INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 133–144, 2003

The MSIIA Experiment: Using Speech to Enhance Human Performance on a Cognitive Task

LAURIE DAMIANOS, DAN LOEHR, CARL BURKE, STEVE HANSEN AND MICHAEL VISZMEG
The MITRE Corporation, m/s K309, 202 Burlington Road, Bedford, MA 01730, USA

Abstract. We performed an exploratory study to examine the effects of speech-enabled input on a cognitive task involving analysis and annotation of objects in aerial reconnaissance videos. We added speech to an information fusion system to allow for hands-free annotation in order to examine the effect on efficiency, quality, task success, and user satisfaction. We hypothesized that speech recognition could be a cognitive-enabling technology by reducing the mental load of instrument manipulation and freeing up resources for the task at hand.

Despite the lack of confidence participants had for the accuracy and temporal precision of the speech-enabled input, each reported that speech made it easier and faster to annotate images. When speech input was available, participants chose speech over manual input to make all annotations. Several participants noted that the additional modality was very effective in reducing the necessity to navigate controls and in allowing them to focus more on the task. Quantitative results suggest that people could potentially identify images *faster* with speech. However, people did not annotate *better* with speech (precision was lower, and recall was significantly lower). We attribute the lower recall/precision scores to the lack of undo and editing capabilities and insufficient experience by naïve users in an unfamiliar domain.

This formative study has provided feedback for further development of the system augmented with speech-enabled input, as our results show that the availability of speech may lead to improved performance of expert domain users on more complicated tasks.

Keywords: augmented cognition, experiment, imagery annotation, multi-modality, speech interface

Introduction

This experiment was designed to measure improvements in human performance by integrating enabling technologies into an existing operational system. Specifically, the experiment examines the effects of speech-enabled input on the Multi-Source Intelligence Integration and Analysis (MSIIA) system (Hansen, 1997, described below) in performing a simple task involving the analysis and annotation of objects in aerial reconnaissance videos. The objective was to demonstrate quantifiable enhancements to human cognitive ability in an operational military environment.

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 optional hands-free input of annotations. We hypothesized that speech recognition could be a cognitive-enabling technology by reducing the mental load of instrument manipulation and freeing up resources for the task at hand.

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 (Defense Advanced Research Projects Agency) 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 mission is to develop and demonstrate quantifiable enhancements to human cognitive ability in diverse, stressful, operational environments.

The Augmented Cognition effort plans to develop and implement strategies of multiple sensory inputs and mixed initiatives between human- and computer-generated interactions. The specific hypothesis is that freeing up more cognitive resources, augmented by computer intelligence to aid in context awareness, will create a better human-computer symbiosis. The goal is to move away from the current paradigm of human learning the system and move towards the human and machine cooperating to solve problems and arrive at decisions.

Knowledge of human cognitive functions has enabled us to imagine automation systems in which the human-computer interface is less obtrusive, and collaboration could more closely resemble human-human interactions (involving anticipation, mixed-initiative interactions, dialogue, gesturing, etc.). One of the challenges facing the Augmented Cognition program is designing these interfaces based on cognition. Historically, graphical user interfaces are organized by application, not by content, and they are largely uni-modal.

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." These emerging multi-modal systems offer "expressive, transparent, efficient, robust, and mobile human-computer interaction." Oviatt and Cohen believe that, for many applications, multi-modal systems will eventually replace standard graphical user interfaces of today's systems.

Mellor et al. (1996) also see the future of human-computer interactions in terms of multiple interaction modes. They state: "The choice of interaction devices and the combination of these devices in a true multi-modal interface will depend on many factors; the individual component performance, user experience and ability, the computer application and fundamental properties of the various interface modalities can all be expected to influence the overall interface design."

Multi-modality does not necessarily entail using speech. Why do we hypothesize that the addition of speech might 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, some believe 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). Cohen suggests that GUIs/DMIs and speech recognition interfaces have complementary strengths and weaknesses that could be utilized in multi-modal systems.

