Technology Evaluations and Performance Metrics for Soldier-Worn Sensors for ASSIST

Brian A. Weiss, Craig Schlenoff, Michael Shneier, and Ann Virts
National Institute of Standards and Technology (NIST)
100 Bureau Drive, Stop 8230
Gaithersburg, MD, USA
{brian.weiss}, {craig.schlenoff}, {michael.shneier}, {ann.virts} @nist.gov

***Abstract*—The DARPA-funded Advanced Soldier Sensor Information Systems and Technology (ASSIST) project is aimed at developing soldier-worn sensors and software to increase a soldier's battlefield awareness during missions, provide them with data collection tools to augment their mission reporting capabilities following their field operations, and supply additional information to intelligence officers to enhance planning for future missions. The NIST-led Independent Evaluation Team is responsible for evaluating the ASSIST technologies developed by the Task 2 research teams. This paper discusses the overall Task 2 technologies of image/video, audio, and soldier activity data analysis capabilities with each participating research team's technologies presented at deeper levels. After understanding the technologies, the five elemental tests (Arabic text translation, object detection/image recognition, shooter localization, soldier state/localization, and sound/speech recognition) are designed and implemented with metrics for the Baseline and Final Phase I Evaluations.**

***Keywords*:** *DARPA, ASSIST, soldier-worn sensors, evaluations, performance metrics, elemental tests, object detection, image classification, shooter localization, sound recognition, speech recognition, soldier state, soldier localization*

I. Introduction

The Advanced Soldier Sensor Information System and Technology (ASSIST) is a Defense Advanced Research Projects Agency (DARPA) supported research and development program. This effort intends to advance and exploit soldier-worn sensors to increase soldiers' battlefield awareness during humanitarian and combat missions, provide enhanced data collection tools to augment mission reporting capabilities following field operations, and supply additional information to intelligence officers to improve mission planning all within military operations in urban terrain (MOUT) environments. [1] This program is split into two efforts:

- Task 1 emphasizes active information capture and voice annotations exploitation hardware. The resulting products will be prototype wearable capture units and the supporting software for processing, logging and retrieval.
- Task 2 stresses passive collection and automated activity/object recognition. The results from this task will be the software and tools that will undergo system integration in later program phases.

The National Institute of Standards and Technology (NIST), along with its subcontractors (Aptima, Inc. and DCS Corporation), was funded to serve as the Independent Evaluation Team (IET) for Task 2, Phase I. In this role, NIST was responsible for:

- Understanding the Task 2 research technologies
- Devising a testing approach for these technologies
- Identifying a MOUT site to evaluate these technologies
- Designing and executing the tests
- Developing performance metrics to analyze the data
- Documenting the test results

Section II presents the tested technologies along with the specific capabilities of the participating research teams, both at the 6-month (Baseline) and at the 12-month (Final Phase I) evaluations. Section III presents the elemental tests including enhancements that were made between the two evaluations and the performance metrics developed to evaluate the tested technologies. Section IV summarizes the paper.

II. Technologies For Evaluation

Task 2 involves developing a range of data capture, analysis, and display technologies. These capabilities are broken down into three data type categories. Within each data type, several "technology elements" are applied to organize, process and present that data. Some of the key technology elements being applied in the ASSIST program are listed below:

"Image/Video Data Analysis Capabilities"

- Object detection/image classification – the ability to recognize objects (e.g. vehicles, people, etc.) through analysis of video, imagery, and/or related data
- Arabic text translation – the ability to detect, recognize, and translate written Arabic text through imagery analysis

"Audio Data Analysis Capabilities"

- Sound recognition/speech recognition – the ability to identify sound events (e.g. gunshots, vehicles, etc.) and recognize speech (e.g. keyword spotting, foreign language identification, etc.) in audio data
- Shooter localization – the ability to identify gunshots in the environment (e.g. through analysis of audio data), including the type of weapon producing those 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 within an environment and characterize their actions (e.g. running, walking, climbing stairs, etc.)

