ITEA Journal 2011; 32: 471–478

A Retrospective Analysis of Lessons Learned in Evaluating Advanced Military Technologies

Craig Schlenoff, Brian A. Weiss, and Michelle Steves
National Institute of Standards and Technology, Gaithersburg, Maryland

For the past 6 years, personnel from the National Institute of Standards and Technology (NIST) have served as the independent evaluation team for two major Defense Advanced Research Projects Agency (DARPA) programs. DARPA ASSIST (Advanced Soldier Sensor Information System and Technology) is an advanced technology research and development program whose objective is to exploit soldier-worn sensors to augment a soldier's situational awareness, mission recall, and reporting capability in order to enhance situational knowledge during and following military operations. TRANSTAC (Spoken Language Communication and Translation System for Tactical Use) is another DARPA program, whose goal is to demonstrate capabilities for rapidly developing and fielding free-form, two-way speech-to-speech translation systems that enable English- and foreign-language speakers to communicate with one another in real-world tactical situations where an interpreter is unavailable. Both of these efforts are concluding, so this article focuses on overall lessons learned in evaluating these types of technologies.

Key words: ASSIST; electronic chronicle; end-user needs; evaluation design; evaluation logistic planning; mission reporting; speech; test environment; test personnel; translation system; TRANSTAC.

Over the past 6 years, the National Institute of Standards and Technology (NIST) has served as the independent evaluation team for two Defense Advanced Research Projects Agency (DARPA) efforts. The first effort, called ASSIST (Advanced Soldier Sensor Information System and Technology), has the objective of exploiting soldier-worn sensors to augment a soldier's situational awareness, mission recall, and reporting capability in order to enhance situational knowledge during and following military operations. The second program, called TRANSTAC (Spoken Language Communication and Translation System for Tactical Use), has the objective of rapidly developing and fielding free-form, two-way speech-to-speech translation systems that enable English- and foreign-language speakers to communicate with one another in real-world tactical situations where an interpreter is unavailable. Between these two efforts, NIST has orchestrated 13 live evaluations involving over 100 military personnel and foreign-language speakers at locations varying from military operations in urban terrain sites to hotel conference rooms.

In this article, we will give a brief description of each of these two DARPA efforts and describe some of the overall lessons learned from our experiences.

DARPA ASSIST and TRANSTAC efforts

This section gives a brief overview of the DARPA ASSIST and TRANSTAC efforts.

ASSIST

Soldiers are often asked to perform missions that can take many hours. Examples of missions include presence patrols (where soldiers are tasked to make their presence known in an environment for a variety of reasons), search and reconnaissance missions, and apprehension of suspected insurgents. After a mission is complete, the soldiers are typically asked to provide a report to their commanding officer describing the most important things that happened during the mission. This report is used to gather intelligence about the environment to allow for more informed planning for future missions. Soldiers usually provide this report based solely on their memory, still pictures, handwritten notes, or grid coordinates that were collected during the mission, provided these tools are available.

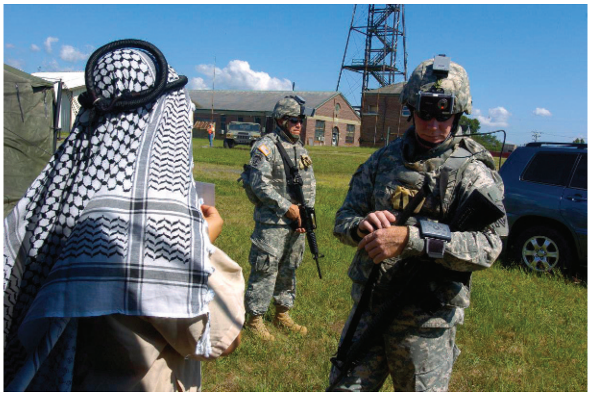

Figure 1. Soldiers using the ASSIST technology.

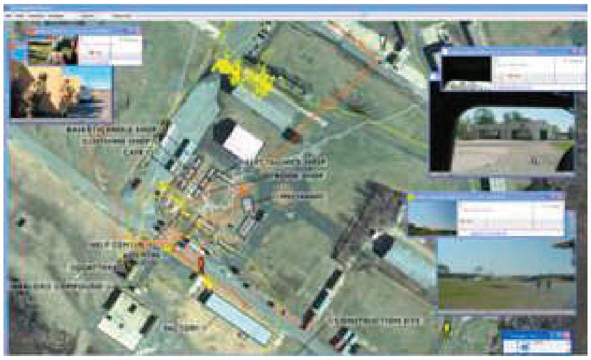

Figure 2. User interface for the ASSIST system.

