Speech “Re-Cognition”?
Investigating Speech as a Cognition-Augmenting Modality

Dan Loehr, Laurie Damianos, Lisa Harper, Carl Burke, Steve Hansen, Michael Viszmeg
The MITRE Corporation
{loehr | laurie | lisah | cburke | swh | mrv}@mitre.org

Abstract

We report on two projects investigating speech interfaces and augmented cognition.

First, we performed an exploratory study to examine the effects of speech-enabled input on an imagery analysis and annotation task. We hypothesized that speech recognition could be a cognition-enabling technology by reducing the cognitive load of instrument manipulation and freeing up cognitive resources for the task at hand. Quantitative results indicate that people did identify images more efficiently and could potentially annotate images faster with speech. However, people did not annotate better with speech (precision was lower, and recall was significantly lower).

Second, we developed a Personal Digital Assistant capable of controlling a search-and-rescue robot. The PDA displayed a map and live video feed of the robot’s camera, and permitted both stylus and speech input. We note differences between speech input in the robot control task and in the imagery annotation experiment. We conclude from both projects that speech is helpful if the cognitive cost of speech recognition delays and errors is outweighed by the cognitive benefit of reducing instrument manipulation.

1. Introduction

Speech has long been an enticing modality, promising to liberate the user from manual computer interfaces. The hope has been that mental effort could be freed from the task interface, allowing humans to concentrate on the task itself. The “recognition” in “speech recognition” refers, of course, to the computer’s ability to transcribe speech. But the word “recognition” is based on the word “cognition”, and although the two senses are unrelated in present usage, perhaps speech *recognition* can indeed benefit user *cognition*.

We report on two projects investigating this possibility. The first is a formal experiment testing a speech interface in an imagery annotation task. The second is a speech interface on a task controlling robots with a Personal Digital Assistant (PDA), in which we informally observed users. We then compare the two projects in terms of subjects’ usage of the speech interfaces.

2. Background and motivation

Military operators are often put into complex human-computer interactive environments that have been shown to fail in stressful situations. The DARPA Augmented Cognition program (2001) proposes to develop technology to enhance human performance using intrinsic capabilities (i.e., brain function) through scientific principles that have previously been inadequately exploited in human-computer system designs. The program will, among other goals, develop and implement strategies of multiple sensory inputs and mixed initiatives between human- and computer-generated interactions. One challenge is designing interfaces based on cognition.

Historically, graphical user interfaces are largely uni-modal, supporting vision. Multi-modal interfaces appear to be an intuitive way to tailor an interface to the user’s cognitive state, instead of forcing the user to adapt to the interface. Oviatt and Cohen (2000) state, "A profound shift is now occurring toward embracing users' natural behavior as the center of the human-computer interface. Multi-modal interfaces are being developed that permit our highly skilled and coordinated communicative behavior to control system interactions in a more transparent experience than ever before."

Multi-modality does not necessarily entail using speech. Why do we hypothesize that the addition of speech would be effective? Graphical user interfaces (GUIs) and direct manipulation interfaces (DMIs) historically have provided interactive environments resulting in increased user acceptance, helping the users concentrate on tasks as the systems become more transparent (Shneiderman, 1983). However, the systems can become truly transparent only if the interface allows for the hands-free, eyes-free interaction provided by speech (Grasso et al., 1998). In addition, GUIs and DMIs are limited in other ways, including support for identifying objects not visible and for identifying and manipulating large sets of objects (Cohen, 1992).

In many situations when conventional GUIs are neither feasible nor desirable, speech can be indispensable. Even when conventional interaction modes are possible, speech-enabled input can be supplementary (Rosenfeld et al., 2001). We can interact through speech while using other facilities (e.g., eyes and hands) since speech does not require focused attention (Rosenfeld et al., 2001). Speech can be used as a shortcut for long navigational paths, to facilitate selection in information-rich environments, and in "hands-busy" and "eyes-busy" situations (Grasso et al., 1998; Rosenfeld et al., 2001; Shneiderman, 2000). If two or more input modes provide parallel or duplicate functionality, users can alternate their use of input modes to reduce the likelihood of errors or resolve existing ones (Oviatt, 2000; Oviatt & Cohen, 2000).

There are arguments against speech being a cognitive-enabling technology. Grasso et al. (1998) note that, since speech is temporary, spoken information "can place extra memory burdens on the user and severely limit the ability to scan, review, and cross-reference information." Shneiderman (2000) argues that "Speech is slow for presenting information, is transient and therefore difficult to review or edit, and interferes significantly with other cognitive tasks". He claims that it is difficult to speak and solve problems at the same time since speaking uses cognitive resources. Speaking and listening are controlled by the same part of the brain that stores information and solves problems, but hand-eye coordination occurs elsewhere in the brain so people can easily type or use the mouse while solving a problem.

It may be that different types of mental processes (e.g. problem-solving) are inefficiently mixed with speaking, while other types are quite compatible. Cognitive research shows that spatial processes (which the MSIIA system requires) may be efficiently mixed with verbal output. Wickens and Hollands (1999) report, “Data from multiple-task studies indicate that spatial and verbal processes… whether functioning in perception, working memory, or response, depend on separate resources and that this separation can often be associated with the two cerebral hemispheres” (Polson & Friedman, 1988).