Presently, there is no single integrated system within the ASSIST program. Instead, several universities and corporations have collaborated to form "research teams". Each organization is developing specific technologies with these components being gradually integrated as a "research team" system. The following subsections provide overviews of the specific "research team" technologies and describe the progression of each team's system from the Baseline Evaluation to the Final Phase I Evaluation.

A. IBM/MIT/Georgia Tech Team ASSIST Technologies

The IBM Team combines researchers from IBM, the Georgia Institute of Technology, and the Massachusetts Institute of Technology. The team's long-term vision for their ASSIST suite is a comprehensive system that captures, analyzes, organizes, and archives data for users (soldiers and intelligence officers) to review and search records to augment military reporting and mission planning. The IBM team's technologies include:

- Image classification – The presence of an object is detected based upon data from image and audio sensors and classified with one or more classes and subclasses. For the Baseline Evaluation, images were classified to contain the presence of *Outdoors*, *Indoors*, *Sky*, *Buildings*, *Vegetation*, *People*, *Weapons*, and *Vehicles*. The IBM team expanded their capabilities for the Final Phase I Evaluation to detect the presence of *Soldiers*, *Commotion*, *Vehicle_civil*, *Vehicle_military*, and *Cars* (in addition to their baseline capabilities).
- Object detection – Objects are detected and localized (a bounding box is created) to a specific region within an image. For the Baseline Evaluation, the IBM team detected and localized *Faces* and *License_plates*. During the Final Phase I Evaluation they could also detect and localize *Clothing_Color*.
- Sound recognition – Recorded audio from the environment that is classified as "non-speech sounds" is further classified into the following: sounds from a car and large truck (Baseline Evaluation) plus single gunshots, machine gunfire, explosions, light trucks, sedans, and transport vans (Final Phase I Evaluation).
- Speech recognition – Keyword extraction is performed on a soldier's speech (English). Keywords that are detected during the Baseline Evaluation include *insurgent*, *target*, *dead*, *shot*, *shots*, *suspicious*, *killed*, *kill*, *fire*, *incoming*, *contact*, *weapon*, *weapons*, *intelligence*, *Intel*, etc. Keywords that are added during the Final Phase I Evaluation include *update*, *A4 millimeter round*, *AK47*, *alpha*, *bravo*, *C4*, *frag-out*, *halt*, *IED*, *M16*, *RPG*, *SIT-REP*, and *tango*.
- Language identification – Capability to identify spoken English, German, Japanese, Mandarin, Spanish, and Hindi (Baseline Evaluation) and later Arabic and French (Final Phase I Evaluation).
- Soldier state identification – Capability to determine when a soldier is performing the following actions: standing, walking, running, driving, and lying down (Baseline Evaluation) along with opening doors, performing a situational assessment from cover, taking a knee, sitting, raising a weapon, shaking hands, crawling, going upstairs and going downstairs (Final Phase I Evaluation).

B. Sarnoff Team ASSIST Technologies

The Sarnoff ASSIST team is composed of three organizations: Sarnoff Corporation, Carnegie Mellon Institute, and Vanderbilt University. Each of these groups is focused on unique technologies that will not be integrated with one another during this phase of the project. This resulted in each organization being treated as a separate team. The following subsections discuss the technologies from these three groups.

1) Sarnoff Corporation's System: Sarnoff is developing an ASSIST system to support soldier localization and object detection. These technologies are discussed further:

- Object detection – An object is detected and localized (specified with bounding boxes) to a region within an image. Sarnoff was able to detect *Vehicles* and *People* at the baseline plus *Faces*, *Weapons*, and *Vehicle_type* for the final evaluation.
- Soldier localization – Capability of locating (in GPS coordinates) the ASSIST-wearer both indoors and outdoors. This capability did not change between the evaluations, rather the software algorithms were refined following the Baseline Evaluation.

