ITEA Journal 2014; 35: 58-64

Performance Testing and Evaluation of Transformative Apps Devices

Anthony Downs[1], Lisa Fronczek, Emile Morse, Brian Weiss, Ian Bashor, Craig Schlenoff
National Institute of Standards and Technology Gaithersburg, Maryland

Transformative Apps (TransApps) is a Defense Advanced Research Projects Agency (DARPA) funded program whose goal is to develop a range of militarily-relevant software applications ("apps") to enhance the operational effectiveness of military personnel on (and off) the battlefield. A team from the National Institute of Standards and Technology (NIST) is responsible for designing and implementing the testing and evaluation methods for ~60 software apps for the Android™-powered smartphone platforms as well as software-hardware interaction. This paper focuses on the new test methods developed by the NIST team for comparing Android™-powered devices in consideration of becoming future TransApps devices. These test methods include Camera Usability, GPS Accuracy and Timing metrics, Compass Accuracy, and Display Usability. The results of these test methods allow the program to be more informed in the purchasing decisions for future TransApps devices.

Key words: Performance Evaluation, Metrics, Android™, Military, Functional Testing, Regression Testing, Usability

1. Introduction

The Transformative Apps (TransApps)[1] effort is a Defense Advanced Research Projects Agency (DARPA[2])-funded program that began in 2010 and is aimed at enhancing the warfighter's effectiveness on and off the battlefield. Specifically, the program is developing a flexible and secure suite of applications ("apps"), enabling direct end-user input, promoting quick fielding and updates, and leveraging pre-existing state-of-the-art commercial-off-the-shelf technology. To accomplish these goals, TransApps has focused its attention on developing a secure Android™ software platform, specialized application software, middleware and tools, a usable application portal, and flexible development processes. The program has made numerous achievements through November 2013 including developing over 60 apps for tactical users and fielding over 4000 handheld devices to warfighters in Afghanistan. Likewise, the program provided both first response and law enforcement personnel with over 150 devices to capture and share information in real-time to support the 2013 Presidential Inauguration and have continued to work with the first responder/law enforcement communities throughout the year for various efforts. The program has received overwhelmingly positive feedback from the warfighter community including leadership personnel. Soldiers have credited the device with not only enhancing their situational awareness, but also enabling them to be successful in dangerous situations. The program continues to actively field additional units in Afghanistan in addition to providing application updates to existing users.

Testing is a critical element of this program, assessing numerous facets. Personnel from the National Institute of Standards and Technology (NIST[3]) have been funded to serve as an independent evaluation team since early 2011 for the TransApps program. Since the program's inception, NIST has been responsible for assessing three key areas:

- Handheld applications
- Client-based applications
- Application MarketPlace

In the past year, the NIST evaluation team has created test methods for comparing potential future TransApps devices (handhelds and tablets) in various key areas such as displays, cameras, Global Positioning System (GPS) and compass. The purpose of these test methods is to provide direct comparisons of devices in specific areas allowing the sponsor to make a more informed choice about which device to use in the future. Likewise, the device test methods provide program leadership with foundational knowledge of key component performance prior to devices being converted to TransApps systems. The NIST evaluation team has extensive experience assessing advanced and emerging technologies; related prior work is discussed in

Section 2. Greater detail on the Transformative Apps program including an overview of the specific technologies tested and the assessments of handheld, client-based, and online marketplace applications can be found in Section 3. A discussion of the new comparative test methods follows in Section 4. Finally, the conclusion is presented and future work is discussed.

2. NIST Advanced Technology Assessment Efforts

Personnel from NIST's Engineering Laboratory (EL) and Information Technology Laboratory (ITL) have extensive experience in evaluating advanced and emerging technologies for the military. NIST led the independent evaluation teams in assessing the DARPA Advanced Soldier Sensor Information System and Technology (ASSIST) technologies (2004-2008)[2,3,4,5] and the DARPA Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) technologies (2006-2010)[6,7,8]. NIST developed the System, Component, and Operationally-Relevant Evaluation (SCORE) framework as the backbone of its test plan design and implementation of these technologies.