In summary, researchers claim that speech interfaces can be effective when used in conjunction with other modalities so that complementary strengths can emerge. These integrated modalities can result in the system interface becoming more transparent to the user. Finally, this transparency can allow the user to shift cognitive resources from the interface to the task at hand. We specifically test this claim in our experiment, described below.

3. The Imagery Annotation Experiment

This experiment examines the effects of speech-enabled input on the Multi-Source Intelligence Integration and Analysis (MSIIA) system (Hansen, 1997) in performing a simple imagery analysis and annotation task. The MSIIA system is an information fusion system that allows imagery analysts to view and annotate multiple streams of visual data for airborne surveillance and reconnaissance activities. We added speech to a component of the system for hands-free input of annotations. Our experiment was designed to test the following hypotheses:

- People can annotate images in video segments *faster* with the MSIIA augmented by speech.
- People can annotate images in video segments *better* with the MSIIA augmented by speech. (Better is defined as more target items annotated and target items more accurately annotated.)
- People *prefer* speech-enabled input to manual input when annotating video images in the MSIIA.

We designed a within-subjects, counterbalanced experiment in which eight participants were asked to identify and annotate images in two different video segments. We controlled one independent variable: input mode (i.e., manual input only versus manual input with the addition of speech-enabled input).

Each participant was tested under both system configurations. The order in which the conditions were used was switched from one participant to the next so as to counterbalance any confounding effects. Under each mode, the participants performed one training trial and one test trial. The training trials were used to familiarize the participants with the task as well as the input mode. Two different video segments were used for the two test trials, and for purposes of this experiment, we assumed that the video segments were approximately equivalent. To account for any slight differences in the video segments, we alternated the order in which the two segments were administered. There were four different treatment conditions (where order matters), as shown in Table 1.

Table 1. Experiment treatment conditions

	1st Test Trial		2nd Test Trial	
	Clip	**Input Mode**	**Clip**	**Input Mode**
1	**C**	Manual	**D**	+ Speech-enabled
2	**D**	Manual	**C**	+ Speech-enabled
3	**C**	+ Speech-enabled	**D**	Manual
4	**D**	+ Speech-enabled	**C**	Manual

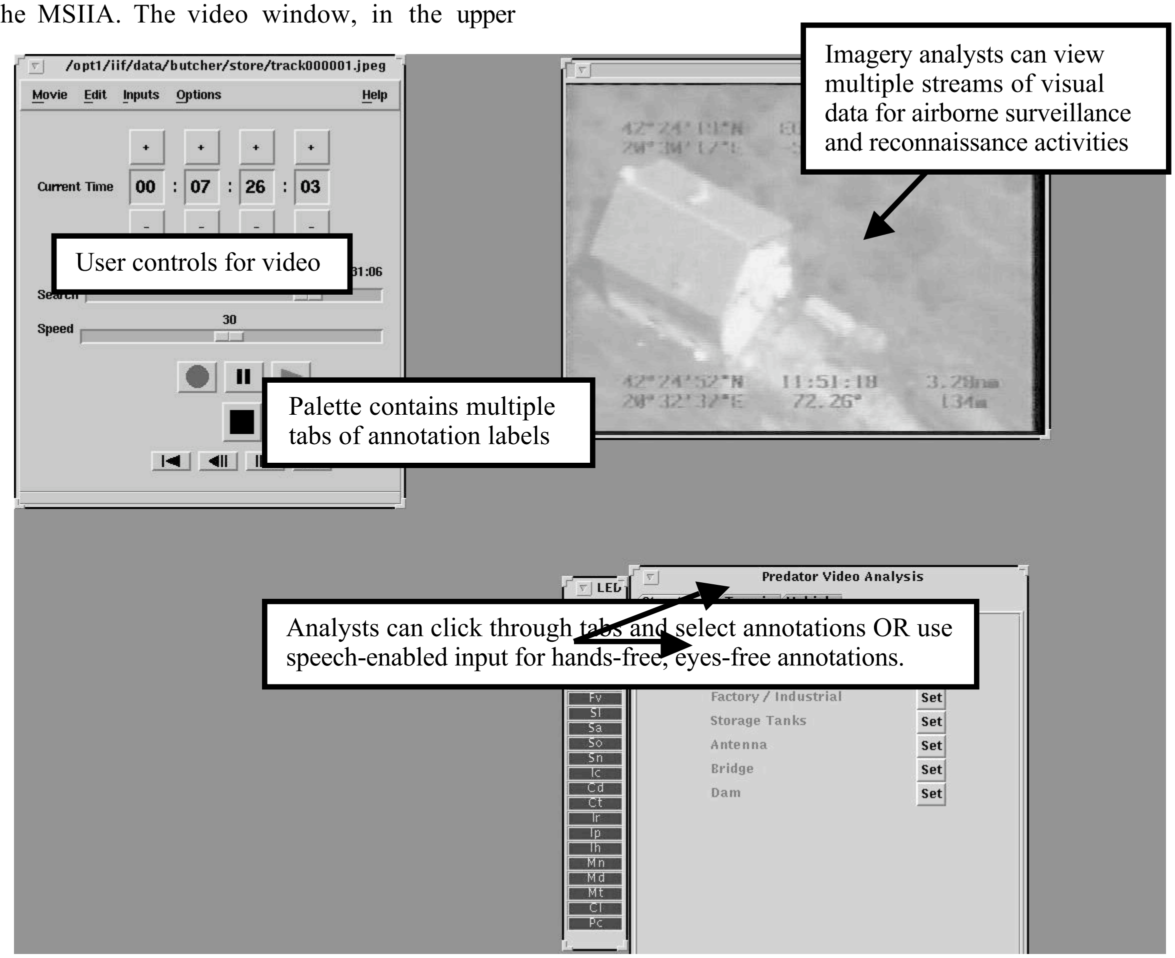
As this was meant to be an exploratory evaluation, we ran the experiment on a small sample size: two participants under each treatment condition. A series of pilot studies was run to help debug the training materials, questionnaires, task complexity, choice and length of video segments, and annotation categories.