2) Carnegie Mellon University's (CMU) System: CMU is developing technologies aimed at extracting and translating

Arabic text from images captured with a typical consumer-grade digital camera. Their technology operates in three stages:

- Arabic text is identified within an image through edge detection, layout analysis and search algorithms.
- Arabic text is extracted from an image using optical character recognition software.
- Arabic text is translated to English using statistical machine translation technology.

Again, there were no capability increases for this technology between the two evaluations, rather software refinements were made to improve each of the three technology stages.

3) Vanderbilt University's System: Vanderbilt University (also referred to as Vanderbilt) is developing a shooter localization technology that detects gunfire, determines bullet trajectory, localizes the shooter, classifies the bullet caliber and identifies the type of weapon being fired. Their current hardware suite consists of 10 acoustic sensors. The system's capabilities are below:

- Shot Localization – Determine bullet trajectories and shooter origins of short-range (≈30 meters) shots using single and multiple shooters along with determining trajectories of long-range (≈100 meters) shots (Baseline Evaluation) plus determining the trajectories of longer-range (200 meters to 300 meters) shots along with determining the trajectories and shooter origins of automatic fire at shorter ranges (Final Phase I Evaluation).
- Shot Classification – Classify shots from an M16, AK-47, and M107 (Baseline Evaluation) plus classifying shots from an M4, M240, and M249 (Final Phase I Evaluation).

C. University of Washington Team

The University of Washington (also referred to as Washington) team consists of the University of Washington, Intel Research Seattle, and Lupine Logic. This team's system is aimed at collecting soldier state data. Specifically, identifying whether a soldier is indoors, outdoors, riding in a vehicle, walking, running, standing, performing a situational assessment from cover, going upstairs and going downstairs at both evaluations. Again, no capability improvements were made between the two evaluations, rather software enhancements are made following Baseline Evaluation.

III. Elemental Tests

The IET developed a two-part test methodology to produce the following three metrics (as stated per DARPA):

- Measure the accuracy of object/event/activity identification and labeling
- Measure the system's ability to improve its classification performance through learning
- Measure the utility of the system in enhancing operational effectiveness

The first two metrics are evaluated through "elemental tests" while the last metric is evaluated through "vignette tests". [2] Elemental tests are developed to test the ASSIST technologies in an "idealistic" environment and allow a focused examination of the specific components. Vignette tests immerse the technologies in realistic military scenarios to assess the systems in more practical, fast-paced, stressed conditions. This paper focuses on the elemental tests.

The elemental tests afford the ability to modify specific variables in a controlled manner to measure the impact of those variables on technology performance within a MOUT environment. Five elemental tests are developed:

- Arabic text translation
- Object detection/image classification
- Shooter localization
- Soldier state/localization
- Sound/speech recognition

Each of these elemental tests is discussed in detail in the following subsections.

A. Arabic Text Translation

The Arabic text translation elemental test was specifically designed to evaluate CMU's ASSIST ability to detect, recognize, and translate Arabic signs. Again, this elemental test seeks to evaluate the three key Arabic text translation capabilities:

- Identify Arabic text in an image
- Extract Arabic text from an image
- Translate Arabic text to English text

1) Test Description – Baseline Evaluation: A single approach was taken in evaluating these three capabilities. This was accomplished by placing six Arabic signs in the MOUT environment and having CMU collect imagery data at two distances (near and far) and five angles (30°, 60°, 90°, 105°, and 135° with 90° being a straight-on view of the sign). Distances were based upon the letter-size of the specific signs with the near distance corresponding to approximately 50 pixels per height of the smallest letter in the sign and the far distance corresponding to approximately 30 pixels per height of the smallest letter of the sign when using CMU's consumer-grade camera. Signs were placed both indoors and outdoors. The location of each sign placed in the environment along with their associated data collection points were

measured with two-centimeter accuracy.

