**2004 Command and Control Research and Technology Symposium**
The Power of Information Age Concepts and Technologies

**TOPIC:** C2 Human Factors Engineering

**Paper Title: The Requirements and Applications of Speech Recognition Technology for Voice Activated Command and Control in the Tactical Military Environment**

Mr. Lockwood Reed
US Army CERDEC
Fort Monmouth, NJ 07703
AMSRD-CER-C2-SS
732-427-2559/FAX 732-532-0134
Lockwood.Reed@us.army.mil

## Abstract

The US Army Communications/Electronics Research, Development and Engineering Center (CERDEC) at Fort Monmouth, NJ has been at the forefront of the research and development of speech recognition technologies for the tactical military environment for over twenty years. This includes the development of specialized techniques and technology to provide reliable performance in high noise environments. Additionally, unique to this technology is the ability to recognize whispered speech that is barely audible at one meter. The whispered speech recognition capability permits voice activation of C2 equipment during covert operations, such as urban house-to-house and room-to-room fighting, which requires the soldiers to maintain both hands weapons and an eyes-alert posture.

This paper will presents a discussion of the unique and specialized requirements for a militarized speech recognizer, as well as the tactical advantages of militarized speech recognition technology as it could be applied in several C2 applications and environments. Additionally, this paper will present the results of a comparison study, which was performed between a custom military speech recognition technology and various manual input modalities, including keyboard and trackball, for activating a selected C2 application. The results of this paper demonstrate a clear superiority of continuous speech recognition over discrete speech recognition in both metrics, and a tradeoff of task execution speed for error rate for continuous speech recognition verses manual input.

**Background: Initial research into speech recognition technology**

The CERDEC C2 Directorate's (then known as the Avionics Research and Development AVRADA C3 Division) first foray into speech recognition technology began in 1979 when this author attended the second meeting of the DOD sponsored Interactive Speech Technical Advisory Committee (ISTAC) [this organization has since become a sub-group of the DOD Human Factors Technical Committee]. The first task was to utilize the extensive analysis capabilities of the C2D audio laboratory to evaluate the effectiveness of COTS speech recognition technology for command and control applications in the Army Aviation environment. Tactical operations of Army helicopters require them to fly NAP-of-the-earth [below treetops at up to 60 Knots]. This requires extreme concentration, with the aviators maintaining their hands on the flight controls at all times and an eyes-out-of-cockpit posture. As one can imagine, manually controlling on-board avionics, while trying not to fly into a tree, can be quite challenging. Once all the available COTS speech recognition technology was evaluated and the deficiencies identified, specifications and requirements could be written. One critical requirement was for operation in high noise ranging from 103-107dBA, for the Blackhawk type helicopter and 110dBA for the Apache helicopter, the two most commonly used aircraft. Other aircraft, such as the CH-47 Chinook and the Heavy-lift Helicopter produced sound levels of 115dBA and 123dBA respectively. The unassisted COTS technology failed to perform reliably at sound levels as low as 80dBA, which presented quite a technological challenge. Various techniques were employed to improve the performance of COTS

technology, including numerous subtractive front-end filtering techniques, with little success. The early COTS technology all used the same basic approach - pattern recognition. Digital patterns for each vocabulary word were generated and used as templates and matched for best fit against the input stream. Trying a radically different technique, we exploited the pattern-based approach, and rather then subtracting the environmental noise we added environmental noise to the template during the template enrollment phase. Initially we simply had test subjects enroll in the target environment: recognition accuracy jumped from the low 70 percentile to over 95%. With a performance improvement like that we knew we were on to something. These results were reported a subsequent ISTAC meeting, and various member organizations confirmed the results. Early on we knew that this approach had it's limitations: first and foremost it was too stressful on the test subjects to enroll in high noise environments; second the recognition technology became too environment dependent. The Air Force group at the Rome Air Development Center (RADC) experimented with reducing the level of the enrollment environment[1], while normal operation was evaluated at normal environmental levels. The RADC work demonstrated that very little performance degradation was noticed for approximately 20 dB differences in environmental levels from enrollment to operation. However, a problem still existed if the spectral content of the environment changed significantly during operation. Helicopter environments are relatively stationary: the environment does not spectrally vary enough during various flight profiles to affect recognition performance. [Extensive environmental data collection was obtained on various aircraft in various flight profiles to verify the affect on recognition performance] However, noise environment of military track vehicles is not nearly as stationary as the helicopter environment, and these variations must be taken into consideration. For our own work we wanted to eliminate the user from having to enroll in the environment altogether, and we wanted to eliminate environmental dependency. Our approach to eliminating the noise during the enrollment session (at least as far as the test subject was concerned) was to electrically mix environmental noise into the test subject input stream to the target speech recognizer. The RADC work indicated that the precise signal-to-noise ratio should not be too critical, and indeed, through electrical mixing, we obtained results comparable to actual acoustical mixing of signal and noise (i.e. test subject in the environment). This also assuaged fears that the Lombard-Effect[2] would affect the results. The Lombard-Effect is simply that people tend to speak louder in high noise, which can affect the spectral content of their voices. Our results indicated that within the parameters of our experiments, the Lombard-Effect did not significantly affect recognizer performance. Achieving our second goal, the elimination of environmental dependence was going to somewhat more difficult. To achieve environment independence, the recognition technology would need to adapt to environmental changes on the fly, during operational use. Up until now we were able to perform all our experimentation and evaluation without the need to modify the recognition algorithms, to move to the next level, namely environmental independence, we needed access to the recognition algorithms. Rather then "re-invent the wheel" and devise our own recognition algorithms, we opted for letting contracts to companies who, according to our evaluations, had the most promising technology and who were willing to address the military environment. [Most companies are focused on the larger consumer market, and refuse to divert engineering resources to military problems, which don't