Participants were asked to review two video clips and look for structures, terrain, and vehicles that might reveal the presence of military or the existence of a possible war zone. They were told to mark each of the identified objects with an appropriate annotation tag. (Guidelines for identifying objects were also provided.) This task was chosen as being representative of real-world airborne reconnaissance activities while being sufficiently simple to allow control over experimental design. For one segment, participants were permitted to use only the manual interface to make annotations; speech and/or manual mode were permitted for the other.

We solicited eight volunteers, all of whom were technical employees. No attempt was made to select participants on demographic characteristics or on computer skills; the volunteers were chosen based on their willingness to participate and on their availability.

For the experiment sessions, we used a Solaris workstation to run the MSIIA system. Before each session, the MSIIA was launched and configured so that each participant had the same view into the system, and the controls were positioned in a standard layout. Figure 1 shows the setup of the MSIIA. The video window, in the upper right hand corner, displays the streaming video clips loaded by the experimenters. The user controls for the video window are to the left. Below the video window is an annotation palette, the device for annotating video images.

The user control window provides controls for playing, pausing, and stopping the video, manually navigating through the video segment (by time or frame increment), and manipulating the video playback speed (in frames per second). In this experiment, the user controls were accessible via manual input (mouse) only.



Imagery analysts can view multiple streams of visual data for airborne surveillance and reconnaissance activities

User controls for video

Palette contains multiple tabs of annotation labels

Analysts can click through tabs and select annotations OR use speech-enabled input for hands-free, eyes-free annotations.

Figure 1. Participant view of the MSIIA system as configured for experiment

The annotation palette provides a direct way to mark or annotate the objects. The palette consists of several pages, or tabs, of the annotation labels available; each tab represents a single category of annotations. Once an object in the video has been identified, a participant finds the appropriate category tab, chooses a tag, and clicks the associated button. Annotations can be made to the video in any state (i.e., play, pause, or stop) and at any speed. Both the tab switching and annotation tag selection in the annotation palette are accessible through either manual input (i.e., clicking and button pressing) or speech-enabled input (described below). Independent of mode, tab selection results in displaying the selected tab. When an annotation is made, no visible changes are apparent in the video, but the selected annotation (whether manual or voice-activated) is indicated by a visual button depress. In this exploratory study, no mechanism for editing or deleting annotations was provided in either mode.

The speech-enabled input component consists of a modified Nuance speech recognizer agent on a separate networked Windows computer. Subjects wore a head-mounted, close-talking microphone while seated in front of the Solaris workstation. A Java interface is used to communicate with the annotation palette and a speech feedback GUI, a small window that provides minimal indication of the state of the speech agent. This allows subjects to select items in the annotation palette verbally, without using the mouse. The modified client accepts a simple grammar consisting of a keyword and one or more identifiers, as follows. (The keyword, the verb *record*, is a substitute for the alternate push-to-talk method that would not allow for completely hands-free annotation.)

record <*tab*> <*annotation tag*>

where either <*tab*> or <*annotation tag*> is optional, but at least one is required. Table 2 lists examples of valid commands and their resulting actions.

Users can make an annotation via speech even when that particular annotation button is not visible on the screen. (In manual mode, a user is required to select the tab, search the annotation list, and physically click on the button next to the desired annotation tag.) This is a main advantage of speech interfaces over GUIs and DMIs; users can identify and select objects not on the screen as well as identify and manipulate objects from large sets (Cohen, 1992).

Upon arrival, participants read and signed a consent form. They were then asked to complete a short questionnaire designed to gather background information on age, gender, professional status, skill/familiarity with imagery analysis, skill/familiarity with the MSIIA, skill/familiarity with speech-enabled input, etc.