The test began with the researcher collecting images of the signs from the various distances and angles. The test then proceeded through three successive stages whereby each was evaluated:

- Sign detection (step 1) – The signs placed in the environment were used to evaluate the ability of the system to extract regions of text.
- Text extraction (step 2) – The regions extracted from the signs in step 1 were processed to extract and localize Arabic characters and words.
- Text translation (step 3) – The output from step 2 was fed into the translation component and the English output was evaluated both quantitatively and qualitatively by a native Arabic speaker.

2) Lessons Learned – Baseline Evaluation: The testing approach taken during the Baseline Evaluation where technology performance of one step is dependent upon the technology performance of a previous step (i.e. successful text extraction that is dependent upon successful sign detection) made it impossible to accurately test the system's text extraction and translation capabilities. The test approach was modified for the Final Phase I Evaluation where the three individual steps of the system were evaluated separately in addition to conducting an overall (step successive) evaluation.

3) Test Description – Final Phase I Evaluation: To enable a comparison with the baseline, three signs were placed in the environment at marked positions so that sets of images could be taken at the same angles and appropriate distances. Of the three signs, one is a sign that was used during the baseline and setup at its original location while the other two signs have never been used before. Images were captured of these signs and the data is put through the three-step process. This was the overall test that enables direct comparison of the system's capabilities between the two evaluations. This process also allowed for the individual evaluation of sign detection (step 1).

In addition, text extraction (step 2) was separately evaluated by feeding Arabic letters and words in "ideal" fonts directly into the optical character recognition (OCR) program. In order to test the ability of the text extraction to deal with more complex backgrounds, two signs with textured backgrounds were used, two signs were input with an image as well as text, one sign included English numbers as well as Arabic text, and two signs had colored backgrounds.

The text translation component (step 3) of the system was tested in a similar way. Fifteen text files were created containing Arabic text taken from real signs. The files were encoded in the required format and input into the program one at a time. Once again, this provided an ideal situation for the translation system, with no misspelled words, no extra characters, and no missing characters.

4) Metrics for Evaluation: The following metrics are identified and used to evaluate this technology:

- Text rows correctly extracted (%)
- Non-text regions found/false alarms (%)
- Characters correctly localized (%)
- Arabic words correct (%)
- English words correct (%)

B. Object Detection/Image Classification

The object detection/image classification elemental test evaluated the following capabilities of the IBM team's and Sarnoff's ASSIST systems:

- Presence detection of objects and states within imagery (IBM)
- Localized detection of objects within specific regions of imagery (IBM, Sarnoff)

1) Test Description – Baseline Evaluation: Prior to the evaluation, the ≈45 meters squared, courtyard (containing 10-single story and two double-story buildings) was setup with objects. Each building contained multiple doors and windows and is populated with various amounts of furniture (e.g. chairs, desks, tables, etc.). Approximately 50 waypoints (using two-centimeter accurate, differential GPS and surveying equipment) were marked to include a range of indoor, outdoor, ground-level, and upper-story locations (including positions in front of doorways, windows, etc.). These waypoints were used to denote imagery collection locations for the ASSIST-wearer, and the locations of additional objects to be placed in the environment.

Additional objects in the environment include vehicles (both civilian and military) with license plates (both US and Iraqi), people (soldiers and civilians dressed in simulated middle-eastern attire), weapons (both US military and foreign that were either carried by people or placed within the environment), Arabic signs, tires (both stacked vertically and resting against buildings), trash piles, barrels, sandbag piles, etc. The following variables were taken into account when selecting the locations of objects and imagery viewpoints:

- Position of ASSIST-wearer
  - Ground level
  - Upper level
- Position of ASSIST-wearer relative to object(s)
  - Both indoors
  - ASSIST-wearer indoors with objects outdoors
  - ASSIST-wearer outdoors with objects

indoors
    - Both outdoors
- Object(s) orientation relative to ASSIST-wearer
    - Above, below, same level
    - Head-on, angled, side-view, rear-view
- Object distance relative to ASSIST-wearer
    - Near (<5 meters)
    - Mid-range (<20 meters)
    - Far (>20 meters)
- Object occlusion relative to ASSIST-wearer
    - Entirely visible
    - Partially occluded by other objects
- Background scene relative to object(s)
    - Objects viewed with other objects close behind them vs. far away
    - Objects viewed with objects behind them with similar colors vs. objects behind them with contrasting colors