return a large “bang for their research buck”]. As it will become evident, our current technology represents the culmination of several years of investment. The technology provides; reliable performance in noise levels up to 115dBA does not require user enrollment in the target environment, will adapt to changing noise environments, and can be configured to permit the user to whisper, speak normally or shout commands.

**Discussion: The evolution of our speech recognition technology**

*Laying the ground work*

The development of our current speech recognition/response system did not occur over night: as the saying goes “Rome wasn’t built in a day”. The technology evolved over years of tech-base development. Beginning in the late 70’s a major effort was undertaken to determine the effectiveness of Army aircraft communication systems, from the radios to the intercom system. The CERDEC (then Electronics Command (ECOM)) launched a six-month data collection exercise at Fort Hood Texas. The goal of the exercise was to evaluate the condition of the communications systems on as many aircraft, and aircraft types as possible. The evaluation included several UH-1 (utility helicopter), OH-58 (observation helicopter), AH-1 (attack helicopter) and OV-1 (fixed wing reconnaissance) aircraft. During the first phase of the data collect, the communication systems of each aircraft were removed and tested for compliance with specifications and re-installed. The second phase, which became the most importance to the Interactive Speech Technology program, was the collection of in-flight sound recording. Each aircraft was flown through various pre-planned maneuvers (i.e. level flight, NAP-of-the-earth, ground-effect and out-of-ground-effect hover for the helicopters, etc.) while recording the ambient sound environment, and simultaneously recording over the interphone systems while subjects read various test passages (i.e. Rainbow passage, Modified Diagnostic Rhyme Test, etc.). This data became the basis for the evaluations of various speech recognition technologies.

*Noise Processing*

It became evident early on that, while overall sound level adversely affected recognition performance, the effect was not linear with increasing sound level. When the sound reached a certain critical level for a given recognition system performance degraded rapidly and then failed to be usable completely. The recognition systems were found to be far more sensitive to changes in the spectral content of the noise environment, then simply to changes in the sound intensity. It is believed that this relationship of performance to the sound characteristics accounts for the improved performance obtained by the additive techniques over subtractive techniques. Most subtractive techniques tested employed fixed filtering which were intended to exploit specific spectral spikes that occur in most aircraft: but these systems could not adequately track the spectral changes as the aircraft sequenced through various maneuvers. The first iterations of the additive technique suffered from the same limitations, but as the technology evolved it became possible to continuously sample the ambient environment and dynamically adjust the internal noise model within milliseconds before use: permitting the additive approach to track the variations of the ambient environment. Subsequent dynamic versions of the

subtractive technique were experimented with, but the dynamic additive technique maintained a significant performance enhancement.

The character of the noise can significantly impact recognizer performance. Our current technology was designed for environments in which the noise is at least quasi-stationary (the noise is predictable and does not vary too rapidly). Although it has demonstrated good performance in moderate levels of impulsive noise (noise spikes), as might occur adjacent to weapons fire, it was not designed to function in that environment per say. However, our group has studied the problem and believes we have solutions, which would permit operation even during weapons fire by the soldier-user. However these technologies are only under development and require further investment if they are to be implemented.

*Gain Management*

The next technical hurdle to overcome was the sensitivity of the recognizer technology to speaker dynamics and microphone placement. The problem is exacerbated when noise cancellation microphones are used. Most noise cancellation microphones are second order differential devices: they employ a diaphragm configuration in which sound is applied to both sides. Ideally, the ambient sound, which is considered 'far-field', immerses the microphone, exerting equal sound pressure on both sides of the diaphragm, and the diaphragm remains motionless: thus producing no electrical output for the ambient component of the overall signal. However, when the microphone is placed close to the speaker lips, the speech sound is considered 'near-field', and the pressure impacts only one side of the diaphragm: thus the microphone produces an electrical output proportional to the speech component of the overall signal. It is this differential characteristic of the noise cancellation microphone which produces a second order effect in signal level as the speaker distance from microphone changes. As the distance from the speaker to the microphone increases, the speaker's component begins to transition from 'near-field' to 'far-field', and the speakers own signal begins to be cancelled by the microphone. Therefore the total signal reduction is the summation of the cancellation effect plus the normal drop-off in signal with distance.