The participants were given a hardcopy of an overview of the experiment and the MSIIA system while the experimenters read it aloud. The ensuing hands-on training session guided the participants through the use of the MSIIA and a simplified identification and annotation task similar to the actual experimental task. The training provided experience both with manual input and speech-enabled input. The training also helped the participants become familiar with the annotation tag sets. A guide with hints on how to identify images was provided as well.

The experimenters read task instructions to the participants and gave them a half-hour time limit to complete each of two trials. Half of the participants started with a trial in manual input mode and then did a trial with the addition of speech-enabled input. The other half did the reverse. The experimenter observed the participant during both sessions, provided assistance, and recorded critical incidents (discussed further below).

After each trial, the participants completed a one-page questionnaire based on that particular trial mode. While the majority of the questions were Likert-scale, several open-ended questions were designed to stimulate brainstorming. After both trials were completed, the participants were asked to answer another set of questions comparing the two modes. Finally, the experimenter asked for questions, comments, and other feedback during a short interview period.

Table 2. Examples of speech commands and ensuing actions

Speech	Ensuing Action
"record terrain"	• Annotation palette displays *Terrain* tab if not already visible
"record terrain valley or trench"	• Annotation palette displays *Terrain* tab if not already visible • Annotation palette presses the <**Set**> button for **VALLEY/TRENCH** annotation tag
"record valley or trench"	(same as immediately above)

4. Methodology: Metrics and data collection

We wanted to test whether the MSIIA system augmented with speech-enabled input would lead to better and faster task performance and that participants would be more satisfied than with mouse-only input. To this end, we defined five high-level metric categories: efficiency, quality, task success, user satisfaction, and usability. These categories were modified from those established by the DARPA Communicator project (Walker et. al., 2001). Each category consists of one or more quantifiable metrics such as time on task, precision, recall, and several user-rated perceptions. A complete listing of categories, their associated metrics, and definitions is detailed in Table 3.

Table 3. Metrics

Category	Metric	Definitions, notes, examples
Efficiency	Time on task Image identification and annotation efficiency	Assumes half hour time limit did not create ceiling effect • Playback speed • Video state (stopped, playing, paused)
Quality	Task outcome (precision)	Precision = (# images accurately marked) / (# images marked)

Task success	Task completion (recall)	Recall = (# images accurately marked) / (# markable images in master key)
	Perceived task completion	Subjective value based on questionnaire
User satisfaction	Task ease	Subjective value based on questionnaire
	User expertise	*Did user know how to use system and each feature?*
	Expected behavior	*Did the system/input mode work as expected for this task?*
	Future use	*Would the participant use the system/input mode again? Regularly?*
Usability	Critical incidents	Critical incident is any event, positive or negative, fatal or non-fatal, which interrupts task execution
	Errors	• Using controls incorrectly • Marking an image and then wanting to edit or remove that annotation • "Wrong path" errors • Using incorrect speech "command" or trying to do or say something system or speech recognizer does not understand
	Repair activities	Attempt to backtrack or correct an error
	User feedback	Comments made during or after experiment

The evaluation focused on both quantitative and qualitative, anecdotal reactions. The pre-experiment questionnaire provided quantitative background information on the participants. The two test trial questionnaires and the final questionnaire provided quantitative data on user satisfaction and perceived task completion.

The MSIIA was instrumented to record each time-stamped annotation event and associated data such as frame number (for purposes of indexing), the annotation tag, playback speed setting (frames per second), and state of the video tool (whether stopped, playing, or paused). Time on each task was recorded as well as overall experiment time. Quantitative background data from the questionnaires (pre-experiment questionnaire, two trial questionnaires, and the post-experiment questionnaire) were tabulated.

The automated logfiles were parsed and analyzed to calculate precision, recall, image identification efficiency (playback speed), and annotation efficiency (play-to-stop ratio). In order to calculate precision and recall, annotations in the logs were compared to a master annotation file. Each annotation was marked as correct, incorrect or missing. Precision scores were computed as the number of images correctly annotated divided by the number of images annotated. Recall was computed as the number of images correctly annotated divided by the total number of correct annotations (in the master annotation file).

During each of the experiment sessions, we observed the participants and made notes of critical incidents, errors, and repair activities related to the task and to usability of the system. (These errors and activities were not automatically recorded for this experiment.) We also recorded participants' comments and questions. Interviews, based on open-ended questions, gathered data on participants' reactions to speech-enabled input as well as to the system and the task itself.

5. Results and discussion

All participants in this experiment were male engineers, ranging in age from 22 to 40. All were inexperienced in areas of domain, task, and specific technology. None had ever been involved in airborne surveillance and reconnaissance activities nor had any performed imagery analysis prior to this experiment. None had seen or used the MSIIA system before, but two had heard of it. Only one participant had used speech-enabled input, and then only once or twice in his work.

5.1. Hypothesis 1: Speech annotation faster?

Our first hypothesis is that people can annotate images in video segments *faster* with the MSIIA augmented by speech.

Quantitative results indicate that participants in this study might be able to annotate images faster with speech. On average, participants spent less time on the task when speech-enabled input was available although the difference was not significant according to a paired, one-tailed distribution t-Test. See Table 4. Image identification efficiency, however, was significantly higher in speech mode. This means that users were able to play the video segment at faster speeds in speech mode. There is some indication that annotation efficiency could also be higher when speech is available; participants paused or stopped the video less often when making annotations.