These missions are often very stressful for the soldiers, and thus there are undoubtedly many instances in which important information is not made available in the report and thus not available for the planning of future missions.

The ASSIST program (Schlenoff 2006) addressed this challenge by instrumenting soldiers with sensors that they can wear directly on their uniform (as shown in *Figure 1*). These sensors include still cameras, video cameras, global positioning systems, inertial navigation systems, microphones, and accelerometers. They continuously record what is going on around the soldiers while on a mission. When soldiers return from their mission, the sensor data are run through a series of software systems that index the data and create an electronic chronicle of the events that happened throughout the time that the ASSIST system was recording (as shown in *Figure 2*). The electronic chronicle includes times that certain sounds or key words were heard, times when certain types of objects were seen, and times that the soldiers were in a specific location or performing certain actions.

With this information, soldiers can give reports without relying solely on their memory. The electronic chronicle will help jog the soldiers' memory on activities that happened that they did not recall during the reporting period, or possibly even make the soldiers aware of important activities that they did not notice when out on the mission. On top of this, the multimedia information that is available in the electronic chronicle is available to the soldiers to include in their reports, which will provide substantially more information to the recipient of the report than the text alone.

Specific technologies being developed include:

- object detection/image classification—the ability to recognize and identify objects in the environment;
- Arabic text translation—the ability to detect, recognize, and translate written Arabic text;
- sound recognition/speech recognition—the ability to identify sound events (e.g., explosions, gunshots, or vehicles) and recognize speech;
- shooter localization/shooter classification—the ability to identify gunshots in the environment; and
- 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.

TRANSTAC

The goal of the TRANSTAC program (Schlenoff et al. 2009) is to demonstrate capabilities for rapidly developing and fielding free-form, two-way translation systems that enable speakers of different languages to communicate with one another in real-world tactical situations without an interpreter.

Several prototype systems have been developed under this program, for numerous military applications, including force protection and medical screening. The technology has been demonstrated on smartphone (shown in *Figure 3*) and laptop platforms. NIST was asked to assess the usability of the overall translation system and to individually assess each component of the system (the speech recognition, the machine translation, and the text-to-speech).

All of the TRANSTAC systems work fundamentally the same. Either English speech or an audio file is fed into the system. Automatic speech recognition processes the speech to recognize what was said and generates a text file of the speech. That text file is then translated to another language using machine translation technology. The resulting text file is then spoken to the foreign-language speaker using text-to-speech technology. This same process then happens in reverse

Figure 3. TRANSTAC systems on a smartphone platform.

when the foreign-language speaker speaks. This is shown in *Figure 4*.

Lessons learned

The rest of this article focuses on some of the overall lessons learned while implementing the evaluations of the technologies described previously. Listed are nine lessons, each with brief explanatory text.

Keep your eye on the ball (the ultimate objective of the evaluation) and make sure your decisions along the way reflect that goal

As evaluation planning proceeds and new approaches and constraints are uncovered, it is often easy to get caught up in the minutiae and lose sight of the big picture. Decisions are often made that solve an immediate challenge but take you further away from the goals that are to be accomplished.

As an example in the TRANSTAC effort, one of the metrics that was used to measure the performance of the systems was a high-level concept-transfer metric that gauged how many concepts could be exchanged in a 10-minute period between the speakers using the system. Once the development teams understood this metric, they started making their systems faster at the expense of accuracy. The English- and the foreign-language speakers sometimes spoke over one another, which would have been highly impractical in a fielded environment but helped them to get though more concepts quicker. They determined that they could maximize their score using this approach even though it is not how they envisioned their fielded systems operating.