Furthermore, in many situations when conventional means of human-computer interaction (via keyboard/mouse or using a display) are neither feasible nor desirable, speech can be indispensable. Even when standard interaction modes are possible, speech-enabled input can be supplementary (Rosenfeld et al., 2001). In fact, it is doubtful that speech could completely replace the keyboard and mouse in all applications (Rudnicky et al., 1994), but the ideal future interface will support a combination of input modalities from which the user can select, specific to a task. The flexibility of multi-modal systems allows users to exercise their own judgment in tailoring their input style to specific tasks and in using a combination of several modes (Oviatt, 2000; Rudnicky, 1993). If two or more input modes provide parallel or duplicate functionality, users can alternate their choice of input modes to reduce the likelihood of errors or resolve existing ones (Oviatt, 2000; Oviatt and Cohen, 2000). From a usability perspective, this flexibility makes multi-modal interfaces better at facilitating error avoidance and recovery than uni-modal speech systems (Oviatt, 2000). Oviatt (1996) demonstrates clear task performance and user preference for multi-modal interfaces over uni-modal speech interfaces: 10% faster task completion, 23% fewer words, 35% fewer task errors, and 35% fewer spoken disfluencies.

It was predicted decades ago that automatic speech recognition and text-to-speech output would revolutionize the human-computer interface, allowing completely natural and error-free communication (Rosenfeld et al., 2001; Waterworth, 1987). Since performance levels of speech input/output devices have not improved according to predictions, "conversational" computers are still not quite reality. Natural interaction is not the only benefit or goal of speech interaction, however. 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 (as can keyboard shortcuts in well-designed manual interfaces), 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).

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." "Speech is slow for presenting information, is transient and therefore difficult to review or edit, and interferes significantly with other cognitive tasks" (Shneiderman, 2000). Shneiderman argues 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. Oshika[1] concurs, noting, "A main aspect in the notion of 'cognition-enabling' is the performance and interaction of the speech interface itself. If the interface requires cognitive effort to use (confirming dialogues, a different way of structuring the input, etc.) and makes errors, then at some point it becomes more trouble than it is worth."

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 spoken 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 and Friedman, 1988). The separation of spatial and verbal resources seemingly accounts for the high degree of efficiency with which manual and vocal outputs can be time-shared..."

In summary, some researchers claim that speech interfaces can be effective when used in conjunction with other modalities so that the complementary strengths of multiple, integrated modalities 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.

The Experiment

Our experiment was designed to test the following hypotheses:

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

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. The test video segments were longer ($\sim$10 minutes each) and more content-rich than the training video segments. 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. Table 1 itemizes the four different treatment conditions (where order matters).

Table 1. The four experimental treatment conditions consist of combinations of two input modes, two video clips, and alternating orders of both.

	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 a set of annotation tags.

Participants were asked to review the 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. Guidelines for identifying specific objects were provided. The participants were told to mark each of the identified objects with an appropriate annotation tag. This task was chosen as being representative of real-world airborne surveillance and reconnaissance activities while being sufficiently simple to allow control over experimental design. For one video clip, 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 dedicated Solaris workstation to run the MSIIA system. Before each session, the MSIIA system 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 system. The video window, in the upper right hand corner, displayed the video clips loaded by the experimenters. The user controls for the video window were positioned to the left. Below the video window was an annotation palette, a tool for selecting tags and annotating images in the video.

The user control window provided 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.

