=287, SE=61.6), t(10)=4.66, p=0.000898. The differences for

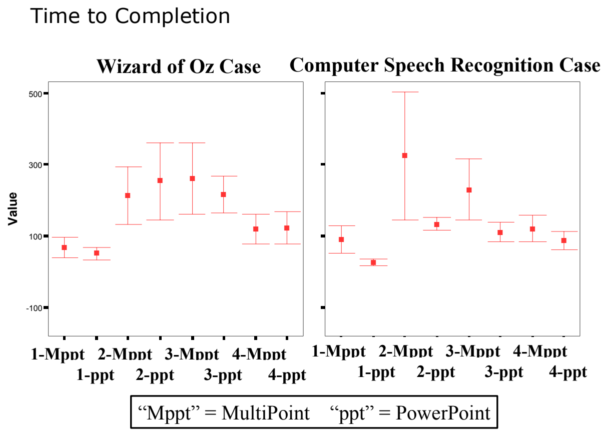


Figure 5. Time to completion with 95% confidence intervals for each of the four tasks across the two groups of participants, showing not significant differences for the WOz case and significant differences for Computer Speech Recognition.

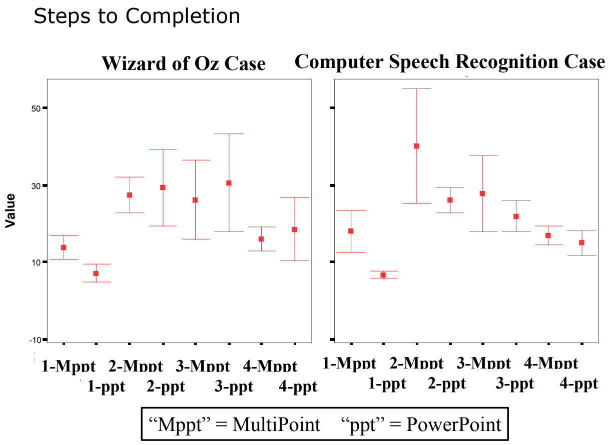


Figure 6. Steps for completion with 95% confidence intervals for each of the four tasks across the two groups of participants, showing not significant differences for the WOz case and significant differences for Computer Speech Recognition.

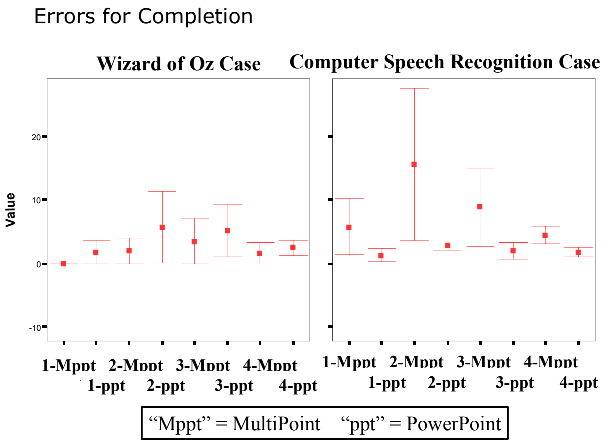


Figure 7. Errors for completion with 95% confidence intervals for each of the four tasks across the two groups of participants, showing not significant differences for the WOz case and significant differences for Computer Speech Recognition.

total steps (MD=16, SE=10.9), t(10)=1.46, p=0.173 and for total errors (MD=7.17, SE=3.44), t(10)=2.09, p=0.064, are not statistically significant.

This analysis suggests that there are some underlying differences between the Wizard of Oz Case group and the Computer speech recognition group. A set of Repeated Measures ANOVA's, varying the tool as the within subjects factor and the type of recognition as the between subjects factor further confirms this, with much more of the variance explained by the difference of within subjects factors and much less by the difference of between subjects factors.

Sum of Time Repeated Measures ANOVA
on Sum of Time with Tool as Within-Subjects Factor and Recognition Type as Between-Subjects Factor
- Both Factors: Pillai's Trace=0.560, F(1,10)=12.7,p=0.005, Total Variance Explained 56%
- Within-Subjects F(1,10)=12.731, p=0.005, Variance Explained 56%
- Between-Subjects F(1,10)=1.332, p=0.275, Variance Explained 12%
Sum of Step Repeated Measures ANOVA
on Sum of Steps with Tool as Within-Subjects and Recognition Type as Between-Subjects Factor
- Both Factors: Pillai's Trace=0.43, F(1,10)=6.75, p=0.027, Total Variance Explained 40%
- Within Subjects F(1,10)=6.746, p=0.027, Variance Explained 40%
- Between Subjects F(1,10)=0.047, p=0.833, Variance Explained 0.5%
Sum of Errors Repeated Measures ANOVA
on Sum of Errors with Tool as Within-Subjects and Recognition Type as Between-Subjects Factor.
- Both Factors: Pillai's Trace=0.707, F(1,10)=24.1, p=0.001, Total Variance Explained 71%
- Within-Subjects F(1,10)=24.168, p=0.001, Variance Explained 71%
- Between Subjects F(1,10)=8.770, p=0.014, Variance Explained 47%