Table 4. Efficiency-related results

Metric	μ_{manual}	μ_{speech}	**Sig.**	**St. dev.**
Time on task	24.38 min	23.31 min	*None*	

Image ID (playback speed)	7.51 fps	15.36 fps	0.01	0.17
Annotation (play/stop)	0.09	0.34	*None*	

Rudnicky reported that there is no evidence supporting speech as an advantageous modality in terms of an aggregate measure such as time on task although it is consistently faster at the level of single input operations (1993). He attributes this difference to the added costs of non-real-time recognition and error correction. Oviatt's research suggests that error detection and correction is the crucial factor in determining task completion times (Oviatt, 1994). Karat et al. (1999) also believe that examining error detection and correction is important in explaining differences in modalities. They have found that measures such as time on task, which include error detection and correction times, favor keyboard/mouse input devices over speech. In their 1999 study examining errors, they showed that the average number of corrections in speech tasks was slightly higher than for keyboard/mouse tasks, and the length of time to correct these errors was much longer in speech tasks. They noted that participants tended to correct keyboard/mouse errors in text entry within a few words of having made it. In contrast, some participants reported they were not always aware of misrecognition in speech tasks. Mellor et al. (1996) plotted task completion times against speech recognizer word accuracy and showed that task completion times decrease with increasing recognizer performance. They believe that speech could potentially provide equivalent performance times to manual mode inputs if word accuracy were closer to 94%, given the task-specific vocabulary. In a different type of task study comparing voice controlled to mouse controlled web browsing, Christian et al. (2000) observed that voice browsing the web (navigating slide shows and hierarchical menus) took an average of 1.5 times longer than mouse browsing even though error rates (for both missed and misinterpreted commands) were low.

Much of error detection involves confidence. When users push a button on a mouse, they can feel quite certain of the result. When users speak to an ASR system, they may experience system errors - errors in which the system output does not match their input - which they do not experience with other devices. (Karat et. al., 1999)

In manual mode, a participant could select the wrong tab in the annotation palette or the wrong annotation tag button or make an annotation in the wrong video frame. The user will notice almost immediately when he selects the wrong tab (desired annotation tags are not visible on the selected tab), and correcting that error simply involves selecting another tab. Selecting an incorrect annotation tag may not be as easily detected since there is no support for visualizing annotations made. Detecting the wrong video frame may be impossible for the same reason. Our experimental system does not support correction of annotation errors other than allowing the user to make other, correct annotations in addition to the incorrect ones (i.e., no delete or edit functions were made available).

Ignoring speech recognition errors for a moment, these same errors of intent also occurred in speech mode. However, an incorrectly selected tab did not always pose a problem unless the user actually wanted to scan the contents (since tab selection was not a prerequisite to selecting an annotation on that tab). Detection of these errors involved a shift of focus and a time delay; the user either had to look in the speech feedback GUI for the recognized text or look at the changes in the annotation palette (tab switch and/or button depress). Some users ignored (or did not notice) the feedback, but others paused to divert their attention to the speech feedback GUI, thus reducing the benefit of having speech-enabled input available - for an eyes-busy situation where users need to focus on the task at hand. Error corrections involved repeating the command or issuing a command for another, correct annotation or, alternately, using the mouse.

Participants in speech mode also experienced other types of errors including forgetting to use the command word ('record'), choosing an annotation that does not exist in the annotation palette, using the wrong phrasing for an annotation, and recognition errors. Again, detection of these errors involved diverting attention from the video window, and correction involved repetition.

5.2. Hypothesis 2: Speech annotation better?

Our second hypothesis is that people can annotate images in video segments *better* with the MSIIA augmented by speech. (*Better* is defined as more targets annotated and targets more accurately annotated.)

Our results do not support this hypothesis. Both quality and task success metrics were lower for the speech task. (Table 5). The average precision score was slightly lower, and the average recall score was significantly lower.

Table 5. Results on quality (outcome) and task success (completion)

Metric	μ_{manual}	μ_{speech}	Sig.	St. dev.
Task outcome (precision)	0.36	0.31	0.05	0.04
Task completion (recall)	0.84	0.81	*None*	

A study by Karat et al. (1999) showed no statistical difference in quality between modalities. In their experiment, users composed text using speech recognition and keyboard and mouse where users had the ability to make corrections in both modalities. We believe that the lower recall and precision scores in our experiment can be attributed to the lack of undo and editing capabilities combined with insufficient experience by naïve users in an unfamiliar domain. According to Karat et al, the most common command used is the undo command. Our participants expressed frustration at not being readily and unambiguously able to identify images in the video. Images were difficult to discern because of poor focus and resolution of the video, level of viewable detail, and lack of familiarity with airborne surveillance tasks. For example, upon seeing a roof-less structure, a participant might immediately think it was a damaged house. As the video camera zooms in for a closer look or gets a different angle, construction materials may become visible, indicating that it is a house under construction. Similarly, vehicles were not often immediately distinguishable.