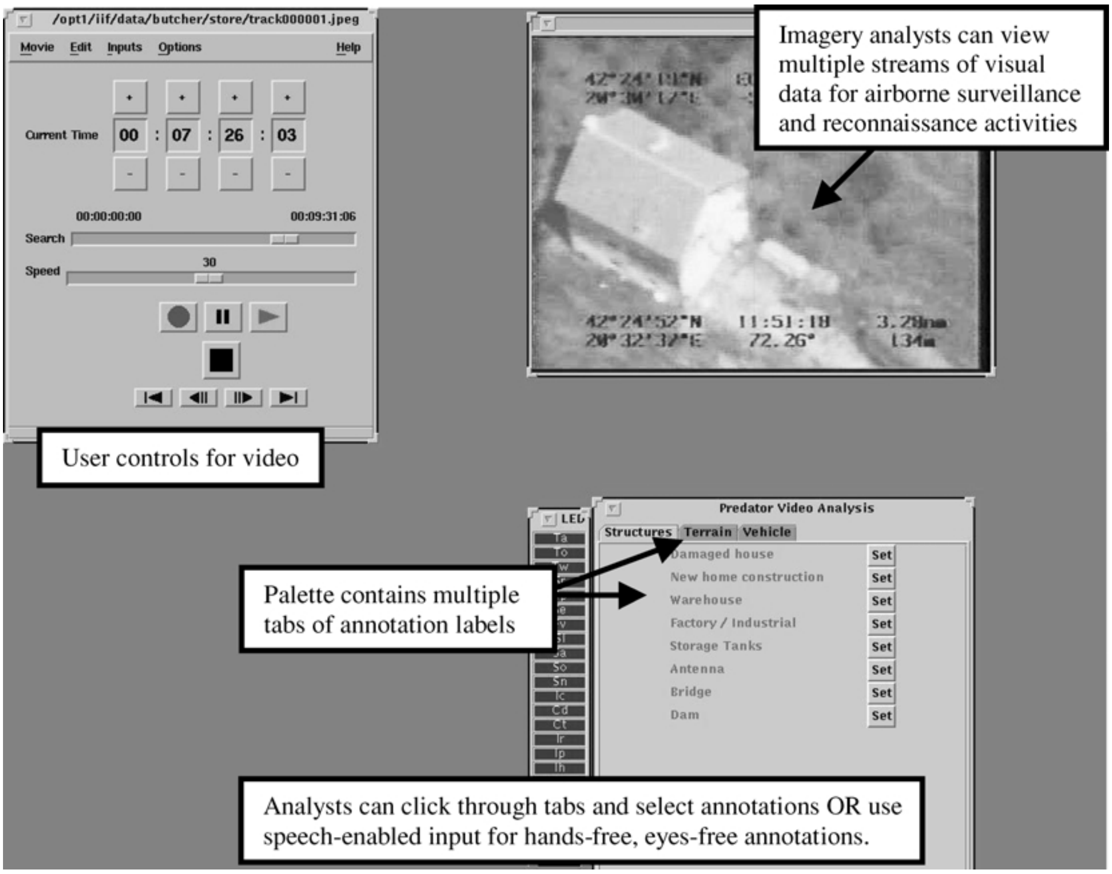
The annotation palette consisted of several pages, or tabs, of the annotation labels available; each tab represented a single category of annotations. Once an object in the video had been identified, a participant would select the appropriate category tab, choose an annotation tag, and click the associated button. Annotations could 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 were accessible through either manual input (i.e., clicking and button pressing) or speech-enabled input (described below). Independent of mode, tab selection resulted in displaying the selected tab. When an annotation was made, no visible changes were apparent in the video, but the selected annotation (whether manual or speech-activated) was 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 consisted of a modified Nuance speech recognizer agent (Nuance, 2002) 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 was used to communicate with the annotation palette and a speech feedback GUI, a small window that provided minimal indication of the state of the speech agent. This allowed subjects to select items in the annotation palette orally, without using the mouse.

The modified client accepted a simple grammar consisting of a keyword and one or more identifiers. (The keyword, the verb *record*, was 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*> was optional, but at least one was required. The vocabulary size was 42 words, comprising a single keyword, 3 tab names, and 28 annotations (an annotation could require multiple words, as in the annotation "valley or trench"). The annotation vocabulary was based on a selected subset of a standard listing of annotations. The tab names were fabricated to represent categories of the annotation subset. Table 2 lists examples of valid commands and their resulting actions.