1.1. SCORE

NIST developed the SCORE framework, a unified set of criteria and software tools for defining a performance evaluation approach for complex intelligent systems. It provides an evaluation framework that assesses the technical performance of a system and its components through isolating and changing variables as well as capturing end-user utility of the system in realistic use-case environments. SCORE is built around the premise that, to get a comprehensive picture of how a system performs in its actual use-case environment, technical performance should be evaluated at the component and system levels.[9,10,11]

SCORE takes a tiered approach to measuring the performance of intelligent systems. At the lowest level, SCORE uses elemental tests to isolate specific components and then systematically modifies variables that could affect the performance of that component to determine those variables' impact. Typically, elemental tests are performed for each relevant component of the system. At the next level, the overall system is tested in a highly structured environment to understand how modifying specific variables impacts the overall system performance. Next, individual capabilities of the system are isolated and tested for both their technical performance and their utility using task tests. Lastly, the technology is immersed in a longer scenario that evokes typical situations and surroundings in which the end-user is asked to perform an overall mission or procedure in a highly-relevant environment which stresses the overall system's capabilities. Formal surveys and semi-structured interviews are used to assess the usefulness of the technology to the end-user.

SCORE is applicable to a wide range of technologies, from manufacturing to defense systems and its elements can be decoupled and customized based upon evaluation goals. It can evaluate technology through the stages of development from conceptual to full maturity, and combines results of targeted evaluations to produce an extensive picture of the capabilities and utility of a system.

3. Application Testing

DARPA selected NIST to be a primary evaluator of the TransApps technologies in 2010. Initially, NIST was tasked to assess the performance of the 1) handheld applications, 2) client-based applications, and 3) online application marketplace. As an independent, third-party evaluation team, NIST personnel presented unbiased and objective performance data to the DARPA sponsor enabling them to make informed decisions regarding application stability and field-worthiness. The NIST team facilitated interactions with developers to understand the scope and intent of the apps as well as embracing the voice of the end user population.

The NIST team leveraged the principles of SCORE to develop and implement test protocols and procedures to yield comprehensive performance assessments at multiple layers. This includes individual assessments of apps while isolated from other apps and global assessments of the apps and their interactions while operating on configurations expected during fielding. A summary of these testing efforts is discussed below. More detailed descriptions of this part of the NIST testing efforts can be found in other papers[12].

Handheld applications enable tactical mobile capabilities for the warfighters using the 60+ apps deployed on the devices. The NIST team tests these apps to make sure the needed functionality is present and working properly before it is sent to the end users locally and overseas. The NIST team employs a user-centered approach to testing handheld devices and apps. Since some of the target users (warfighters overseas) are not available for real-time testing, NIST gathers insights on use cases and ideas from briefings with recently returned soldiers and teleconferences with the technical and training liaisons for deployed units. With these gathered insights and ideas, NIST has developed typical workflows and use cases for the apps. Expert and heuristic reviews[13] are the primary day-to-day methods that

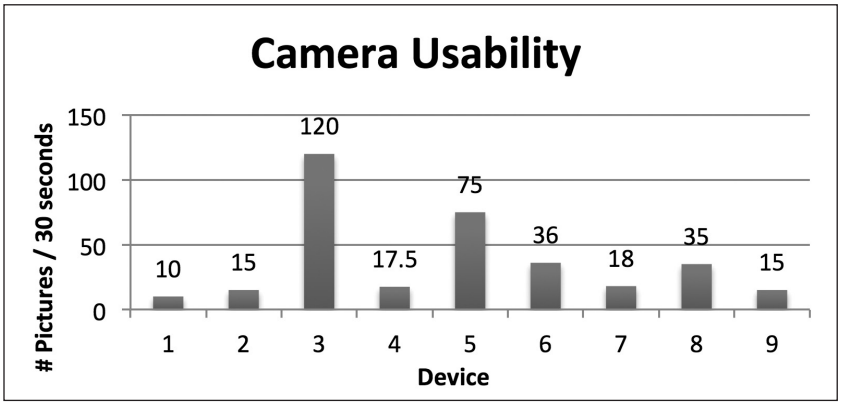

Figure 1: Camera Usability

the NIST team employs to evaluate apps. Periodically, the NIST team performs full regression testing to ensure all desired functionality is present.