Table 1. Results of Repeated Measures ANOVA's on tool and recognition type as contributing factors to the variance for the performance metrics, time, steps, errors, across all four tasks.

Correlational Analysis

We performed a set of correlations between the time, steps, and errors data and the responses to satisfaction in the post-survey. One results that came out of this analysis is that participants in the WOz case rated the speech recognition performance higher if they finished faster and lower if they finished slower (*Pearson Correlation*=0.821, *p*=0.045). Ease of Use ranking was directly related to the ranking of the likelihood that the participant would use MultiPoint to create the first draft of a presentation across both WOz case (*Pearson Correlation*=1.00, *p*=0.000) and Computer speech recognition case (*Pearson Correlation*=0.857, *p*=029).

Few of the other correlations were statistically significant, and thus it is difficult to make additional claims about the link between the performance with the tool and the satisfaction rankings.

Additional General Observations

All participants initially referred to the cheat sheet provided which listed the available speech commands. Half of the participants did not initially switch modes when starting with the speech commands. It was clear that moding was something that they had to learn. But they soon adapted, and learned the repetitive pattern of "dictation mode," "…," "command mode." Switching modes was generally disliked by the participants, but most did not have trouble using it. A few participants asked if specific commands were available, such as "cut" and "paste." The commands that they asked for can form the basis of commands to add to MultiPoint in the future.