/opt1/iif/data/butcher/store/track000001.jpeg
Movie Edit Inputs Options Help
Current Time 00 : 07 : 26 : 03
00:00:00:00
00:09:31:06
Search
Speed
30
42°24'52"N 11:51:18 3.28nm
20°32'32"E 72.26° 134m
LED
Ta To Tw So Sn Ic Cd Ct Ir Ip Ih
Predator Video Analysis
Structures Terrain Vehicle
Damaged house Set
New home construction Set
Warehouse Set
Factory / Industrial Set
Storage Tanks Set
Antenna Set
Bridge Set
Dam Set

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.

Users could make an annotation via speech even when that particular annotation was not visible on the screen. (In manual mode, a user was required to select the tab, search the annotation list, and physically click on the button next to the desired annotation tag.) This is an 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). However, well-designed manual interfaces can also provide some of these features.

Table 2. Examples of speech commands and ensuing actions.

Spoken command	Physical action
"record terrain"	• Annotation palette displays *Terrain* tab if not already visible
"record terrain valley or trench"	• Annotation palette displays *Terrain* tab • Presses the <**Set**> button for **valley/trench** annotation tag
"record valley or trench"	(same as above) • Annotation palette displays *Terrain* tab • Presses the <**Set**> button for **valley/trench** annotation tag

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 system, 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 system 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 also included. 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.

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 just mouse input. We defined five high-level metric categories: efficiency, quality, task success, user satisfaction, and usability. These categories were adopted and modified from those established by the DARPA Communicator project (Walker et al., 2001). Each category consisted 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.

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 system 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.

Table 3. Categorized metrics used as a basis for this experiment.

Category	Metric	Definitions, notes, examples
Efficiency	Time on task	Assumes the half hour time limit did not create ceiling effect
	Image identification and annotation efficiency	• 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

Table 4. Efficiency-related results supporting Hypothesis 1. Users were able to play the video at faster speeds when speech was available. A paired, one-tailed distribution t-Test was used, on a sample size of 8, to determine significance.

Category	Metric	μ_{manual}	μ_{speech}	Resulting relationship	Significance	St dev
Efficiency	Time on task	24.38 min	23.31 min	$\mu_{\text{manual}} > \mu_{\text{speech}}$	None	
	Image ID (playback speed)	7.51 fps	15.36 fps	$\mu_{\text{manual}} < \mu_{\text{speech}}$	0.01	0.17
	Annotation (play/stop)	0.09	0.34	$\mu_{\text{manual}} < \mu_{\text{speech}}$	None	

The automated logfiles were parsed and analyzed to calculate precision, recall, image identification efficiency (playback speed), and annotation efficiency (play-to-stop ratio). 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. Where more than one annotation was made for the same image, the first was ignored (attributed to an error), and the second was considered. 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 annotations in the master annotation file.

During each of the experimental 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.

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 in his work, and then only once or twice.

Hypothesis 1. *People can annotate images in video segments* ***faster*** *with the MSIIA system 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.

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 attributed 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 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.

Our experimental system did not support direct correction of annotation errors (i.e., no delete or edit functions were available). However, the user could make other, correct annotations in addition to the incorrect ones, in essence making a meta-correction. Thus, the above discussion of error correction times can still be considered relevant to our experiment, and our participants made annotation errors in a number of ways. In manual mode, a participant could select the wrong tab in the annotation palette or the wrong button associated with an annotation tag or make an annotation in the wrong video frame. The user would notice almost immediately when he selected the wrong tab (desired annotation tags were not visible on the selected tab), and correcting that error simply involved selecting another tab. Selecting an incorrect annotation tag might not be as easily detected since there is no support for visualizing annotations made. Detecting the wrong video frame might be impossible for the same reason.

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 did 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 to other windows, and correction involved repetition. We did not record errors (mouse or speech), detection time, or correction time in this experiment.

Hypothesis 2. *People can annotate images in video segments* ***better*** *with the MSIIA system augmented by speech* (*Better* is defined as target items more accurately annotated.)

Our results do not support our second hypothesis that people can annotate *better* with speech. Both quality and task success metrics were lower for the speech task. See Table 5. The average precision score was slightly lower, and the average recall score was significantly lower.

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

Table 5. Results on quality and task success. There is no supporting evidence that people can annotate better with the addition of speech in this experiment. A paired, one-tailed distribution t-Test was used, on the sample size of 8, to determine significance.