Client-based applications consist of tools developed for encrypted laptops (clients) or servers to interface with the handheld devices. These apps are used pre-mission for data creation and transfer to handhelds and post-mission for transfer from handhelds and data analysis. Data downloads and uploads between handhelds and clients are achieved with the Sync Services and TransApps Maps (TA Maps). The Sync Services allow information to be transferred to the handhelds for the next mission and download data from the handhelds after the mission. TA Maps also has a powerful tool to create drawings, with a variety of tools ranging from simple lines, polygons and grids to specialized tactical graphics. Numbers and text can be added to drawings for use during tactical operations. The NIST team has a few client-specific testing practices; for data upload (client to handheld), the NIST team creates complex files (like drawings) and sets of files to be synced to the devices and for data download (handheld to client), all of the media and files collected/created during the mission is checked by the members who collected it to ensure that all transfers happened correctly.

MarketPlace, a web-based military app store, serves the needs of a broad range of users (military and civilian) allowing users to: 1) get the latest TransApps applications and imagery for their handhelds and client laptops, 2) learn more about the program, 3) access tutorials, and 4) interact with other users, support, developers, and project personnel. This site makes use of dashboard-style pages – easy to read, graphical, real-time interfaces that are dynamically created based on the user's access level. Since MarketPlace needs to serve many user groups, it is a complex website to evaluate. It is important to simulate the roles of the various users and carry out the same tasks in the same conditions as the actual users to provide more realistic results. After the more direct tests are run, it is important to compare the site with its intended usage. Does it meet the needs of the various users? Does the site restrict information to the user's access tier? Are user roles maintained across the site? Does the site support the technical team that outfits the handhelds both here in the US and abroad? The testers have access to special testing accounts with varying levels of permissions. Regression testing with these accounts ensures that the website provides the functionality needed even as new functions are added.

The NIST team reports its findings and testing results in the form of weekly reports and bug reports. The weekly reports are sent to the upper level program management, and include a list of show-stoppers and watch points, which are itemized and prioritized. The bug reports include bugs found in the apps, suggestions for new features, and ideas for improvements and are submitted via an online bug-reporting tool. In this way, the bugs, etc. are sent to the developers so they can be fixed. Through this process the NIST team learns what the intended and required functionality of the apps are as well as what to watch out for in further testing.

4. Device Testing Using Comparative Test Methods

The NIST evaluation team was tasked to develop test methods for specific key areas of possible future TransApps devices. These test methods and their results will allow the program sponsors to compare the devices available for purchase in these key areas and be more informed about the strong and weak points of any given devices. In each area the devices are analyzed, and the sponsor can apply the program's desired priorities to each area to decide what device(s) should be purchased. NIST took advantage of the expertise in creating repeatable and reproducible test methods using readily available materials in developing the test methods

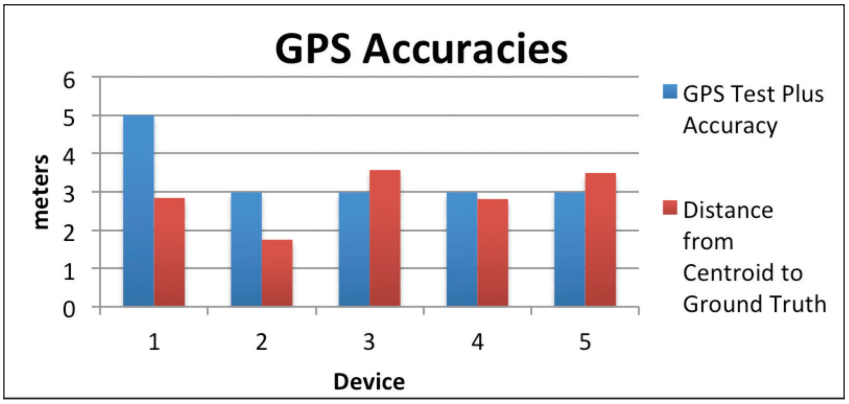

Figure 2: GPS Accuracies

described below. Samples of results for these test methods are shown in the sections below with devices numbered for anonymity.