When users had speech available, they often blurted out the first thing that came to mind, resulting in recorded annotations that were not always correct. Since editing and undo capabilities were not provided, participants could not correct errors. In manual mode, users more often paused the video to divert their attention to the annotation palette where they were forced to click through tabs and search lists for specific annotation tags. During that process, they were often visually reminded of tags they might not have remembered, and they had ample time to think about the image and change their mind. (This explanation was provided via interviews.) It is not the case, however, that participants made more annotations in speech mode. In fact, they made slightly fewer annotations (an average of 28.6 annotations in manual mode versus an average of 25.6 annotations in speech mode) which tended to be correct less often than manual annotations.

5.3. Hypothesis 3: Speech annotation preferred?

Our third hypothesis is that people *prefer* speech-enabled input to manual input when annotating images in video segments in the MSIIA.

Participants liked the speech-enabled input. They felt it made it both faster and easier to annotate images in the video clips. The user reports are consistent with the efficiency results in Table 4. Note that statistical significance is shown only for user ratings of annotation speed.

Mellor et al. (1997) compared task completion times to ASR performance, and observed that users preferred speech-enabled input even when ASR performance was poor. In fact, speech-enabled input was favored over other modes of input despite its lower performance.

Table 6 itemizes user satisfaction results from questionnaires completed both during and after the experiment. After each trial, participants were asked to respond to questions based on that particular mode. At the end of the experiments, the participants were asked to compare the two modes in terms of annotation ease, annotation efficiency, and navigation. Normalized scores for each question are shown for both modes. These scores are combined into overall scores corresponding to our user satisfaction metrics: task ease, user expertise, and expected behavior. Several participants commented that the second modality (speech) was very effective in reducing the necessity to navigate controls and in allowing them to focus more on the task. Having to use the annotation controls manually diverted their attention.

Overall, however, participants found the task easier to do in manual mode and also believed that their annotations were more accurate. We have no explanation as to why participants thought the overall task easier to do without speech, but perceptions on accuracy are consistent with recall and precision measures in Table 5. Negative comments centered on the lack of confidence participants had for the accuracy and temporal precision of the speech-enabled input. The recognized speech appeared as text in the speech feedback GUI, but the delay was considerable enough that participants often doubted their annotation was recorded in the correct frame. One participant commented that he never even bothered to wait for feedback for some of the longer strings, and so he was never quite sure that all of his annotations were correctly captured or even captured at all. In general, participants were more confident of image identification and annotation accuracy in manual mode.

Table 6. Normalized USER SATISFACTION results from questionnaires during and after experiment

Cat	Metric	μ_{manual}	μ_{speech}	Sig.	St. Dev.	Preferred mode	
						During	After
User satisfaction	Overall Task Ease[1]	0.65	0.61	*None*		Manual	
	Ease in image ID	0.49	0.54	*None*		Speech	
	Ease in finding tags	0.71	0.54	*None*		Manual	
	Ease in annotating	0.77	0.82	*None*		Speech	Speech
	Speed in annotating	0.70	0.86	0.05	0.11	Speech	Speech
	Correct image ID	0.39	0.34	*None*		Manual	

	Correct annotating	0.64	0.5	*None*		Manual	
	Enough time	0.84	0.77	*None*		Manual	
	Ease in performing task	0.64	0.54	0.02	0.08	Manual	
	Less navigation needed					*N/A*	Speech
	Overall User Expertise[2]	0.67	0.71	*None*		Speech	
	Training	0.71	0.79	*None*		Speech	
	Navigating	0.63	0.63			Equal	
	Overall Expected Behavior[3]	0.80	0.73	*None*		Manual	
	System responded	0.81	0.72	*None*		Manual	
	Knew what system was doing	0.78	0.75	*None*		Manual	
	Future Use	*Data incomplete*[4]					

1. Roll-up of scores pertaining to ease, speed, time allotment, and task.
2. Roll-up of scores from questions on training and navigation.
3. Roll-up score on system response and expectations.
4. This question was asked orally of some of the participants but not all.

Participants felt they were less sure of system behavior when using speech. They did not always know what the system was doing and felt the system did not always respond as expected. These sentiments are related to the previously mentioned feedback issue.

We did not ask every participant whether he would use the system again. Those we asked said they would use the system with speech again if several things were improved. Desired improvements included: bug fixes in the video playback mechanism, speech-enabled input added to high-level navigation controls (i.e., stop, pause, and play), annotation visualization, the ability to edit or delete annotations, and better feedback from the speech recognizer. These are detailed in the following section on usability.

Rudnicky (1993) noted users preferred speech to other modalities (keyboard and scroll bar) even when it was less efficient in terms of overall task time and error detection and recovery. Participants possibly based their preference on input time rather than on overall task time. However, longer utterances resulted in a decreased preference for speech. Perhaps users are willing to ignore additional costs of errors only to a certain degree.

6. Usability

During the experiment sessions, we asked each of the participants to comment on the system, the task, and the input mode, and also to speak freely about what he was doing. In addition, an observer recorded critical incidents including errors and attempts at recovery.