Normally signal variation due to speaker dynamics and microphone placement can be corrected with some form of automatic gain control (AGC), however common AGC implementations will dramatically degrade recognizer performance. The common AGC degrades performance by destroying the correlation in signal continuity from sample to sample as it adjusts its gain in response to the fluctuations of input signal level. The common AGC is a non-deterministic device, in that there is no information provided to the recognizer, which it can use to generate the inverse of the process. Because of the problems of speaker dynamics and microphone placement, some form of AGC was essential: but it would have to be a deterministic AGC. The implementation employed in our recognizer is deterministic: it is an integrated component of the recognition algorithm. Future implementations of the recognizer will employ technology, which will accommodate the full range of speaker dynamics, eliminating the need for an AGC.

*Recognizer Activation*

One of the difficulties of implementing any of the current forms of speech recognition technology is informing the recognition system as to when it should be processing speech input (i.e. when the user is addressing the recognizer and when the user is simply talking on a communications system, or speaking to adjacent individuals). One of the easiest and most effective means to eliminate confusion is to utilize a simple press-to-talk (PTT) switch. However in some applications it is either undesirable or unworkable to utilize a PTT. We developed an alternate approach, which we utilized in our Systems Test-bed for Avionics Research (STAR) aircraft. We implemented one of our speech recognition technologies into the Airborne Digital Avionics System (ADAS) – a networked system to control the on-board avionics via a MIL-STD-1553 bus. In the STAR aircraft, as in most military aircraft, there exists a PTT mounted on the cyclic flight control. The pilot selects the communication system via the intercom panel and can then activate the selected system via the cyclic PTT. It would have been awkward to require the pilot to select the speech recognition system (SRS) before each use, and then be required to re-select a particular communications system. In the helicopter environment the pilot or copilot can either communicate on a radio system or the intercom, therefore at all other times they would not normally have any reason to be speaking. Our solution was to provide a "press-to-off" function, utilizing the existing cyclic PTT. The SRS interpreted the pressing of the PTT as an indication to 'stand down', as the user was speaking to the communications system, thus avoiding the problem of the SRS misinterpreting conversation intended only for the communications system as commands. While the press-to-off function avoided the problem of the SRS interpreting speech intended for the communications system as commands, there still was the possibility that pilots, might vocalize to themselves. Therefore, in addition to the press-to-off function, we configured the SRS with a word-switch, or activation utterance. The SRS would continually monitor the pilot's microphone listening for a command phrase starting with the activation utterance. When the pilot or copilot spoke an acceptable phrase the SRS would output the appropriate commands to ADAS and effect control over the referenced subsystem. A typical command might be "Vic on, hydraulic backup pump on, disengage". In this example "Vic on" was the activation utterance, which must precede all command phrases. In addition to reducing the chances of false activations of the SRS, the activation utterance also provided a means to affect control of side-tone. In high noise environments it is necessary for the user to wear a headset, as hearing protection and means to monitor the communications system. When a headset encloses users ears, their own voice is heard primarily through a bone conduction path, which results in muffling the sound of the users voice (loss of high frequencies, and reducing the sound level). A solution to the muffling effects of the headset is to electrically inject the users voice into the headset: this is known as side-tone. The side-tone not only prevents the muffling effect of the bone conduction path, but amplifying the users voice prevents the user from shouting, which will change the spectral content of the voice, reducing recognition performance and fatiguing the user's voice. Although the SRS is continually monitoring the user's microphone, continually providing side-tone could produce hearing damage over extended exposure. Therefore, when the SRS detects the activation utterance, the SRS turns on the side-tone for the remainder of the command phrase, until the word

"disengage" is detected, terminating the recognition sequence and the side-tone. At all other times the communications system provides the side-tone as necessary.

Error correction and prompting is a further consideration. For the STAR SRS implementation we devised an effective and novel means of user feedback and error correction. If the user completed the example command phrase, the system would respond with: "backup pump on". However if the user paused in mid-phrase, for example saying: "Vic on, hydraulic". After a short delay the system would respond with the last thing recognized: "hydraulic", minimizing the prompting. If, however, the system did not understand something in the command phrase, such as the word "backup" in the example: without delay the SRS would respond with "hydraulic", immediately informing the user that it did not recognize anything after "hydraulic". At this point the user would only need to repeat the unrecognized portion of the command phrase: "backup pump on". The same process would occur for an unrecognized word anywhere in the command phrase.