4.1 Camera Usability

One of the most important uses people use cell phones for is as a camera, and this holds true for the TransApps users as well. The pictures and videos collected with the camera are used in numerous apps for photo profiles as well as being attachable to chat messages, and location based events. So, naturally this is something that is important for usability testing. For usability of the camera on a device, the test method calls for taking as many pictures as possible within a 30 second window of time. The stock camera is opened and then the shutter button is pressed as fast as possible for the 30 seconds. Then, the tester goes into the gallery to count the number of pictures that were taken. This procedure is run a number of times for each device to ensure the repeatability and reproducibility of the results. The average number of pictures able to be taken is reported for each device allowing the sponsor to compare. This provides a good comparison of the camera hardware in addition to the normal technical comparison of file size, resolution etc. for pictures. A sample of the results is shown with devices in Figure 1 showing a range of about 10 pictures (~ 1 pic/3 s) to 120 pictures (~ 4 pics/1 s). While some of these improvements in camera can be attributed to software tweaks, the results show what is possible for each device to do in hardware.

1.2. GPS Accuracy & Timing

GPS location is a heavily used function on TransApps devices and other military equipment. The GPS functionality is used on the TransApps devices for providing the user's location for pictures, videos, text notes, bearings to other points, routes, and many other things. The two main portions of the GPS functionality are GPS accuracy and the time required to obtain a GPS signal. On the NIST grounds there are a number of National Geodetic Survey (NGS) benchmarks that provide the NIST team with highly accurate positional reference (latitude, longitude, and elevation)[14]. These NGS benchmarks are available in a large number of other locations as well, allowing the test method to be reproducible in other locations. For GPS accuracy, a number of GPS readings are taken at the NGS locations at varying times of day, weather conditions, and across multiple days. These readings are then used to analyze the average accuracy and distances from the ground truth NGS points. The accuracy results are reported as radial distances based on latitude and longitude without elevation considerations. Sample results are shown in Figure 2 showing 4 devices with similar average accuracies and 1 worse, as well as the corresponding distances to ground truth ranging from 1.8 to 3.6 m.

For the time to obtain GPS signal, there is a well-known measure called time to first fix (TTFF) that is used by the industry to compare GPS functionality. While this is a widely used set of measures, it does not match perfectly with the use case for which TA program intends these devices. A modified time metric was determined to be the time to obtain a GPS accuracy of ~9.1 m, when the GPS signal is of sufficient quality to be used by the device to geo-locate media. This GPS accuracy is determined using a GPS recording app like GPS Test[15]. This metric is determined for all three start-up states of the GPS functionality: Hot Start (when the GPS has been on and connected to the GPS satellites within about an hour), Warm Start (no GPS for between a couple hours and a day), and Cold Start (no GPS for more than a couple of days). Each of these states will mean a differing amount of data needed from the satellites and thus a different time to get the GPS fix. These start states can also be artificially caused by using an application like NMEA Recorder[16] to remove the GPS