The evaluation team identified this issue and is now reconsidering using that metric at all. The test subjects in previous evaluations have consistently stated that they would happily sacrifice some translation time for greater accuracy. If this metric were continued, the

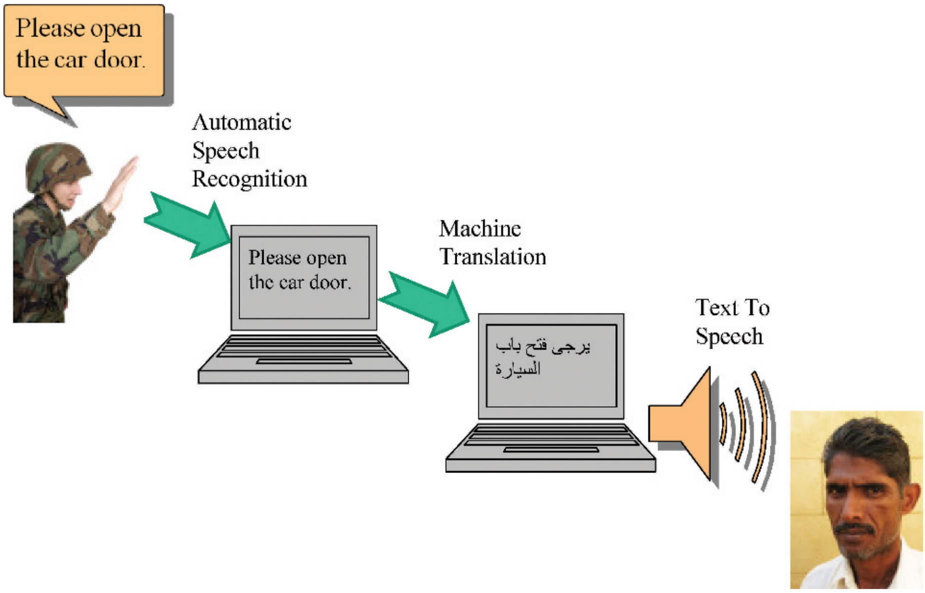
Please open the car door.
Automatic Speech Recognition
Please open the car door.
Machine Translation
يرجى فتح باب السيارة
Text To Speech

Figure 4. How speech translation works.

TRANSTAC systems would progress in a way not aligned with the goals of the program as a whole.

Deeply understand the needs and wants of the technology end users

It is usually a straightforward process to understand the exact needs and wants of technology end users in the case of testing systems that already have been fielded, where end users can categorically state what they like, what they don't like, and what they would improve. Extracting end-user needs and wants is nontrivial when it comes to testing emerging technologies with end-user groups that have yet to be specifically determined, exact use cases that have yet to be finalized, and precise usage procedures that are unclear. During the evaluation design process, it is critical for evaluation team members to speak with representatives of the intended end-user population to thoroughly understand the related challenges they face without the technology and the constraints they are bound by when presented with a new piece of equipment to carry into the field.

NIST TRANSTAC evaluation team members met with soldiers and marines on many occasions to deeply understand the challenges they faced when communicating with foreign-language-speaking personnel without a machine-translation technology. One of the most significant communication challenges currently faced is unreliable interpreters, including those that don't show up for work on time, are limited in their translation skills, or have ulterior motives when facilitating dialogue between U.S. and foreign forces. Other significant challenges include the general unavailability of interpreters. This leads to soldiers and marines attempting to have conversations with foreign-language speakers using extremely limited vocabularies. All of these challenges can lead to misunderstandings, damaged relationships, and in some instances, injuries or loss of life.

Knowledge of this challenge was also complemented by clear statements from soldiers and marines that they wanted a communication tool that was easy to use, fast and accurate with translations, small, lightweight, and durable enough to stand up to frequent use in harsh environments. This insight provided the evaluation team with a clear idea of the soldiers' and marines' needs and wants.

Realize that utility and technical performance assessments are both very important perspectives

Technology evaluations can take many forms, yielding varying types and amounts of data. Data output can yield two unique types of information: quantitative technical performance and qualitative utility assessments. Each piece of data offers unique insight into a technology's overall behavior, individual functionality, and benefit to the end user. Quantitative evaluations can offer detailed information about a system's overall functionality along with specific performance metrics related to inherent components and capabilities. Determining a technology's means of failure at the system level is an important process. Overall failures can lead to individual component or capabilities testing to identify the point of failure and determine which variables or parameters are responsible for the failure. Quantitative metrics also provide a basis of comparison among multiple evaluations and technologies. Likewise, qualitative metrics enable the evaluation team to assess the perceived worth and value the technology has to the test subjects representative of the target user population. This type of insight complements the quantitative data. For example, a technology could be 100% accurate in its function, yet if it is too heavy to carry, users will seldom use it and will therefore place a low value on it. Individually, both of these data types paint very contrasting pictures. It is important that the data be viewed together to get a complete understanding.