Imagery was collected from 25 viewpoints that were distributed across 10 waypoints, most of which had multiple viewpoints at different orientations. Labeled doormats were placed at each waypoint to indicate the ASSIST-wearer's orientation for imagery collection. Each team collected a single image at each of the 25 viewpoints. The IET also collected imagery data from each viewpoint using its own consumer-grade, digital camera.

2) Lessons Learned – Baseline Evaluation: Several improvements were realized following the Baseline Evaluation. A greater quantity and diversity of objects (e.g. people in a wider range of attires, etc.) including clutter (e.g. wires hanging from buildings, more trash, etc.) should be added. The elemental test should also provide data collection points across a larger area of the MOUT. Another issue was that imagery collected from upper level locations allowed the ASSIST systems to capture data outside of the control area whereas ground locations only allowed imagery out to a very finite distance.

3) Test Description – Final Phase I Evaluation: This elemental test evolved to address the shortcomings of the first evaluation. First, the test area was expanded so that data collection viewpoints were added in both the courtyard and the warlord compound (≈100 meters x ≈60 meters with three, single-story buildings and a double-story building). Overall object density and diversity was increased as more objects (specifically, people and vehicles) were added to the environment (additional GPS waypoints were surveyed). Data viewpoints were also modified so that imagery was only collected from ground level to better control the viewing area.

4) Metrics for Evaluation: The imagery data that each team captured with its ASSIST system was used as both data for experimental analysis and as ground-truth. If an object could be viewed within a team's image, then it was evaluated against the team's processed data (e.g. if a vehicle is visible in a team's image, then the team would be evaluated whether it could detect the vehicle or not). Likewise, if a human is not able to discern an object from viewing an image, then the team was not evaluated against that object. Output data includes:

- Positive identification (true positive) - an object was correctly identified
- Negative identification (false positive) – an object was identified that is not present
- Missed identification (false negative) – an object was not identified that is present
- Total instances of presence (total presence) – sum of positive identifications and missed identifications
- Total identifications – sum of positive identifications and negative identifications

The following metrics were applied based upon the output data:

- Positive identifications over total presence (%)
- Missed identifications over total presence (%)
- Positive identifications over total identifications (%)
- Negative identifications over total identifications (%)

C. Shooter Localization

The shooter localization elemental test evaluated the capabilities of Vanderbilt's ASSIST system to:

- Detect gunshots
- Calculate a bullet's trajectory
- Localize a shooter's origin
- Classify the caliber of bullet being fired
- Identify the specific weapon being fired

1) Test Description – Baseline Evaluation: This test was conducted at Aberdeen's outdoor firing range, due to restrictions against live fire at the MOUT site. A "zero line" and four firing lines (≈25 meter, ≈50 meter, ≈100 meter, and ≈200 meter) were marked on the range. The ASSIST system's acoustic sensors were placed on and behind the zero line, and randomly covered an area that was ≈30 meters squared. Five targets were set up behind the sensor region. Simple, wooden-walled structures (single-story and two-story) with windows were constructed at the firing lines and in the sensor region to simulate the buildings and obstructions that would be found in a MOUT environment, and to provide unique shooter positions through windows, next to walls, and on upper levels.