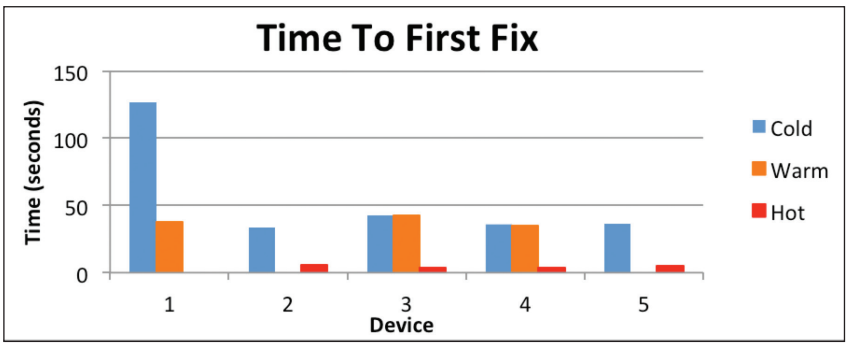

Figure 3: Time To First Fix

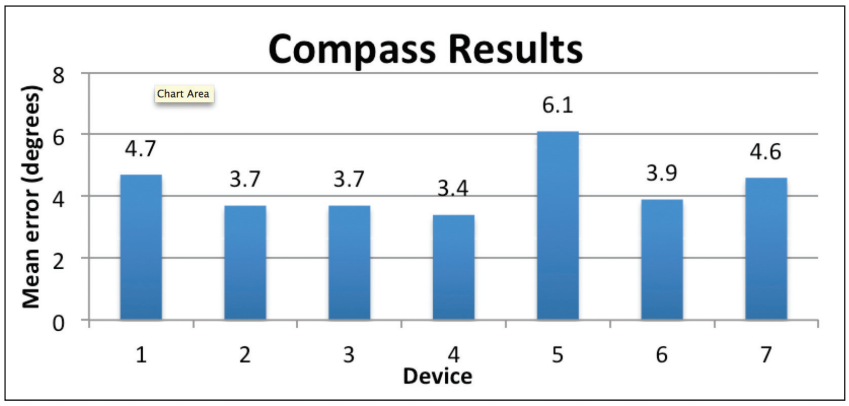

Figure 4: Compass Results

almanac and/or ephemeris data, effectively forcing the GPS functionality to re-obtain that data from the GPS satellites fresh. The times for each start state on each are recorded, in multiple weather and time conditions similar to the GPS Accuracy testing to ensure consistency among results. Sample results are shown in Figure 3 showing cold starts ranging from 120+ seconds on the worse end down to ~ 33 seconds on the better end and roughly equal warm start times and hot start times.

1.3. Compass Testing

A compass function is available in almost all of the stock Android™ devices that are sold and it can be useful to know in which direction the device is pointing. It is used in the TransApps devices to indicate direction within the camera and as part of a user's location icon within maps applications, in addition to being used for determining the bearing to other locations. For NIST's compass testing, a course is laid out using five points (a central point and four points in different directions), at least one of which should be a NGS point like described above. The four points should each be in the northwest, northeast, southeast, and southwest quadrants and should be between 25 and 125 m distant from the central point. To minimize potential bias in the results, the four points should not be exactly at any of the cardinal or ordinal directions. The four points should also be well defined and/or marked to be easily sighted from the central point. The tester stands at the central point, aims the device toward the other four points and records the compass readings. For completeness, each device is tested for the four points, in both portrait and landscape orientation, three times each for a total of 24 readings per device. To avoid user bias, the order of these individual readings is randomized. The readings are also performed with a military lensatic compass as an additional baseline. The results are averaged for each device and reported in the form of mean error value and standard deviation of the error values. Sample mean error values are shown in Figure 4 showing a range of 3.4° to 6.1° off from the true compass readings.

1.4. Display Usability

The usability of the display is a very important function of the handheld devices in tactical activities. The display is used for everything on the device when the screen is on, and for the TransApps devices, often used in outdoor lighting conditions ranging from nighttime use where light must be kept low to daytime use where the screens can be unusable due to not enough light being output from the screen. The metric for this test