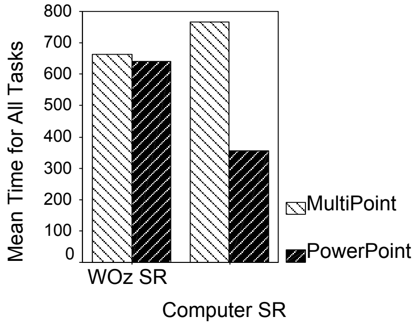


Figure 8. Time for completion in the two groups of participants

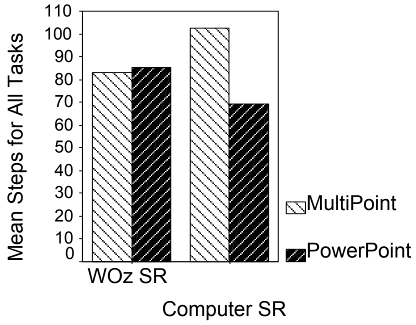


Figure 9. Steps for completion in the two groups of participants

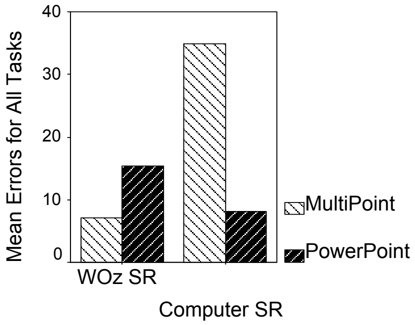


Figure 10. Errors for completion in the two groups of participants

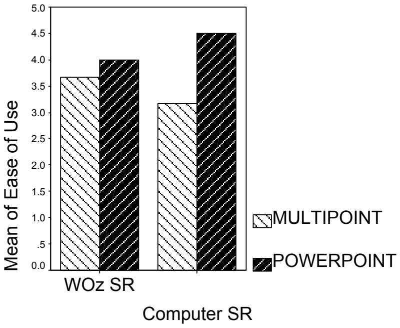


Figure 11. Ease of use rankings in the two groups of participants

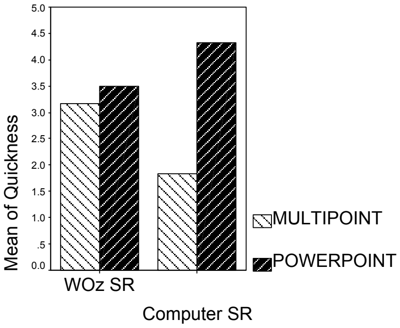


Figure 12. Quickness rankings in the two groups of participants

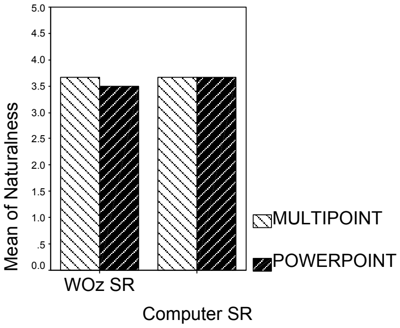


Figure 13. Naturalness rankings in the two groups of participants

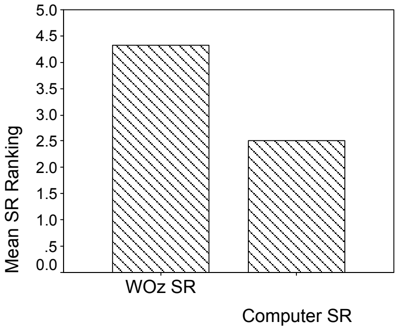


Figure 14. Speech Recognizer rankings in the two groups of participants

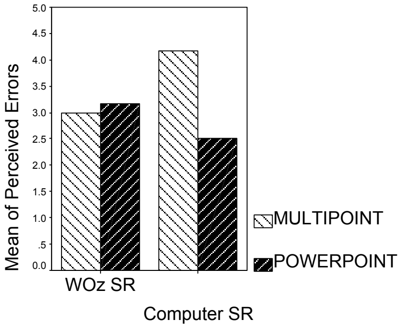


Figure 15. Perceived errors in the two groups of participants

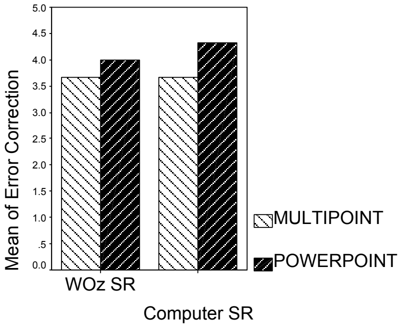


Figure 16. Rank of ease of error correction in the two groups

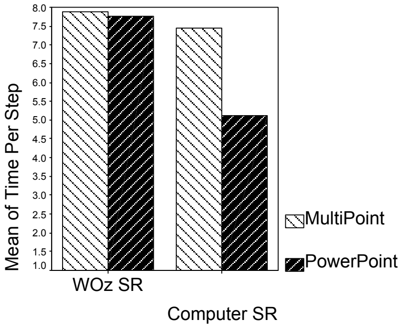
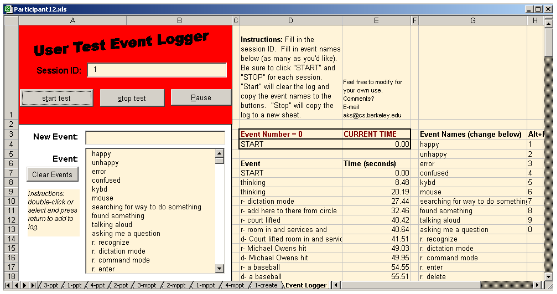


Figure 17. Time per step for each of the two groups of participants

Repeated Patterns and Criticial Incidents

In addition to summary performance and survey data, we collected a set of timing and behavioral data for each participant accomplishing each task. This data was collected in the User Test Event Logger which allows the test administrator to transcribe the recorded speech commands as well as input in behavior data (Figure 18).



User Test Event Logger

Session ID: 1

Instructions: Fill in the session ID. Fill in event names below (as many as you'd like). Be sure to click "START" and "STOP" for each session. "Start" will clear the log and copy the event names to the buttons. "Stop" will copy the log to a new sheet.

Feel free to modify for your own use. Comments? E-mail aks@cs.berkeley.edu

New Event:

Event: happy, unhappy, error, confused, kybd, mouse, searching for way to do something, found something, talking aloud, asking me a question, r: recognize, r: dictation mode, r: command mode, r: enter

Instructions: double-click or select and press return to add to log.

	D	E	G	H
3	Event Number = 0	CURRENT TIME	Event Names (change below)	Alt+
4	START	0.00	happy	1
5			unhappy	2
6	Event	Time (seconds)	error	3
7	START	0.00	confused	4
8	thinking	8.48	kybd	5
9	thinking	20.19	mouse	6
10	r- dictation mode	27.44	searching for way to do somethin	7
11	r- add here to there from circle	32.46	found something	8
12	r- court lifted	40.42	talking aloud	9
13	r- room in and services and	40.64	asking me a question	0
14	d- Court lifted room in and servic	41.51	r: recognize	
15	r- Michael Owens hit	49.03	r: dictation mode	
16	d- Michael Owens hit	49.95	r: command mode	
17	r- a baseball	54.55	r: enter	
18	d- a baseball	55.51	r: delete	

Figure 18. The User Test Event Logger used in this study to collect the speech commands used as well as behavior data.