Five to seven shooter positions (both practice and test positions) were marked at each firing line. All positions on the firing range (sensors, targets, shooter positions and wall

corners) were localized to within two-centimeter accuracy using differential GPS. The following variables were considered in the placement of shooter positions:

- Shooter positioning relative to walls at the firing line
  - From a clearing
  - Next to a wall
  - From within a structure with the weapon's barrel protruding out of a window
- Obstructions between the firing line and sensor field
  - Positions obstructed by walls that could occlude a weapon's muzzle blast and/or shockwave from a subset of the sensors
  - Positions with clear line of sight to the sensors

A shot matrix was developed for 200+ shots with the following variables considered:

- Weapon and caliber [M16, M4, & M249 (5.56mm), AK47 & M240 (7.62mm), M107 (50 caliber)]
- Firing lines (≈25 m, ≈50 m, ≈100 m, ≈200 m)
- Shooter positions from the four firing lines
- Rounds per test (single shot vs. 3-round bursts)
- Number of shooters (single shooter vs. multiple shooters)
- Weapons fired by multiple shooters (same weapon vs. different weapon)
- Bullet trajectory (shots that crossed in between a majority of sensors vs. shots that passed by very few sensors near the perimeter)

Shots were fired at each of the four firing lines and data was collected.

2) Lessons Learned – Baseline Evaluation: Following the Baseline Evaluation, several enhancements were recognized that would improve the operational relevance and expand the complexity of this elemental test. First was that there is little operational relevance in testing from the ≈25 meter firing line. Additionally, shooters (particularly snipers) will typically fire from within structures where their weapon's barrel is not protruding out of a window/opening. Also, shooters will sometimes strafe up towards a target whereby they can see their bullets hit the ground in front of their target and adjust their trajectory accordingly.

3) Test Description – Final Phase I Evaluation: This later evaluation addressed all of the lessons learned from the Baseline Evaluation. First was the elimination of the ~25 m firing line and the addition of the ≈300 m firing line. Second was to add a shooter position at each firing line from within one of the wooden structures that forced the weapon barrel to be recessed at least 1 m to 2 m from a window. Lastly, targets (additional to those placed behind the sensors) were placed in front of the sensor region. The shot matrix was updated with ≈250 shots.

4) Metrics for Evaluation: The ASSIST system's output data was evaluated against the following three metric categories:

- Detection (broken down by firing line, shooter position, and weapon caliber plus variants to evaluate multiple shooter detections)
  - Shot detections over all shots fired (%)
  - Trajectory detections over all shots detected (%)
  - Shooter origin detections over all shots detected (%)
- Localization (broken down by firing line, shooter position, bullet caliber, and single shot vs. 3-round burst)
  - Shooter origin (m) – accuracy and precision
  - Trajectory angle (degrees) – accuracy and precision
  - Target crossing (m) – accuracy and precision
- Classification (broken down by firing line, specific weapon, and specific bullet caliber)
  - Correct shot detections by weapon (%)
  - Correct shot detections by bullet caliber (%)

D. Soldier/State Localization

The goal of the soldier state/localization elemental test was to evaluate the ASSIST systems' ability to localize a soldier within indoor and outdoor environments, and to characterize their actions. The IBM, Sarnoff, and Washington teams participated in this test, with each team outputting different information (see Section II for further detail).

1) Test Description – Baseline Evaluation: There were four test runs, each of which was performed twice. Each run exposed the system to a different level of difficulty for soldier state / localization identification. Each run required a soldier, shadowed by a researcher wearing the ASSIST system, to traverse a predefined path of waypoints in a scripted fashion. Run 1 was only outside in open areas. Run 2 was also outside but included some tight, GPS-restricted areas. Run 3 was both outside and inside, but did not force an elevation change. Run 4 was predominantly inside and traversed two floors of a building (one of the ground and the other elevated one story).

Approximately 60 waypoints were marked (including indoor, outdoor, ground-level and upper level points) with two-centimeter accuracy using differential GPS and surveying equipment. Poles were placed in cones at each waypoint. Signs attached to the poles indicated a letter for each waypoint in a run (e.g. A, then B, etc.), gave a brief description of the action to be performed at the waypoint and on the way to the

next waypoint (e.g. "lie down for 10 seconds then run", "go up stairs", etc.), and provided an arrow pointing to the next waypoint. The actions scripted were dictated by the superset of stated capabilities by all three teams' ASSIST systems.