*Whispered/Shouted Speech*

Anticipating the need for utilizing speech recognition during covert operations, such as urban room-to-room fighting, we have demonstrated a means of configuring our current technology to respond reliably to whispered speech. The initial implementation did not require any modifications to the current architecture, although modification of certain database files would optimize enrollment and possibly even improve the already acceptable performance. The current approach involves simply generating an alternate set of enrollment templates for whispered speech. Our current technology permits multiple template sets to be active simultaneously. Therefore normal and whispered template sets can exist side by side, allowing the user either mode of operation. Additionally shouted speech is handled in a similar manner. While this can provide an impressive demonstration, it is not the ultimate solution we envision. We have devised a far more sophisticated approach, which does not require alternate template sets. Unfortunately, as with most technological advances, the implementation of the more sophisticated approach is awaiting addition investment, as it will require modification and addition to the current architecture – but it is eminently achievable.

*Multiple Speaker Confusion*

Speech recognition has application to many diverse environments, however each application environment generally engenders unique requirements. In addition to requiring most of the capabilities of our current tactical SRS, the mobile command post imposes an additional requirement: the need to avoid multiple speaker confusion. In the generally cramped quarters of a mobile command post, users will be shoulder-to-shoulder, engaged with their respective applications. Under these circumstances, cross talk from an adjacent user is inevitable. In addition, many times there is a preference to use some form of mounted 'boom-microphone', as opposed to a headset microphone, which can only exacerbate the problem. The solution I have devised actually exploits the geometry of the problem and turns it into an asset: by codifying the problem as a signal flow diagram. The concept has been reviewed, by signal processing specialists and found to be a workable solution to the problem. We are currently seeking support for the

development of this technology from our Tactical Operations Center group, as they have expressed interest in our SRS technology.

**SRS Comparative Testing**

The bottom line question for Speech Recognition, or any technology: is there any benefit to the user? In an effort to get an initial answer to the question, a comparison test was performed at Fort Benning, Ga. The test comprised a command and control task, comparing two forms of voice activation and manual activation. The two forms of voice activation were "isolated (or discrete)" word recognition (words are spoken in isolation, requiring distinct pauses in between: no co-articulation) and "continuous speech" recognition. For clarification "continuous speech" was originally intended to mean the user could speak continuous, co-articulated, phrases, with optionally interspersed pauses. Unfortunately, some manufactures took the name "continuous speech" literally, forcing the user the complete a phrase without any pauses: pauses would cause misrecognition. In an effort to clarify the situation, this writer coined the term "natural speech" to represent the capability of a recognition system to correctly recognize continuous speech, with optional, arbitrarily inserted pauses: the original intent of "continuous speech". All references in this document to "continuous speech" can be read as "natural speech".

*Test Configuration*
The test platform consisted of a V2 LC Unit (transportable computer comprising a main unit: housing the CPU, Display, Hard-drive, and accessory cards and a keyboard and trackball unit) (**Figure 1**). The test software was an application called Brigade and Below, Command and Control (B2C2), a Force XXI Battle Command Brigade and Below (FBCB2) application pre-cursor. The test task was to perform the Call-for-Fire B2C2 messaging sequence. **Figure 2** is a composite of data entry pages 1 and 2, for a Call-for-Fire message. The following is the test script: **Create Call-for-Fire**, **Type-of-Fire Immediate-Suppression**, **Enemy-Description Armors**, **Enemy-Location Map** (move cursor to location on map) **Here**, **Enemy-Size Five**, **Page 2**, **Target Number Alpha Five One Three Four**, **Type-of-Munition High Energy**, **Method-of-Control When-Ready**, **Send-Report**. For manual entry the subject followed the same order as the speech input, but utilized the trackball and keyboard to enter the information into the corresponding fields on the display screen. Eighteen subjects were evaluated. Each subject's speech pattern was enrolled and each subject was trained on how to manually and vocally enter the test scripts. Each subject was give sufficient time to practice both manual and vocal message entry, to minimize the training curve. The test consisted of each subject entering a scripted message three times for each input modality (Manual, Continuous Speech and Isolated Word). It is important to note that this evaluation favored the manual entry, as it was performed in a static laboratory environment. Due to limited time and resources a dynamic test in a moving vehicle was not possible. This test was considered a worst-case comparison for speech recognition technology: simulating the user stopping the vehicle before entering data. A test moderator carefully recorded the time to complete each task run and the errors made. In the manual mode, any attempt which failed to select a target "radio-button", including the number of times it had to be repeated, or any mistyping was considered an error. In voice mode, any misrecognition

or lack of response by the SRS or any misspeaking by the test subject was considered an error. The users corrected errors by simply repeating the command.