NIST's evaluations of advanced technologies have demonstrated a need to collect both types of data. In both the ASSIST and TRANSTAC programs, evaluations were conducted of technologies that had yet to be finalized and deployed to actual end users. This means that the evaluation team's analysis of the collected quantitative and qualitative data was crucial to informing the technology developers and program sponsors on the current state of the systems, including specific successes and areas for improvement. Across both programs, quantitative data were captured that assessed individual technology components, capabilities, and systems. For example, component-level evaluations of the TRANSTAC systems' automatic speech recognition, machine translation, and text-to-speech demonstrated specifically which of these components produced errors ultimately leading to system errors. Also, both programs captured qualitative data at the capability and system levels. For example, capability-level evaluations of the ASSIST technologies enabled the evaluation team to capture specific feedback from soldiers about which technology capabilities (e.g., real-time data sharing or image annotation) were of the most value, easiest to use, etc. Likewise, this specific information, coupled with the other collected data, enabled the evaluation team to paint a clear picture of the technologies' current state.

The NIST evaluation teams have employed an evaluation approach that captures a range of quantitative and qualitative data. This allows the creation of a definitive picture of the technologies' current successes, shortcomings, and areas that must be improved.

Understand that there are often multiple approaches to evaluating a technology, so it is crucial to identify those that will achieve the overall evaluation goals, given the test constraint

There are many approaches for evaluating systems. For any particular evaluation effort there are also various constraints that much be considered, e.g., logistical, budgetary, and programmatic concerns. Method selection must consider these concerns; otherwise, the assessment effort and results may be compromised in undesirable ways. NIST's evaluation framework advocates identifying evaluation goals and user requirements, and then identifying evaluation methodologies that support those test parameters. Once the set of evaluation methodologies that can support the evaluation have been identified, then method selection can be further refined by other logistical parameters, such as availability of qualified personnel to design and conduct the assessment, type of testing environment needed to execute the test, mechanisms needed to collect the data, and data-analysis considerations, e.g., whether time and resources exist to code many hours of video data. Approaches that do not have contingency avenues for high-risk elements should be avoided if possible. For example, if an approach calls for a specific test environment, e.g., military operations in urban terrain, but there is a high probability that the test will be bumped from the site, a feasible fallback location is needed. If no reasonable fallback location is available, alternate approaches should be considered or a determination should be made that test delays are acceptable.

Understand the interactions of the technology with the test environment and the test personnel to be mindful of the technology's ideal operating conditions and its boundaries

The performance of the system under test is greatly and directly related to the environment in which it is being tested and the personnel that are using the system. Slight changes to either one of these factors can often have a significant effect on how well the system performs. For example, the competency of the end user in operating systems similar to the ones being tested can be the difference between success and failure. In addition, the end user's experience in scenarios where the technology would be useful and understanding of how the technology can be best applied is also a critical factor.

Apart from the individual user, many other variables can play a significant role in how well a system performs. In the case of the TRANSTAC systems, these variables may include background noise, distance between the microphone and the speaker, glare issues, dustiness of the environment, wind conditions, dialects of the speakers, etc. Almost none of these variables are true-false; there are various levels that must be understood.

No matter how familiar one gets with a type of technology, nobody knows a specific system better than its developer. However, the developer also has a vested interest in ensuring that the system works as well as possible. For both the DARPA TRANSTAC and ASSIST efforts, regular interaction occurred between the evaluation team and the developers of the technologies. In every case, the developers provided suggestions for the best ways to test the systems and the most appropriate variables to vary. In parallel with this, the evaluation team always spoke with the end users of the technologies (primarily military personnel) to better understand the environments in which the technologies were expected to be used, including variables such as background noise, temperature, weather conditions, etc. Understanding that the technologies were still under development and not yet ready to be fielded, the evaluation team took both sides into consideration and tried to find the proper balance between realism and the known shortfalls of the systems.

Realize that the background and experience of the test subjects can greatly affect their impression of the systems under test

Test subjects—those individuals using a technology during an evaluation in which qualitative or quantitative data are collected—greatly affect data quality by their actions during the test. Their actions are dictated by both the technology training they receive prior to the evaluation and their specific backgrounds and experiences. The latter may include experiences with similar technologies or experiences within the operating environments within which the technologies under test are envisioned for use.

