Performance Evaluation of DARPA's Advanced Soldier Sensor Information Systems and Technology (ASSIST) Program

Soldiers are often asked to perform missions that can last many hours and be under great stress. After a mission is complete, the soldiers are asked to provide a report describing the most important events that happened. This report is used to gather intelligence about the environment and local/foreign personnel to allow for more informed planning for future missions. Soldiers usually provide this report based solely on their memory, still pictures, and Global Positioning Systems (GPS) data from handheld units. Due to the extreme stress that they encounter, there are undoubtedly many instances in which important information is missed and not available for the planning of future missions.

The ASSIST (Advanced Soldier Sensor Information System and Technology) program addressed this challenge by instrumenting soldiers with a suite of sensors that they can wear directly on their uniforms. During the mission, sensors continuously captured various aspects of what is going on around the soldier. When soldiers returned from their mission, the sensor data was run through a series of Artificial Intelligence (AI)-based software systems that indexed the data and created an electronic chronicle of the events that happened while the ASSIST system was recording. With this information, soldiers could give more accurate reports without relying solely on their memory.

The AI-based algorithms had enabled numerous functionalities, including:

- "Image/Video Data Analysis Capabilities"
    - Object Detection / Image Classification – the ability to recognize and identify objects (e.g., vehicles, people, license plates) through analysis of video, imagery, and/or related data sources.
    - Arabic Text Translation – the ability to detect, recognize and translate written Arabic text (e.g., in imagery data).
    - Change Detection – the ability to identify changes over time in related data sources (e.g., identify differences in imagery of the same location at different times)

- "Audio Data Analysis Capabilities"
    - Sound Recognition / Speech Recognition – the ability to identify sound events (e.g., explosions, gunshots, vehicles) and recognize speech (e.g., keyword spotting, foreign language identification) in audio data.
    - Shooter Localization / Shooter Classification – the ability to identify gunshots in the surroundings (e.g., through analysis of audio data), including the type of weapon producing the shots, and the location of the shooter.

- "Soldier Activity Data Analysis Capabilities"
    - Soldier State Identification / Soldier Localization – the ability to identify a soldier's path of movement around an environment and characterize the actions taken by the soldier (e.g., running, walking, climbing stairs)

For AI systems such as these (often termed autonomous or intelligent systems) to be successful, they must be comprehensively and quantitatively evaluated to ensure that they will function appropriately and as expected in a wartime environment. The National Institute of Standards and Technology (NIST) was tasked with evaluating these AI systems based on three metrics:

1. The accuracy of object/event/activity identification and labeling,
2. The ability to improve classification performance through learning, and
3. The utility of the system in enhancing operational effectiveness.

NIST developed a two-part test methodology to produce its performance measures. Metrics 1 and 2 were evaluated through component and system-level technical performance evaluations and metric 3 was evaluated through system-level utility assessments. The technical performance evaluations were designed to measure the progressive development of ASSIST system technical capabilities; and the utility assessments were designed to predict the impact these technologies will have on warfighter performance in a variety of missions and job functions. In specifying the detailed procedures for each type of evaluation, NIST attempted to define evaluation strategies that would provide a reasonable level of difficulty for system and soldier performance.

At the component level, the ASSIST systems were broken into components that implemented particular capabilities. For example, to test the Arabic translation capability, the system was split into an Arabic text identification component, an Arabic text extraction component (to localize individual text characters), and a text translation component. Each aspect was tested separately to assess their individual impact on the system.

At the system level, each ASSIST system was evaluated as a black box, where the overall performance of the system was evaluated, independent of the performance of the individual components. A single score was provided for the overall system, which represented the capability of the system to accomplish the overall task.

Also at the system level, a test was performed to assess the utility of the system in enhancing operational effectiveness for after-mission reporting. All the systems that were evaluated under this program were in relatively early stages of development, therefore a formative evaluation approach was appropriate. NIST was particularly interested in assessing utility of the system to the war fighter. As such, the evaluation team was concerned with impacts on both their processes and products. To reflect this perspective, user-centered metrics were employed. NIST attempted to identify metrics that would help answer such questions as: What information do infantry soldiers want and/or need after completing a mission in the field? How well are information needs met, both from the soldiers' perspective and the S2's (Staff 2 - Intelligence Officer) perspective? What were the ASSIST contributions to mission reporting with respect to user-stated information needs?

The Aberdeen Test Center Military Operations in Urban Terrain (MOUT) site at the Aberdeen Test Center in Aberdeen, Maryland was chosen to carry out this evaluation. The site was chosen for the following reasons:

- Ground truth – Aberdeen was able to provide ground truth at selected locations to 2 cm accuracy. This allowed the evaluation team to gain a very good representation of what

actually happened in the environment, thus allowing the team to provide a good benchmark against which to compare the system output.

- Realism – the MOUT site contained approximately 20 buildings set up to resemble an Iraqi village.
- Infrastructure for testing – the site was instrumented with numerous cameras (both indoors and outdoors) to allow the evaluation team to gain a better understanding of the surroundings during testing.
- Soldier availability – the site was able to provide a small squad of active duty soldiers to participate in the evaluation.

To further facilitate an operationally-relevant test environment, the MOUT site was augmented with objects, persons, and background sounds, whose placement and behavior were scripted. The purpose was to provide an environment that would exercise the different ASSIST systems' capabilities as they detected, identified and/or captured various types of information. NIST included the following elements in the utility assessments: foreign language speech detection and classification, Arabic text detection and recognition, detection of shots fired and vehicular sounds, classification of soldier states and tracking their locations (both inside and outside of buildings), identifying objects of interest including vehicles, buildings, people, etc. The soldiers' actions were not scripted as they moved through each exercise because the tests required the soldiers to act according to their training and experience.