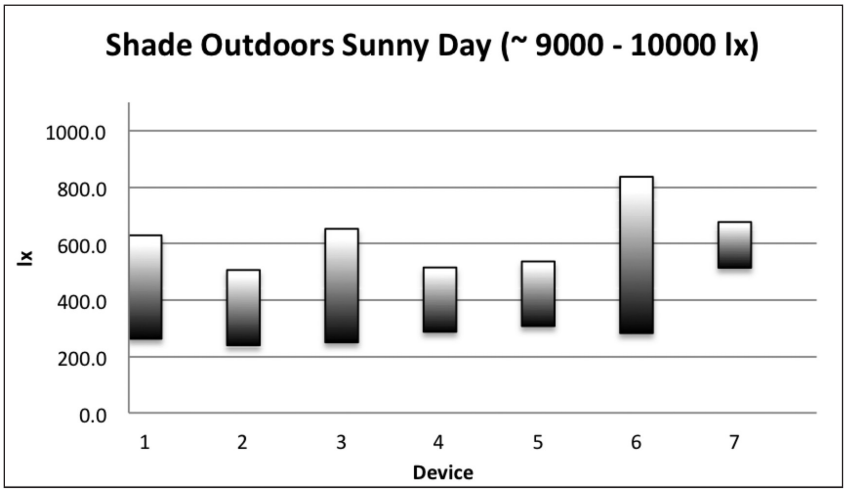

Figure 5: Shade Outdoors Sunny Day (~9000 - 10000 lx)

method is a combination of brightness and contrast produced by the screens in various ambient light conditions. A light meter is used to measure the lux level of the white and black coming off of the screen at maximum brightness in ambient light levels ranging from a dark room to full sunlight. At least six readings are taken for each color in each light level and averaged together to account for anomalous outliers. The difference between the average black and white levels provides a usable contrast estimate where a larger difference indicated greater ease in being able to distinguish between black and white on the screen. The results of this testing (black level, white level & usable contrast) is provided in the form of stacked bar charts to make it easier to visualize and compare the devices at each ambient light level. An ideal device would have a black level of as close to 0 lx as possible and as large a usable contrast value as possible. A sample stacked bar chart is shown in Figure 5 showing a range of usable contrast of 162 to 614.5 lx and black levels ranging from ~240 lx (device 2) on the better end up to ~ 515 lx (device 7) on the worse end.

5. Conclusions And Future Work

The NIST testing team has made (and continues to make) a significant impact on the TransApps program in their extensive and detailed testing of the applications. Since the NIST team began assessing applications for this program in 2011, NIST test feedback and reports have offered program leadership greater insight into the capabilities and limitations of the TransApps technologies. This has led to more informed, quicker fielding of the technology, and allows the program personnel overseas to be more knowledgeable of the latest technology iterations before updating their devices.

NIST testers have developed test methods for handheld device components that will allow the program sponsors to compare devices across a number of key areas. The Camera Usability results provide data on what the camera hardware is capable of with regards to speed. The GPS Accuracy and Time Metrics provide data about the accuracy of the GPS location provided by the device as well as the speed of obtaining sufficiently accurate GPS locations. The Compass Accuracy results provide data about accuracy of the compass data for applications that use it on the TransApps devices. The Display Usability results provide an indication of how easy it is to use a device's screen in various ambient light conditions. The sponsors can then apply the program's priorities to the comparisons and determine the best device for the program's needs.

The backbone of the NIST team's effort continues to be its detailed-oriented, 'leave no stone unturned' mentality coupled with independent, third-party objectivity to offer the program sponsor unbiased and thoughtful feedback detailing technological capabilities and limitations. NIST is pleased to contribute to this effort and expects the technology's evolution to continue to enable end-users to work safer and more efficiently in challenging and threatening environments. ❑

ANTHONY DOWNS is a mechanical engineer at the National Institute of Standards and Technology (NIST). He has a Bachelor of Science degree in mechanical engineering from the University of Maryland at College Part. He has been developing test methods for seven years at NIST and has received a United States Department of Commerce Silver Medal Award for Meritorious Federal Service in 2011 and a United States Department of Commerce Bronze Medal

Award for Superior Federal Service. Anthony has been a member of the American Society of Mechanical Engineers since 2002. E-mail: anthony.downs@nist.gov

Endnotes

[1] anthony.downs@nist.gov; phone 1 (301) 975-3436; fax 1 (301) 990-9688; www.nist.gov/el/isd

[2] The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. Approved for Public Release, Distribution Unlimited.

[3] Certain commercial companies, products, and software are identified in this article to explain our research. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the companies, products, and software identified are necessarily the best available for the purpose.

References

1. DARPA. "Transformative Apps Program Description Page." http://www.darpa.mil/Our_Work/I2O/Programs/ Transformative_Apps.aspx (accessed November 30, 2013).