NIST's involvement in six TRANSTAC technology evaluations from 2007 to 2010 has highlighted the fact that the impressions of the soldiers and marines selected as test subject are greatly influenced by their specific backgrounds and experience. A specific example of this can be seen in assigning evaluation scenarios to marines and soldiers. The evaluation team goes to great lengths to assign test subjects scenarios with which they have intimate knowledge, based upon their own deployment experiences and interactions with foreign personnel. Since the evaluation scenarios are categorized within six domains, the soldiers and marines are queried to see how their experiences correlate. For example, a civil-affairs marine would reasonably be assigned the civil-affairs scenarios and could also be paired with some of the facilities-inspections scenarios, based upon their experiences. Conversely, an infantry officer would most likely be suited for the vehicle-checkpoint/traffic-control-point, combined-training, and combined-operations scenarios. Allowing test subjects to use the TRANSTAC systems to facilitate dialogues they are intimately familiar with supports the capture of targeted feedback. The test subjects will have high confidence in stating what worked well and what needs to be remedied with the technology in order for the system to be successful in an actual situation. Likewise, if test subjects are paired with scenarios with which they have little familiarity, then their dialogue struggles have great potential to negatively influence their perception of the technology.

Be cognizant that the structure and content of the technology training and the feedback requests of the test subjects greatly influence the test subjects' perceptions

Any training provided to subjects on the technology to be tested will have an impact on their interaction with the system and subsequently on their perceptions of the technology. Decisions regarding the amount and type of training required to achieve the test objectives must be made. Complex systems can present additional challenges in attempting to train participants. Some questions to be addressed are: How much training is needed? How long will it take and what is the schedule impact? Where will training take place? If training is conducted in the test environment, will that impact the test results in undesired ways? What training materials are needed, e.g., scenario content or task content? Are the training materials different from or similar to the test materials, and what is the impact of that? Who can provide appropriate, unbiased training? The developers know their systems the best, but they are not unbiased. Testing personnel may not be qualified to conduct training for complex systems.

Removing interactions between system-developer personnel and test subjects can help with controlling those influences on the test subjects; however, there may be advantages of system-developer involvement that lead the evaluation designers to consider having the developers involved during the evaluation period. For example, it may be beneficial to the sponsoring program to have its developers see and learn firsthand how their systems are received and hear subjects' concerns. Also, as mentioned before, the systems may be sufficiently complex that only the system developers can provide adequate training, or be so prototypical in nature that only the developers can set some configuration options (because these controls may not yet have been exposed at the user interface). For off-the-desktop systems, various physical configurations may need to be fitted to each test subject each time the system is deployed. In any of these cases, a simple inquiry of "So, how was it?" and the resulting discussion can have an impact on what the subject ultimately reports in their official assessment feedback. When system developers have access to the test subjects during the testing period, appropriate ground rules need to specified and enforced to control the effect of these influences.

Note that there are often multiple options available to assess specific metrics, so it is critical to identify those options that are optimal for producing the desired assessments

There are typically quite a few measures that can be collected for use in assessing any particular metric. Which measures or assessors are selected may have an impact on what is collected and reported; therefore, careful attention should be paid to these choices. Additionally, some measures are more or less difficult to collect, some are more costly to collect than others in terms of resources needed, some are logistically more difficult to put in place, and so on. Choices here can impact the cost of the assessments as well as the logistical feasibility of completing the data collection and analysis for assessment, so careful attention to these considerations during the measure-selection process is prudent.

For example, when obtaining feedback from subjects, two examples of assessments could be free-form and Likert-type survey responses. Free-form responses typically consist of open-ended responses that need to be coded or categorized for analysis. Likert-type responses to well-formed queries allow quantitative assessment of the data. Assessments for the latter type of data can often be much faster to perform than analysis of free-form responses, and can give quite different perspectives of the same experience interaction.

A case in point is documented in Steves and Morse (2009). In the early stages of the TRANSTAC evaluations, utility data were collected solely via survey instruments. Although a combination of Likert-type response questions and free-form inquiries was used, the free-form responses became repetitive and sparse over the course of the evaluation period. Adding semistructured interviews and the resulting gathered data provided very rich insights into the survey-based data and the user experience overall. However, the cost to collect and analyze the additional data was definitely greater.

Be mindful that your metrics and evaluation approach may need to evolve over time

It is typical for evaluation requirements and concerns to evolve over time, especially if the time span in which the assessments are performed is long or if there are a large number of unknowns at the beginning of the design phase. As more is learned about the system and user requirements, initially envisioned approaches may need to be modified to provide useful assessment of the system. For example, in testing a prototype system, the initial assessment goals may include user testing, but as more is learned, it may be determined that the user interface is not sufficiently developed for users. In this case, another approach could be used, such as expert review, to provide some formative feedback for developers regarding how to move forward to support their eventual users effectively. Understanding of the system, its requirements, its state of development, and user requirements may impact the initial assessment vision, as that vision may not have had the benefit of the understanding gained during the initial design phase.