Before the execution of each run, the soldiers and the researchers walked the path of the run to become familiar with the route and actions. During the run, three observers captured the time that the ASSIST-wearer reached the waypoints and performed the specified actions. Observers also noted any inconsistencies in the actual actions of the ASSIST-wearer relative to the scripts. This data allowed the IET to accurately capture ground truth and measure the ASSIST system's accuracy in localizing the ASSIST-wearer and identifying actions.

2) Lessons Learned – Baseline Evaluation: Several test concerns were noticed following this initial evaluation. Although each of the four runs was performed twice, the individual runs were relatively short in time and distance covered. Also, the range of actions was relatively limited.

3) Test Description – Final Phase I Evaluation: This elemental test was refined to address the concerns highlighted from the Baseline Evaluation. Instead of having each team perform the four original runs twice, the four original runs were performed in reverse and two additional runs were added (for a total of six runs). Performing the original four runs in reverse still provided a means of comparing data between the two evaluations. Run 5 involved a loop around a large portion of the MOUT complex, in which each action occurred for a longer period of time and distance. Run 6 (also run in a larger MOUT area) included much more driving and going up and down stairs. Also, each of the four original runs had some of their actions supplemented with more complex actions (e.g. raise weapon for 10 seconds, drag a sandbag, etc.). To account for these additional runs, more GPS waypoints were surveyed with the same two-centimeter accuracy.

4) Metrics for Evaluation: Soldier state accuracy was calculated by comparing ground truth times of ASSIST-wearer actions to the actions identified by each ASSIST system. All overlapping time periods were analyzed for correspondence. For example, if the ground truth showed that the ASSIST-wearer was walking from 0 seconds to 5 seconds and running from 5 seconds to 10 seconds, and the ASSIST system showed that the ASSIST-wearer was walking from 0 seconds to 7 seconds and running from 8 seconds to 10 seconds, the time periods were analyzed independently for correspondence. In this case, there would be a match from 0 seconds to 5 seconds and from 8 seconds to 10 seconds, with an incorrect detection from 5 seconds to 7 seconds. Specific state metrics include:

- Correctly identified movement vs. stationary (%)
- Correctly classified type of movement (%)
- Incorrectly classified type of movement (%)
- Unclassified soldier movements (%)
- % Correctly identified indoor vs. outdoor activity

Soldier localization accuracy was calculated by comparing the ground truth location of waypoints to locations returned by the ASSIST systems at specific times. Observers noted the exact time that the ASSIST-wearer reached each waypoint. These location and times were then compared to the data returned from the ASSIST system. To account for human error and non-exact clock time between systems a 4-second window (2 seconds before and 2 seconds after the exact time) were introduced when comparing the locations. The location returned by the ASSIST system that was closest to the ground truth location within this time window was used in the analysis. Specific localization metrics include:

- Accuracy (m) of mapping all soldier movement
- Accuracy (m) of mapping all outdoor movement
- Accuracy (m) of mapping all indoor movement

E. Sound / Speech Recognition

The goal of the sound recognition test was to evaluate the ASSIST system's ability to detect certain sounds in the environment. Since only the IBM team has the ASSIST capabilities to perform this type of test, the sounds and speech presented in this test are aligned with the team's technology.

1) Test Description – Baseline Evaluation: To conduct this elemental test, the following sound events were scripted to occur in the environment at specified times relative to the start of a given evaluation run:

- A soldier fired blank bullet rounds (5.56mm, 7.62mm, and 50 caliber)
- A soldier standing next to the ASSIST-wearer spoke one of ten text phrase which incorporated some combination of the team's stated-capability keywords
- A person in the environment either spoke or played a digital voice recording of people speaking the stated-capability languages
- Vehicles were driven past the ASSIST-wearer while either accelerating or decelerating

The variables for this elemental test were as follows:

- Distance between the sound source and the ASSIST-wearer
- Speakers were stationary or moving (e.g. a person speaking a language in the environment could be stationary, walking away from the ASSIST-wearer, or walking towards the ASSIST-wearer).
- The level of ambient noise that was in the environment. For this condition, ambient noise was

either low (i.e. no additional ambient noise was added) or high (i.e. ambient noise was produced by a generator located ~7m from the ASSIST-wearer).
- ASSIST-wearer was stationary or moving
- Overlapping vs. non-overlapping sounds. In non-overlapping runs, each sound event was separated by a few seconds. In overlapping runs, multiple sounds occurred in the same time segment (e.g. a gunshot, a person speaking a language, etc.).