2. Schlenoff, Craig I., Steves, Michelle P., Weiss, Brian A., Shneier, Michael O., and Virts, Ann M., "Applying SCORE to Field-Based Performance Evaluations of Soldier-Worn Sensor Technologies." *Journal of Field Robotics – Special Issue on Quantitative Performance Evaluation of Robotic and Intelligent Systems* 24 (2007): 671-698.

3. Schlenoff, Craig I., Weiss, Brian A., Steves, Michelle P., Virts, Ann M., and Shneier, Michael O. Overview of the First Advanced Technology Evaluations for ASSIST, In Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop, 2006, Gaithersburg, MD.

4. Weiss, Brian A., Schlenoff, Craig I., Shneier, Michael O., and Virts, Ann M. Technology Evaluations and Performance Metrics for Soldier-Worn Sensors for ASSIST, In Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop, 2006, Gaithersburg, MD.

5. Schlenoff, Craig I. ASSIST: Overview of the First Advanced Technology Evaluations, In Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop, 2006, Gaithersburg, MD.

6. Schlenoff, Craig I., Weiss, Brian A., Steves, Michelle P., Sanders, Greg, Proctor, Frederick, and Virts, Ann M., Evaluating Speech Translation Systems: Applying SCORE to TRANSTAC Technologies, In Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop, 2009, Gaithersburg, MD.

7. Weiss, Brian A. and Schlenoff, Craig I., Performance Assessments of Two-way, Free-Form, Speech-to-Speech Translation Systems for Tactical Use, In Proceedings of the 2010 Annual International Test and Evaluation Association (ITEA) Symposium, September 2010, Glendale, AZ.

8. Weiss, Brian A., Schlenoff, Craig I., Sanders, Greg, Steves, Michelle P., Condon, Sheri, Phillips, Jon, and Parvaz, Dan, Performance Evaluation of Speech Translation Systems, In Proceedings of the 6th edition of the Language Resources and Evaluation Conference, May 2008, Marrakech, Morocco.

9. Schlenoff, Craig I. 2010. Applying the Systems, Component and Operationally-Relevant Evaluations (SCORE) Framework to Evaluate Advanced Military Technologies. The ITEA Journal of Test and Evaluation, 31(1): 112-120.

10. Weiss, Brian A. and Schlenoff, Craig I., The Impact of Evaluation Scenario Development on the Quantitative Performance of Speech Translation Systems Prescribed by the SCORE Framework, In Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop, 2009, Gaithersburg, MD.

11. Weiss, Brian A. and Schlenoff, Craig I., Evolution of the SCORE Framework to Enhance Field-Based Performance Evaluations of Emerging Technologies, In Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop, 2008, Gaithersburg, MD.

12. Weiss, Brian A., Fronczek, Lisa J., Morse, Emile L., Kootbally, Zeid, and Schlenoff, Craig I. 2013. Performance Assessments of Android-Powered Military Applications Operating on Tactical Handheld Devices. In Proceedings SPIE 8755, Mobile Multimedia/Image Processing, Security, and Applications 2013 Conference, June 26, Baltimore, MD.

13. Nielsen, Jakob. 1994. Heuristic Evaluation. In Usability Inspection Methods, ed. Jakob Nielsen and Robert L. Mack, New York: John Wiley & Sons.

14. NOAA NGS. 2013. "National Geodetic Survey Data Explorer." http://www.ngs.noaa.gov/CORS-Proxy/NGSDataExplorer/ (accessed July 25, 2013).

15. Google. 2013. "GPS Test – Google Play." https://play.google.com/store/apps/details?id=com.chartcross.gpstest&hl=en (accessed July 25, 2013).

16. Google. 2013. "NMEA Recorder – Google Play" https://play.google.com/store/apps/details?id=com.mephisto.nmearecorder (accessed July 25, 2013).

Acknowledgement

This work was supported by the Defense Advanced Research Projects Agency (DARPA) TransApps program led by the Program Manager, Mr. Doran Michels.