For example, in both projects the systems were evolving over time. Improvements to existing capabilities were made and new features added between evaluations. This required that changes in what was assessed—and, at times, in how it was assessed—be made. In particular, an early TRANSTAC platform was a laptop; in the field, it was a laptop in a backpack, where the screen could not be viewed and the systems would overheat easily. In the last evaluations, the platform was a smartphone. This meant that field evaluations could be more realistically situated in later evaluations.

Keep the high-level objective of the evaluation in mind and be flexible as modifications need to be made.

Discussion

In this article, we describe the evaluation approach that has been applied to two DARPA-funded efforts over the past 6 years and focus on nine lessons that have been learned during that time. This is not meant to be a comprehensive list of all the factors that should be considered when evaluating these types of systems, but instead represents some of the most critical ones as determined by the authors.

The main lesson described in this article is that additional effort put into the design and logistics planning of the evaluation up front can pay off quite a bit as the evaluation progresses. The design stage of the evaluation is critical, and decisions made during that time have a huge effect on how successful the evaluation will be. Bad decisions in the design can be very difficult to fix later on. This can be compared to the cycle of manufacturing product development: Problems that are identified and resolved in the design stage of a product can cost orders of magnitude less to fix than those same problems if they are not identified until the manufacturing or distribution phases. ❑

Craig Schlenoff is the acting group leader of the Knowledge Systems Groups in the Intelligent Systems Division at the National Institute of Standards and Technology (NIST). His research interests include performance-evaluation techniques applied to autonomous systems and manufacturing. He has served as the program manager for the Process Engineering Program at NIST as well as the director of ontologies at VerticalNet, Inc. He leads numerous million-dollar projects dealing with performance evaluation of intelligent systems. He received his bachelor's degree in mechanical engineering from the University of Maryland, College Park, and his master's degree from Rensselaer Polytechnic Institute. E-mail: craig.schlenoff@nist.gov

Brian A. Weiss is a mechanical engineer at NIST. His focus is the development and implementation of performance metrics to quantify technical performance and assess end-user utility of intelligent systems. His schooling has been through the University of Maryland, including a bachelor of science in mechanical engineering and a professional master of engineering; he is working towards his doctor of philosophy degree in mechanical engineering. His research is targeted at developing the Multi-Relationship Evaluation Design (MRED) framework as a tool that will handle uncertainty in its automatic design of evaluation test plans. E-mail: brian.weiss@nist.gov

Michelle Steves is an information-systems analyst at NIST. Her research interests include utility and usability assessments for desktop and off-the-desktop systems. Recent efforts have included assessments of soldier-worn sensor systems, spoken-language translation systems, a latent-print examination system, and biometrics data-gathering systems. She has a bachelor degree in mathematics and computer science from Western Maryland College, now known as McDaniel College. E-mail: michelle.steves@nist.gov

References

Schlenoff, C. 2006. ASSIST: Overview of the first advanced technology evaluations. In *Proceedings of the 2006 Performance Metrics for Intelligent Systems (Per-*

MIS) Conference, August 21–23, 2010, Gaithersburg, MD, ed. Madhavan, R, and E. Messina, Gaithersburg, MD: National Institute of Standards and Technology.

Schlenoff, C., B. Weiss, M. Steves, G. Sanders, F. Procter, and A. Virts. 2009. Evaluating speech translation systems: Applying SCORE to TRANSTAC technologies. In *Proceedings of the 2009 Performance Metrics for Intelligent Systems (PerMIS) Conference,* September 21–23, 2009, Gaithersburg, Md, ed. Madhavan, R. and E. Messina, Gaithersburg, MD: National Institute of Standards and Technology.

Steves, M., and E. Morse. 2009. Utility assessment in TRANSTAC: Using a set of complementary methods. In *Proceedings of the 2009 Performance Metrics for Intelligent Systems (PerMIS) Conference,* September 21–23, 2009, Gaithersburg, Md, ed. Madhavan, R., and E. Messina, Gaithersburg, MD: National Institute of Standards and Technology.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency ASSIST and TRANSTAC programs (program manager: Mari Maeda).

DARPA disclaimer

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.