This data gives us some information for considering behavioral analysis, in particular the identification of critical incidents, which are represented in the graphs (Figure 19, Figure 20) as more steeply sloped points.

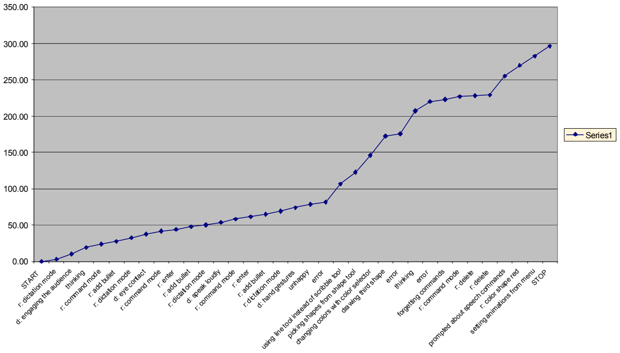


Figure 19. Behavior graph for Participant 2 on task 2. This participant was in the Wizard of Oz recognition group.

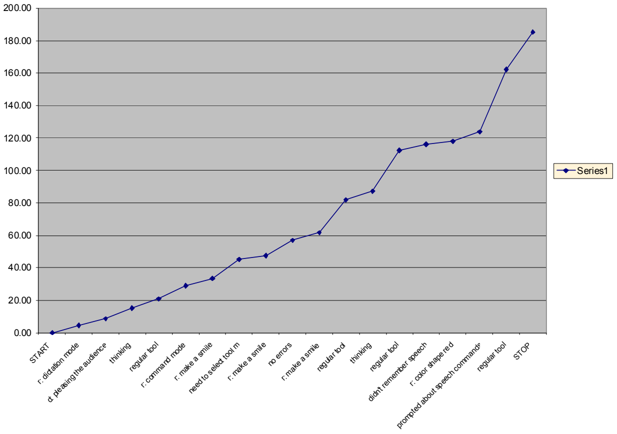


Figure 20. Behavior graph for Participant 2 on task 4. This participant was in the Wizard of Oz recognition group.

A common pattern seen in all of these graphs is a set of steadily upward sloping commands, representing the participant in a steady flow accomplishing the tasks. Places where the graph slopes upward are interruptions in the participant's flow. Most of these were caused by either problems with computer speech recognition performance, or confusion about the way to accomplish a certain tasks. The former can be remedied by improving the computer speech recognition, through a different style of interaction, such as press to speak. The confusion can be remedied by additional training on available commands. One would expect a steady slope without steep slopes among participants who use MultiPoint over longer periods of time.

CONCLUSIONS

MultiPoint with Wizard of Oz recognition performs about as well and ranks about the same in satisfaction as PowerPoint. This holds true even though all participants were new to MultiPoint. There is some chance that MultiPoint performance numbers will improve if participants are given more training or get more experience with it.

MultiPoint with computer speech recognition performs more poorly and leads to lower satisfaction than PowerPoint. Most of the downside can be explained by computer speech recognition errors. Improving computer speech recognition performance either through a different speech recognition system or a different interaction technique will likely improve MultiPoint's results.

Building presentations is indeed a viable task for multimodal interaction. However, novice computer users are more likely than expert computer users to be attracted to a multimodal interface for presentation building.

Recognition performance needs to be quite high to ensure good performance and user satisfaction in a multimodal application such as MultiPoint. Participants need to feel that the multimodal application that they are using is proceeding smoothly and quickly. They do not like waiting for recognition results.

REFERENCES

[COHEN]. Cohen, P. R., Johnston, M., McGee, D., Oviatt, S. L., Clow, J., Smith, I. "The Efficiency of Multimodal Interaction: A Case Study." *Proceedings of the International Conference on Spoken Language*, 1998.

[MORAN] Moran, D. B. and A. J. Cheyer, L. E. Julia, D. L. Martin, and S. Park, "Multimodal User Interfaces in the Open Agent Architecture," in *Proc. of the 1997 International Conference on Intelligent User Interfaces*, (Orlando, Florida), pp. 61--68, 1997.

[OVIATT] Oviatt, S., "Multimodal Interfaces for Dynamic Interactive Maps." *Proceedings of CHI'96*, pp. 95-102, 1996.