Category	Metric	μ_{manual}	μ_{speech}	Resulting relationship	Significance	St dev
Quality	Task outcome (precision)	0.36	0.31	$\mu_{\text{manual}} > \mu_{\text{speech}}$	0.05	0.04
Task success	Task completion (recall)	0.84	0.81	$\mu_{\text{manual}} > \mu_{\text{speech}}$	None	

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 (such as cars, vans, buses, and trucks) 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 other than by making multiple annotations on the same image. 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 in annotations made in manual mode.

Hypothesis 3. *People* ***prefer*** *speech-enabled input to manual input when annotating images in video segments in the MSIIA system.*

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.

An important question is how users choose among multiple candidate input devices. In our study, when speech input was available, participants chose speech over manual input to make all annotations. Rudnicky (1993) reported that users preferred speech to other modalities (when given the choice among speech, keyboard, and scroll bar as input devices) 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. He ruled out the novelty effect by showing that the preference for speech increased over time. However, longer utterances resulted in a decreased preference for speech. Perhaps users were willing to ignore additional costs of errors only to a certain degree. Mellor et al. (1996) 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 our 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 additional modality (speech) was very effective in reducing the necessity to navigate controls and in allowing them to focus more on the task. They remarked that having to use the annotation controls manually was a diversion of attention. In fact, while participants used the navigational controls to stop, pause, and rewind the video less often, on average, when speech was available, Table 7 shows that there is no statistically significant evidence to support these participants' perceptions.

Overall, however, participants found the task easier to do in the 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,[2] 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.

Participants felt they were less sure of system behavior when using speech mode. They did not always

Table 6. Normalized user satisfaction results from questionnaires administered during and after experiment. A paired, one-tailed distribution t-Test was used, on the sample size of 8, to determine significance.

						Preferred mode	
Category	Metric	μ_{manual}	μ_{speech}	Sig	St dev	During	After
User satisfaction	OVERALL task ease[a]	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[b]	0.67	0.71	None		Speech	
	Training	0.71	0.79	None		Speech	
	Navigating	0.63	0.63			Equal	
	OVERALL expected behavior[c]	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[d]					

[a] Roll-up of scores pertaining to ease, speed, time allotment, and task.
[b] Roll-up of scores from questions on training and navigation.
[c] Roll-up score on system response and expectations.
[d] This question was asked orally of some of the participants but not all.

Table 7. The effects of mode use on navigation of video controls. While participants paused, stopped, and rewound the video more often when speech was not available, results using a paired, one-tailed t-Test show no significant difference.

Category	Metric	μ_{manual}	μ_{speech}	Resulting relationship	Significance	St dev
Efficiency	Navigational factor (i.e., count of pause, stop, rewind)	11.29	8.88	$\mu_{\text{manual}} > \mu_{\text{speech}}$	None	

know what the system was doing and believed that the system did not always respond as expected. These sentiments are closely 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. The noted list of 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[3] annotations, and better feedback from the speech recognizer. These are detailed in the following section on usability.

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.

As discussed, when speech input 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 exhibited this behavior 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 speak while moving their gaze back to the video.

Participants noted 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. (These reports are consistent with the data analysis shown in Table 4, where the play-to-stop ratio is higher in speech mode although not significantly so, and playback speed is significantly higher.) Speech-enabled input allowed the participants to focus their eyes and attention better on the playing 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 accuracy of the speech recognizer. To validate that their speech was correctly 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 and where 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.

As reported earlier, some participants also wanted the ability to undo an annotation, particularly in speech mode, where they were more likely to say something quickly and incorrectly. In contrast, manual mode, which required pausing video playback and clicking through the annotation palette, gave the participants more time to think about the correct annotation.

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 it 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-enabled input.

Conclusion

The results of our experiment suggest that adding speech to a multi-windowed video annotation system such as the MSIIA might enable people to identify images in videos *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 having speech available and felt that it made it both easier and faster to annotate the images because the second modality kept their eyes free and hands free so that they could focus more on the task. When speech input was available, participants chose speech over manual input to make all annotations although some participants used the mouse to select tabs.