*Test Results*

A summary of the results appears in **Figures 3 & 4**. **Figure 3** shows the results of three time comparisons: Isolated Word vs. Continuous Speech, Manual vs. Continuous Speech, and Isolated Word vs. Manual. The results of **Figure 3** exclude the trials containing errors, thus eliminating the additional time it would take to correct the errors. As indicated the same tasks took 83% longer to perform by isolated word recognition as compared to continuous speech; manual mode took 92% longer to perform the same task as compared to continuous speech; and isolate word recognition was only 6% faster then manual mode. The results of **Figure 3** are based on 913 manual operations, 850 isolated utterances (words) and 904 continuous utterances (words). **Figure 4** show the results for the same modal comparisons, with the exception that the time to correct errors is included. The results shown in **Figure 4** demonstrate that even factoring in the errors (manual 5, isolated 68, and continuous 14) manual mode still requires 74% more time to complete a task then continuous speech; isolated word required 101% more time to complete a task, and now took 19% longer then manual to complete a task. It is interesting to note the greater number of errors during isolate word recognition (68) as compared to continuous speech (14). At first this may seem counter intuitive, as one would expect the recognizer to be more accurate because of the silence delineations separating the words. However, if one considers the human as part of the system, and how unnatural it is to speak in an isolated manner, the rise in errors becomes obvious: no matter how much training was allowed the users found it impossible to maintain a completely isolate mode of speech and continually reverted to continuous speech during portions of a given trial. The isolated word recognizer was unable to recognize continuous speech, and provided no response, which was scored as an error. Additionally, this tendency toward high error rates for isolate word recognition would only be exacerbated under stressful conditions. In addition to the quantifiable information collected, the test subjects were asked to provide subjective scores to the following questions: On a scale of 1 to 5, where 1 is easiest and 5 is hardest; How easy was it to use (continuous recognition, isolated recognition, and manual entry)? The average score for the 18 participants breaks down as follows: continuous (1.3), isolated (2.5), and manual (2.3). The participants were also asked a related question, which attempted to ascertain how confident the subject felt with the respective entry modalities: On a scale of 1 to 5, where 5 is the most comfortable and 1 is the least; how comfortable did you feel with (continuous recognition, isolated recognition, and manual entry)? The average score for the 18 participants breaks down as follows: continuous (3.7), isolated (2.5), and manual (3.1). It is interesting to note that, contrary to popular belief continuous speech recognition was on par with manual entry in this static test, which favored manual input. Experiments performed by other agencies, including the Air Force[3], have shown that speech is relatively insensitive to movement, as compared to manual operation. Given the certain increase in manual entry error rate for a moving vehicle, it is not unreasonable to predict that the scores would shift, further favoring continuous speech recognition over manual entry in a dynamic environment.

## Conclusions

Since 9/11 the world has become quite a different place. Our military faces new and formidable challenges. The rapid pace and success of "Iraqi Freedom" lends credence to the effectiveness of highly mobile forces, and confirms the need for command-on-the-move technologies. In addition, as evidenced in Iraq and Afghanistan, the battles will be fought from street-to-street, house-to-house, room-to-room and cave-to-cave. A soldier can become a statistic "in a heartbeat" if is he is even momentarily distracted from maintaining a hands-on-weapon, eyes-alert posture, through having to manually interact with some tactical system. Additionally, soldiers will need the capability to interact with these tactical systems, while running and firing. Speech recognition technology has demonstrated the capability to provide hands-free, eyes-free activation of tactical system. Further it has also demonstrated its ability to operate under harsh and high noise battlefield environments. While no system can claim to operate flawless, under all battlefield conditions, there is sufficient evidence to conclude that the current state-of-the-art tactical SRS technology, can fulfill 90% of the current needs, to enable faster more intuitive Soldier/machine interaction, resulting in increased task accuracy, reduced task time, and ultimately yielding greater survivability and lethality. Additional, technology has been identified which, for a minimum investment, can improve the current technology to pickup most of the remaining 10%.

### *Into The Future*

As good as our current technology is, there is still one speech recognition system which is 'head and shoulders' above any available commercial or military technology – 'the human'. However, years of research in human cognition, and brain physiology have yielded many clues that I believe can lead to a radically new approach to speech recognition. Technologies exist that, when fully developed, will form the basis of this new technology, which I have christened Advanced Cognitive Interactive Speech Technology (ACIST). Currently the core technologies are not controlled by any one entity. My approach has been to establish a consortium of organizations from industry, academia and the military. The 'core group' has been formed and we have had our initial meetings to begin the structuring of a program and identify potential resources. The goal of the ACIST initiative is to develop an interactive speech technology, which will equal, and in some instances surpass human performance under the same environmental conditions. The bold objective of ACIST is to engender the same level of user confidence in machine speech recognition that has heretofore been reserved for human-to-human communication.

## References


1. "Speech Enhancement for Improved Recognition" By Michael Heffron, Rome Air Development Center, Minutes of the Voice Interactive Systems SUBTAG, 7 March 1983.