There were five runs of increasing complexity with each run performed twice. During the early runs, there was little or no ambient noise, the ASSIST-wearer was stationary, and there were no overlapping sounds. During the later runs, there was a lot of ambient noise, the ASSIST-wearer was moving, there were overlapping sounds, and the sounds in the environment were moving to and from the ASSIST-wearer.

Ground truth locations of the ASSIST-wearer and the sounds in the environment were measured based upon known points. Before the test, the locations of certain points in the environment were mapped out to two-centimeter GPS accuracy. When stationary, the ASSIST-wearer remained at one of these specified points in the environment; when moving, the ASSIST-wearer moved between these points. Similarly, the scripted sounds were generated at these locations or moved between them.

2) Lessons Learned – Baseline Evaluation: Following this evaluation, improvements were sought. The only realization was that the five runs were conducted in the same open environment. This meant that the environmental acoustics (potential presence of echoes, etc.) was not considered to be a variable.

3) Test Description – Final Phase I Evaluation: This later evaluation added two runs, each in a different part of the MOUT site as compared to the original five runs. This allowed the environmental acoustics to become an evaluation variable. The sixth run was outdoors, in a more confined area; closely surrounded by concrete walls. The seventh run was predominantly indoors. Additional keywords were also added to the soldier-spoken texts based upon the team's additional capabilities.

4) Metrics for Evaluation: This evaluation can be broken down into the following categories: sounds and speech recognition. The metrics applied for sound recognition:

- Correctly classified all sounds (%)
- Incorrectly classified all sounds (%)
- Unclassified sounds (%)
- Correctly classified sounds (broken down by vehicles, gunshots, foreign languages) (%)
- Incorrectly classified sounds (broken down by vehicles, gunshots, foreign languages) (%)
- Unclassified sounds (broken down by vehicles, gunshots, foreign languages) (%)

The metrics applied for speech (keyword) recognition were:

- Correct keyword identifications (%)
- Missed keyword identifications (%)
- Incorrect keyword identifications (%)

IV. Conclusion

The IET successfully designed and implemented these five elemental tests for the DARPA ASSIST's Task 2, Phase I Baseline Evaluation and Final Phase I Evaluation. Metrics were consistently applied to the ASSIST teams' output elemental test data to achieve the DARPA-required high-level metrics:

- Measure the accuracy of object/event/activity identification and labeling (determined from data collected from each elemental test evaluation)
- Measure the system's ability to improve its classification performance through learning (demonstrated in comparing data between the baseline and final evaluations)

It is anticipated that the ASSIST program will continue for at least 3 more years and the NIST-led IET expects to continue to implement and improve upon its tests and metrics for future evaluations.

Acknowledgement

This work was supported by the Defense Advanced Research Projects Agency (DARPA) ASSIST program (POC. Mari Maeda).

Disclaimer

Certain commercial software and tools are identified in this paper in order to explain our research. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the software tools identified are necessarily the best available for the purpose.

References

[1] DARPA, "Advanced Soldier Sensor Information System and Technology (ASSIST) Proposer Information Pamphlet," http://www.darpa.mil/ipto/solicitations/open/04-38_PIP.htm, 2006.

[2] M. Steves "Utility Assessment of Soldier-Worn Sensors Systems for ASSIST", To appear Proceedings of Performance Metrics for Intelligent Systems Workshop, August 2006, Gaithersburg, MD.