This formative study indicates that speech-enabled input, as a second modality, could potentially 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.

Acknowledgments

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

Notes

[1] Beatrice Oshika, Personal communication, January 22, 2002.

[2] An anonymous reviewer suggests that participants might have found the task easier to do without speech because it allowed them to avoid making a decision on which input modality to use.

[3] The MSIIA system does provide the ability to delete annotations. However, the experimenters found that including this capability would have added an extra layer of unnecessary complexity to both the task and the associated training of naïve users.

References

Christian, K., Kules, B., Shneiderman, B., and Youssef, A. (2000). A comparison of voice controlled and mouse controlled web browsing. *Proceedings of the Fourth International ACM Conference on Assistive Technologies*, Arlington, VA, pp. 72–79.

Cohen, P. (1992). The role of natural language in a multimodal interface. *Proceedings of the Fifth Annual ACM Symposium on User Interface Software and Technology: UIST'92*, Monterey, CA, pp. 143–149.

DARPA Augmented Cognition Program (2001). Available at http://www.darpa.mil/ito/research/ac/index.html.

Grasso, M., Ebert, D., and Finin, T. (1998). The integrality of speech in multimodal interfaces. *ACM Transactions on Computer-Human Interaction*, *5*(4):303–325.

Hansen, S. (1997). MSIIA hunts predator in Bosnia. *The Edge, The MITRE Advanced Technology Newsletter*, *1*(1). Available at http://www.mitre.org/pubs/edge/march_97/second.htm.

Karat, C.-M., Halverson, C., Horn, D., and Karat, J. (1999). 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, pp. 568–575.

Mellor, B., Baber, C., and Tunley, C. (1996). Evaluating automatic speech recognition as a component of a multi-input device human-computer interface. *Proceedings of the Fourth International Conference on Spoken Language Processing: ICSLP 96*, Philadelphia, PA, pp. 1668–1671.

Nuance Communications Inc. (2002). Available at http://www.nuance.com.

Oviatt, S. (1994). Interface techniques for minimizing disfluent input to spoken language systems. *Proceedings of Conference on Human Factors in Computing Systems: CHI '94*, Boston, pp. 205–210.

Oviatt, S. (1996). Multimodal interfaces for dynamic interactive maps. *Proceedings of Conference on Human Factors in Computing Systems: CHI '96*, New York, pp. 95–102.

Oviatt, S. (2000). Taming recognition errors with a multimodal interface. *Communications of the ACM*, *43*(9):45–51.

Oviatt, S. and Cohen, P. (2000). Multimodal interfaces that process what comes naturally. *Communications of the ACM*, *43*(3):45–53.

Polson, M. and Friedman, A. (1988). Task-sharing within and between hemispheres: A multiple-resources approach. *Human Factors*, *30*:633–643.

Rosenfeld, R., Olsen, D., and Rudnicky, A. (2001). Universal speech interfaces. *Interactions*, pp. 34–44.

Rudnicky, A. (1993). Mode preference in a simple data-retrieval task. *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*, Amsterdam, The Netherlands, pp. 71–72.

Rudnicky, A., Hauptmann, A., and Lee, K. (1994). Survey of current speech technology. *Communications of the ACM, 37*(3).

Shneiderman, B. (1983). Direct manipulation: A step beyond programming languages. *Computer*, *16*(8):57–69.

Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM*, *43*(9):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., and Whittaker, S. (2001). DARPA communicator dialog travel planning systems: The June 2000 data collection. *Proceedings of EUROSPEECH 2001*, Aalborg, Denmark, pp. 1371–1376.

Waterworth, J. and Talbot, M. (1987). *Speech and Language-Based Interaction with Machines*. New York: John Wiley & Sons, pp. 54–56.

Wickens, C. and Hollands, J. (1988). *Engineering Psychology and Human Performance*. 3rd ed., Eaglewood Cliffs, NJ: Prentice-Hall, p. 451.