2. J.C. Junqua and Yolanda Angelade, "Acoustic and perceptual studies of Lombard speech: application to isolated-words automatic speech recognition", Proc. ICASSP 90, 841-844, Albuquerque, NM, April 1990.

3. “AFTI/F-16 Voice Command Systems: Status Report” By Major Stephen F. Gray, Edwards Air Force Base, Minutes of the Voice Interactive Systems SUBTAG, 7 March 1983.

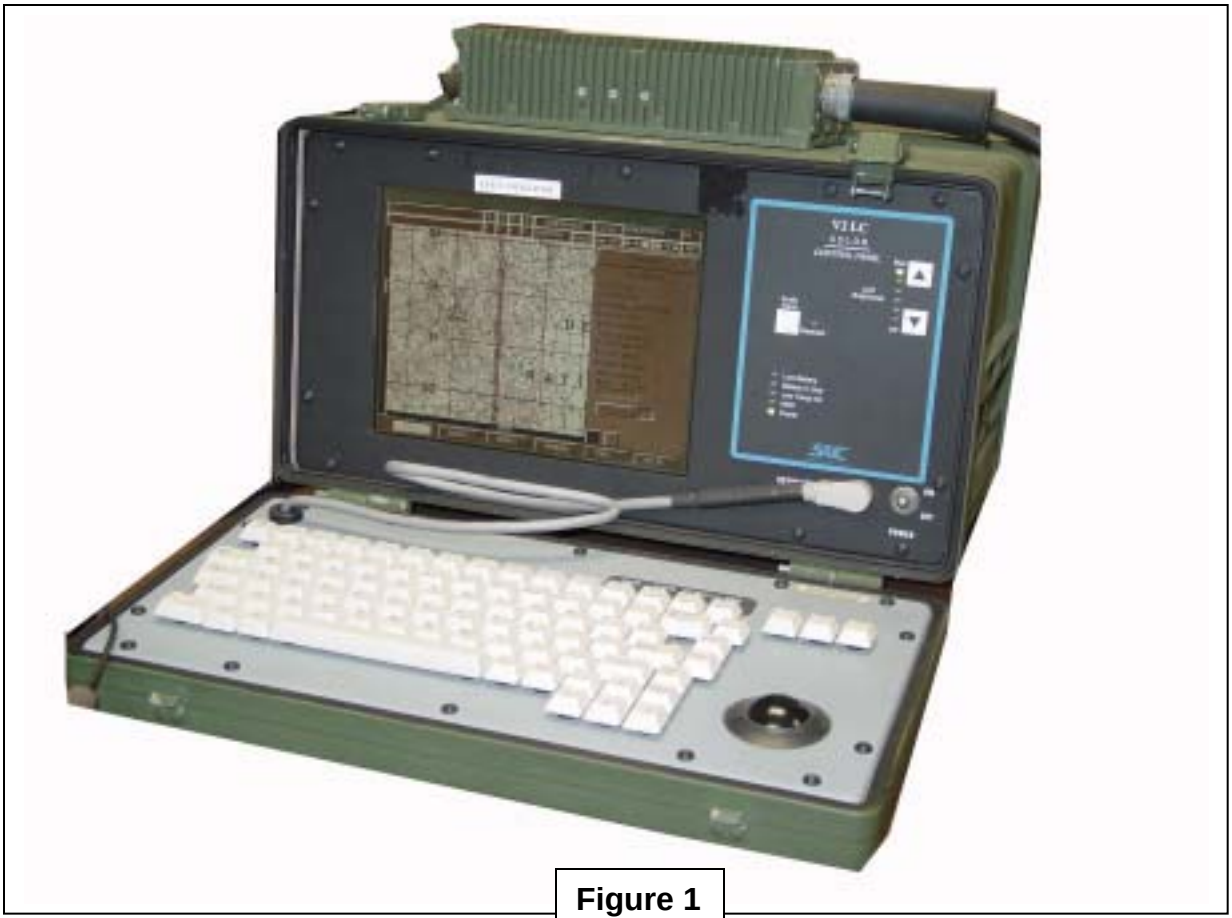

Figure 1

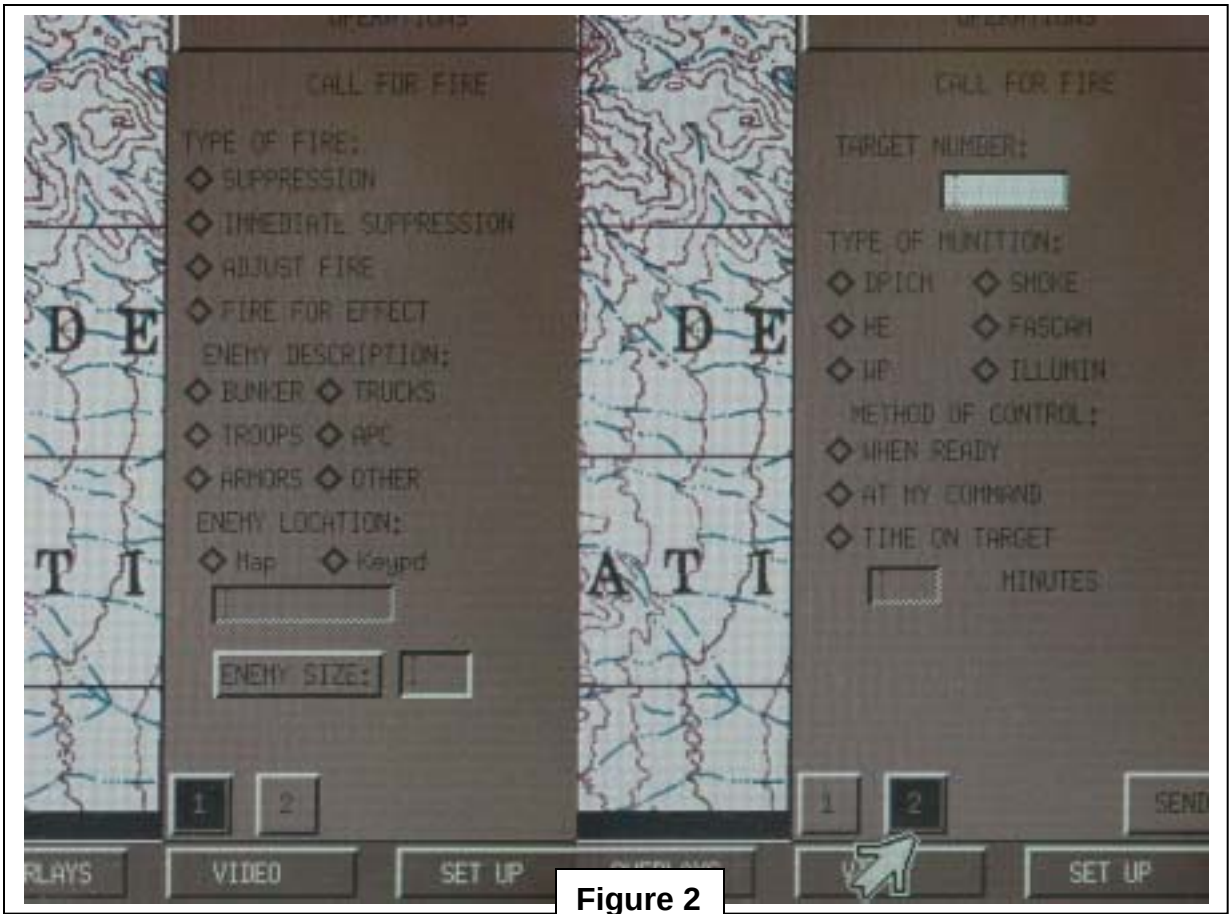


Figure 2

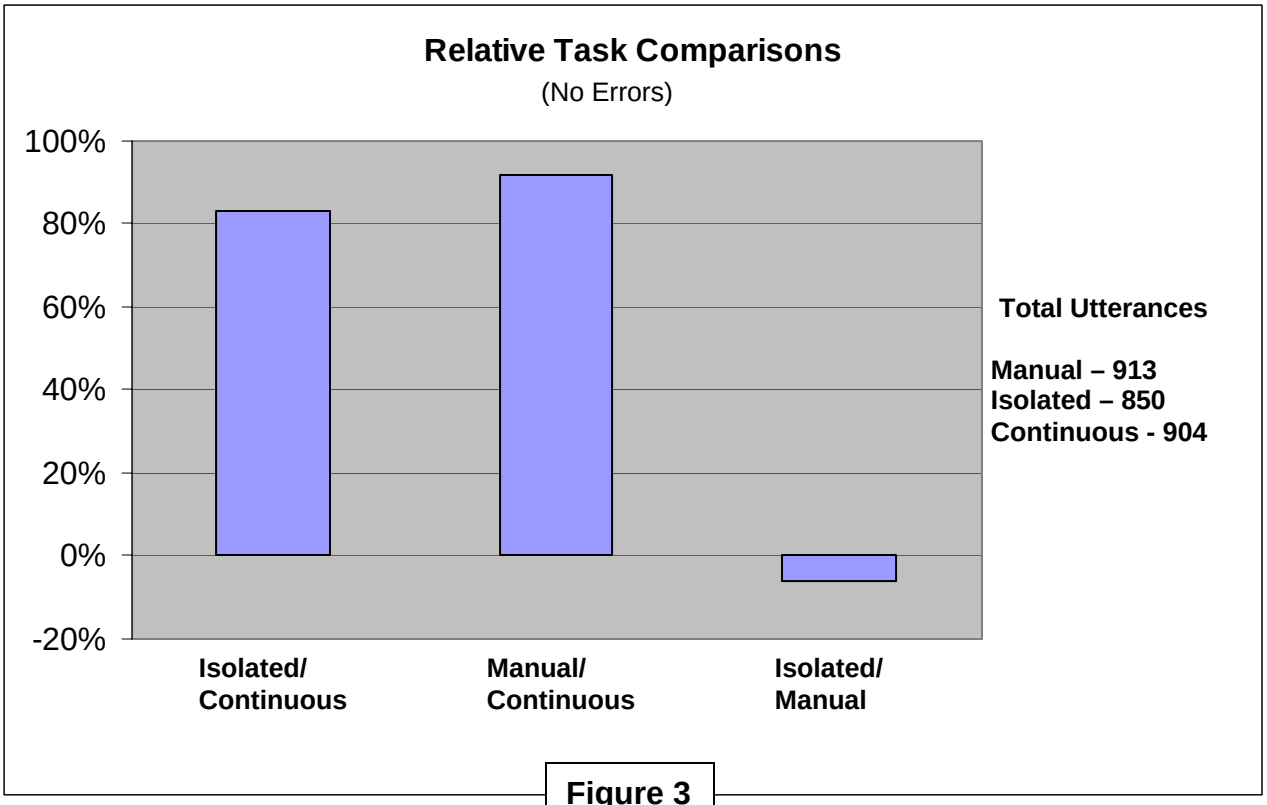


**Figure 3**

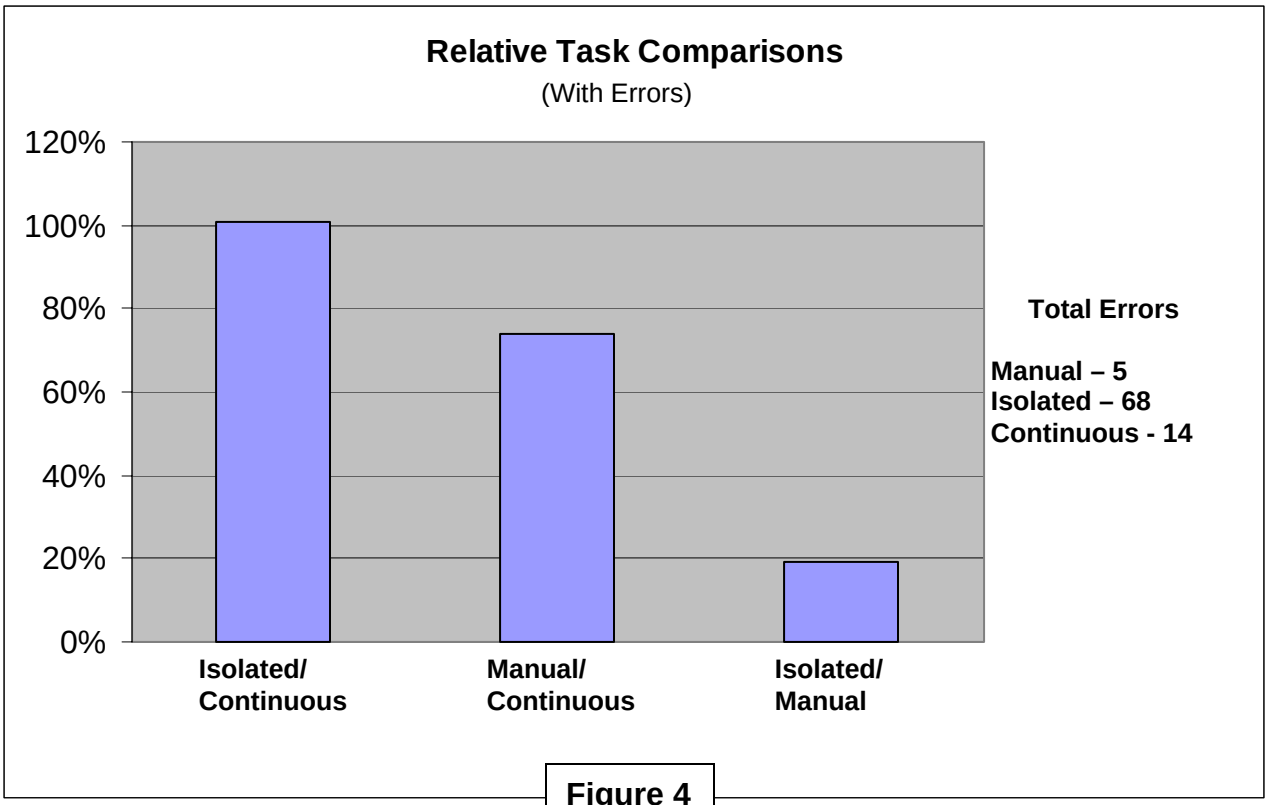


**Figure 4**