When speech mode was available, participants chose speech over manual input to make all annotations. Several participants, however, occasionally used the mouse to switch between tabs on the annotation palette but then used speech to make the actual annotations. Those who did this claimed that the physical action of switching tabs with a mouse click gave them more time to read the contents of each tab since they were not quite familiar with the annotation tag sets. Once they found the desired annotation tag, they used speech to select it since they could use voice while moving their gaze back to the video.

Participants commented that they tended to pause or stop the video much less often when speech was available and were able to play the video at a higher speed. Speech-enabled input allowed the participants to focus their eyes and attention better on the streaming video. As one user reported, "I started and stopped the video less with speech. I was watching faster and pausing less. In manual mode, I kept missing the video if I tried to navigate the tabs on the annotation palette."

One of the biggest concerns was the lack of confidence users had for the speech recognizer precision. In order to validate that their speech was accurately recognized, the users were forced to wait for visual feedback of a text string in the speech feedback GUI window. Because the recognized string appeared several seconds later, users were not sure when the annotations had been made. In manual mode, the users assumed the annotation was made immediately after they clicked the button next to the selected annotation. In speech mode, button presses were visually simulated after speech recognition, but this was both after a delay and also not necessarily in the visual field of the user. Likewise, users believed that annotations could be made as fast as they could click the annotation tag buttons, but they were not convinced that all of the spoken annotations were actually recorded because of the time lag. Since there was no support for annotation visualization, participants could not determine in which video frame the annotation had been made and even whether the annotation had been made at all.

Participants speculated on how speech-enabled input could be even more effective. In addition to wanting faster and more accurate speech recognition, they wanted the ability to use speech shortcuts for long annotation tag names or synonyms for vocabulary they were less likely to remember. Some users guessed that speech would probably be more useful for a similar but more complicated task, i.e. a more complex annotation palette with more categories (tabs) and a larger tag set. Of course, success with this more complicated task assumes that users are domain experts - more knowledgeable about imagery identification and more familiar with the tag set.

Most of the participants also expressed a desire to have speech-enabled input available for video navigation at a high-level (e.g. to control play, stop, and pause functions). None felt that the speech recognition was accurate and fast enough to use for more fine-grained navigation such as manipulating a slider to alter speed or to advance one frame at a time.

All participants said they would like to use speech-enabled input for performing a task similar to the imagery analysis and annotation task if speech recognition were faster and more accurate, and if there were better feedback mechanisms for both speech recognition and annotation visualization. In addition, an editing or undo capability would be highly beneficial at whatever level of confidence users had for the speech recognizer.

7. Speech in Human-Robot Interaction

We now compare results from this experiment with our project controlling a robot with a PDA.

Christiansen and colleagues (Christiansen, 2002) are investigating using teams of robots to find victims in mock search-and-rescue scenarios, using NIST's Standardized Test Course for Urban Search and Rescue Robots. In collaboration, we have developed a wireless PDA interface to control the robots. The PDA displays a map of the rooms being searched, the current location of the robot, and a window with a live video feed from the robot's camera. There are three ways to control the robot. Users can touch the PDA stylus on a location in the map, which causes the robot to move to that location. Users can also use pre-set menus to send the robot to pre-defined locations on the map, as well as to issue directional commands such as "turn left", "rotate 45 degrees", or "stop". Finally, users can issue speech commands to invoke any of the menu commands (the PDA supports speech recognition). The robot acknowledges the command verbally, via speech synthesis on the PDA. All interactions between human and robot (stylus, menu, speech recognition, or speech synthesis) are mediated by a dialogue manager (Burke et al 2002).

This interface is similar to the MSIIA in that both require monitoring of a video stream and manipulation of a GUI with an input device or speech. Our intuition was that the PDA interface was so tiny that users would prefer speaking to it than to manipulating the stylus. Yet in contrast, and unlike the MSIIA, we have found that robot controllers typically abandon speech quickly in favor of stylus and menus. We speculate three reasons for this.

- The rate of incoming information on the robot video feed is slower than that of the MSIIA, as the robot moves very slowly, and is often stopped. Thus, there is more "downtime" to use the GUI; the robot video is not as "eyes busy".
- The field of view on the PDA is so small that people can use the menus or map and still keep the video feed in view. In contrast, the MSIIA uses a large screen with menus not in the same field of view as the video. Thus, the PDA permits overloading the visual field, akin to heads-up cockpit displays.
- The operator task is different. Whereas the MSIIA requires active annotation of the video, the robot controller merely needs to keep an eye out for obstacles (to avoid) and victims (to approach), and then to direct the robot appropriately. Thus, there is less penalty (in terms of task completeness) for the robot controllers to use the GUI.

Like the MSIIA system, the robot interface had inevitable delays and errors in speech recognition. MSIIA users apparently felt the drawbacks of speech were overcome by its benefits in a demanding, "eyes-busy" task. Robot controllers had a less demanding task, and therefore less incentive to put up with the delays and errors of speech recognition.

