The MSIIA Experiment:
Reducing Cognitive Load through Speech-Enabled Input

Laurie Damianos, Dan Loehr,
Carl Burke, Steve Hansen, Michael Viszmeg
The MITRE Corporation

MITRE performed an exploratory study to examine the effects of speech-enabled input on the Multi-Source Intelligence Integration and Analysis (MSIIA) system in performing an 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 the system to allow for hands-free input of annotation identification and tagging in order to examine the effect on efficiency, quality, task success, and user satisfaction.

As part of a project supporting the DARPA Augmented Cognition program, we hypothesized that speech recognition can be a cognition-enabling technology, by reducing the cognitive load of instrument manipulation and freeing up cognitive resources for the task at hand. In particular, we set out to test the following hypotheses: People can annotate images *faster* and *better* with the MSIIA augmented by speech, and people prefer speech-enabled input to manual input for such a cognitive task.

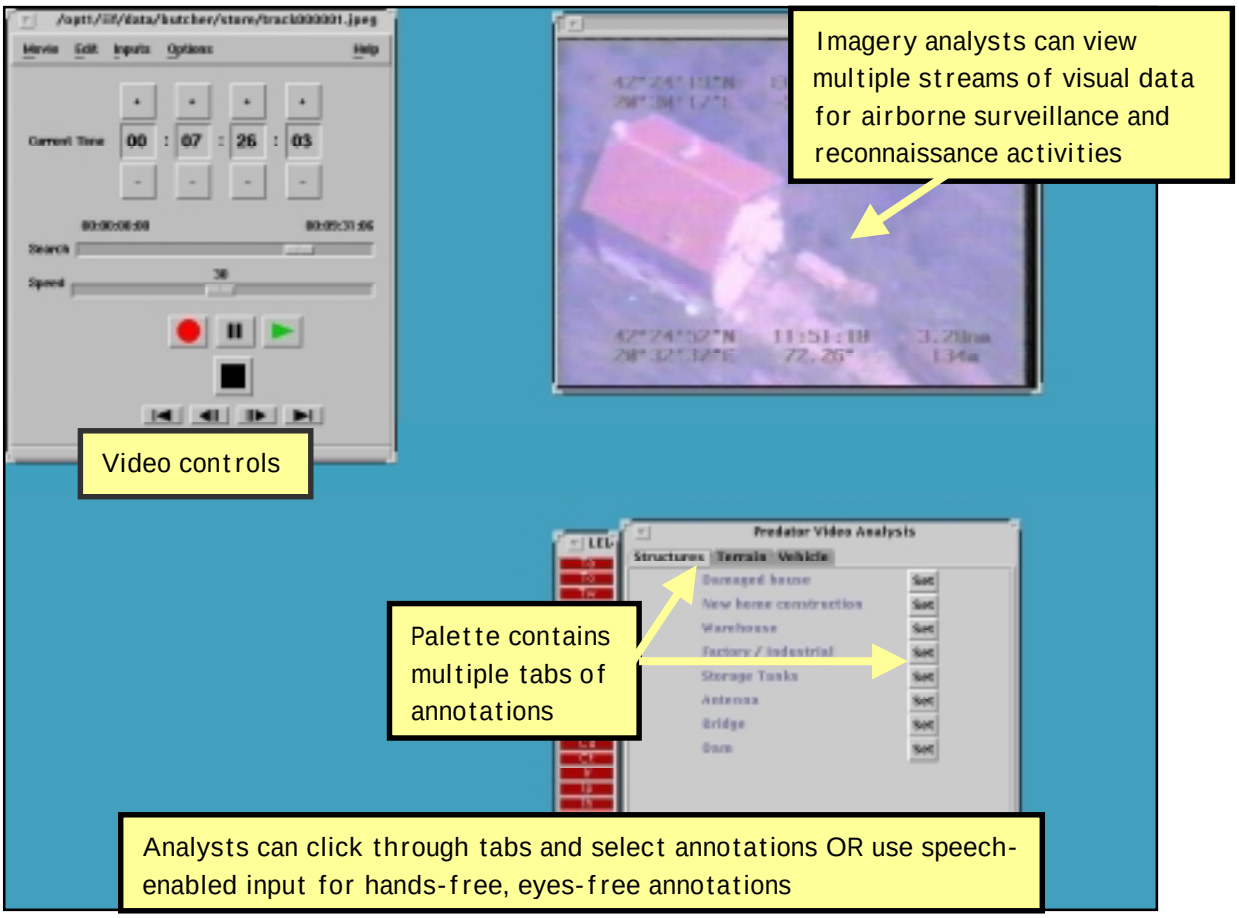

Figure 1 Participant view of the MSIIA system as configured for experiment

We designed a within-subjects, counterbalanced experiment in which each participant was 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 (~10 minutes each) and more content-ful 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, we alternated the order in which the two segments were administered.

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. They were told to annotate each of the identified objects with the appropriate tag from the Annotation Palette. Guidelines for identifying objects were also provided.

For the experiment sessions, we used a dedicated 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 all in the same place. During the session, an experimenter loaded each of the video segments for the participant.

The speech-enabled input component consisted 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 was used to communicate between the speech recognizer and both the Annotation Palette and a feedback GUI, a small window which provided minimal indication of the state of the speech agent. This allowed subjects to select items on the Annotation Palette via speech only.

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.[1] 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, associated metrics, and definitions is shown in Table 1.

[1] 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*, EUROSPEECH 2001.

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 video stream)
	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 1 Metrics

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 in the video clips. Several participants noted that the second modality was very effective in reducing the necessity to navigate controls and in allowing them to focus more on the task. Quantitative results show that people did annotate images *faster* with speech (shorter time on task, faster video playback speed, and significantly fewer stops/pauses while annotating). (See Table 2, below.) However, people did not annotate *better* with speech (precision was lower and recall was significantly lower). (See Table 3.) 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.

Category	Metric	Relationship of means	μ_{manual}	μ_{speech}	Significance	Stdev
Efficiency	Time on task	**μ_{manual}** > μ_{speech}	24.38 min	23.31 min	*None*	
	Image ID (playback speed)	μ_{manual} < **μ_{speech}**	7.51 fps	15.36 fps	0.01	0.17
	Annotation (play/stop)	μ_{manual} < **μ_{speech}**	0.09	0.34	*None*	

Table 2 Efficiency-related results supporting sub-hypothesis 1: *People annotate images faster with speech.*

Category	Metric	Relationship of means	μ_{manual}	μ_{speech}	Significance	Stdev
Quality	Task outcome (precision)	**μ_{manual}** > μ_{speech}	0.36	0.31	0.05	0.04
Task success	Task completion (recall)	**μ_{manual}** > μ_{speech}	0.84	0.81	*None*	

Table 3 Results on quality and task success which do not support sub-hypothesis 2: *People annotate images better with speech.*

This small, formative study has provided feedback for further development of the MSIIA system augmented with speech-enabled input, as our results show that speech-enabled input may lead to improved performance of expert domain users on more complicated tasks. This exploratory work indicates we have not yet fully tested our hypothesis that speech recognition can be a cognition-enabling technology. Towards this goal, refinements for future experimentation include improving speech feedback and annotation correction mechanisms, adding speech to more controls, increasing the complexity of the annotation tag set, and either using real imagery analysts or providing more domain training.