8. Conclusion

The results of our experiment show that adding speech to a multi-windowed video annotation system such as the MSIIA enabled people to identify images in streaming video more efficiently and might enable people to annotate images faster than with just the mouse alone. Participants were not able to annotate better with speech, however, and we believe this can be attributed to poor feedback and visualization, the unavailability of undo and editing capabilities, and also the lack of task and domain expertise. Users liked speech and felt that it made it easier and faster to annotate images because it kept their eyes hands and free so they could focus more on the task.

This formative study indicates that speech-enabled input may lead to improved performance and increased user satisfaction of naïve and expert domain users on more complicated tasks. We believe that we have not fully tested our hypothesis that speech recognition can be a cognition-enabling technology. We plan to use the feedback from this experiment to improve the MSIIA system with speech-enabled input by adding speech to more components of the system, enabling access to correction mechanisms, supporting visualization of representations of annotations, and improving feedback. In addition, to simulate real world tasking, for future experimentation we will increase the complexity of the annotation tag set, provide better domain training to naïve users, and eventually use real imagery analysts as participants in our studies.

Unlike the MSIIA system, users dispreferred speech when controlling robots with a PDA. We surmise that robot controllers, with a less demanding task, had less incentive to put up with the delays and errors of speech recognition. We conclude from both projects that speech is helpful if the cognitive cost of speech recognition delays and errors is outweighed by the cognitive benefit of reducing instrument manipulation.

9. Acknowledgements

The MSIIA experiment was funded was funded by the Defense Advanced Research Projects Agency (DARPA) Augmented Cognition project, under contract number F19628-99-C-0001. The robot PDA project was funded by the MITRE Corporation.

10. References

Burke, C., Harper, L., Loehr, D. “A Dialogue Architecture for Multimodal Control of Robots”. International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, 2002.

Christian, K., Kules, B., Shneiderman, B., and Youssef, A. "A Comparison of Voice Controlled and Mouse Controlled Web Browsing", Proceedings of the fourth international ACM conference on assistive technologies, Nov. 2000.

Christiansen, A. “Robot Platoon Command and Control”. URL: http://www.mitre.org/technology/tech02/tech_02/briefings/intelligent_information/christiansen_presentation/index.htm

Cohen, P. "The Role of Natural Language in a Multimodal Interface", UIST '92, 1992, pages 143-149.

DARPA Augmented Cognition Program, http://www.darpa.mil/ito/research/ac/index.html, 2001.

Grasso, M., Ebert, D., and Finin, T. "The Integrality of Speech in Multimodal Interfaces", *ACM Transactions on Computer-Human Interaction*, Vol. 5, No. 4, Dec. 1998, pages 303-325.

Hansen, S. "MSIIA Hunts Predator in Bosnia", *The Edge, MITRE Advanced Technology Newsletter*, Vol. 1, No. 1, Mar. 1997.

Karat, C-M., Halverson, C., Horn, D., and Karat, J. "Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems", Proceedings of Conference on Human Factors in Computing Systems: CHI '99, Pittsburgh, PA, 1999, pages 568-575.

Mellor, B., Baber, C., and Tunley, C. "Evaluating Automatic Speech Recognition as a Component of a Multi-Input Device Human-Computer Interface", ICSLP ‘96.

Oviatt, S., "Interface Techniques for Minimizing Disfluent Input to Spoken Language Systems", Proceedings of Conference on Human Factors in Computing Systems: CHI '94, Boston, 1994, pages 205-210.

Oviatt, S. "Taming Recognition Errors with a Multimodal Interface", *Communications of the ACM,* Vol. 43, No. 9, Sept. 2000, pages 45-51.

Oviatt, S. and Cohen, P. "Multimodal Interfaces that Process What Comes Naturally", *Communications of the ACM*, Vol. 43, No. 3, Mar. 2000, pages 45-53.

Polson, M., Friedman, A. “Task-sharing within and between hemispheres: A multiple-resources approach”, *Human Factors*, 1988, Vol. 30, pages 633-643.

Rosenfeld, R., Olsen, D., and Rudnicky, A. "Universal Speech Interfaces", *Interactions*, Nov./Dec. 2001, pages 34-44.

Rudnicky, A. "Mode Preference in a simple data-retrieval task", INTERACT '93 and CHI '93 conference companion on Human factors in computing systems, Apr. 1993.

Shneiderman, B. "Direct manipulation: A step beyond programming languages", *Computer*, 1983, Vol. 16, No. 8, pages 57-69.

Shneiderman, B. "The Limits of Speech Recognition", *Communications of the ACM*, Vol. 43, No. 9, Sept. 2000, pages 63-65.

Walker, M., Aberdeen, J., Boland, J., Bratt, E., Garofolo, J., Hirschman, L., Le, A., Lee, S., Narayanan, S., Papineni, K., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., Rudnicky, A., Sanders, G., Seneff, S., Stallard, D., Whittaker, S. "DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection", Proceedings of EUROSPEECH 2001.

Wickens, C. and Hollands, J., *Engineering Psychology And Human Performance*, 3rd ed., Prentice-Hall, 1988, page 451.