Lessons Learned in Evaluating DARPA Advanced Military Technologies

Craig Schlenoff
NIST
100 Bureau Drive, Stop 8230
Gaithersburg, MD 20899
301-975-3456
craig.schlenoff@nist.gov

Brian Weiss
NIST
100 Bureau Drive, Stop 8230
Gaithersburg, MD 20899
301-975-4373
brian.weiss@nist.gov

Michelle Potts Steves
NIST
100 Bureau Drive, Stop 8940
Gaithersburg, MD 20899
301-975-3537
michelle.steves@nist.gov

ABSTRACT

For the past six years, personnel from the National Institute of Standards and Technology (NIST) have served as the Independent Evaluation Team (IET) for two major 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 to enhance situational knowledge during and following military operations in urban terrain (MOUT) environments. This program stresses passive collection and automated activity/object recognition that output algorithms, software, and tools that will undergo system integration in future efforts. TRANSTAC (Spoken Language Communication and Translation System for Tactical Use) is another DARPA advanced technology and research program whose goal is to demonstrate capabilities to rapidly develop and field free-form, two-way speech-to-speech translation systems enabling English and foreign language speakers to communicate with one another in real-world tactical situations where an interpreter is unavailable. Several prototype systems have been developed under this program for numerous military applications including force protection, medical screening and civil affairs. Both of these efforts are concluding and as such this paper will focus on overall lessons learned in evaluating these types of technologies.

Categories and Subject Descriptors

D.2.8. [**Software**]: Metrics – *Performance Measures*

General Terms

Measurement, Performance, Experimentation, Human Factors, Languages

Performance Metrics for Intelligent Systems (PerMIS)'10, September 28–30, 2010, Baltimore, Maryland, USA.

Keywords

DARPA, ASSIST, TRANSTAC, performance evaluation, lessons learned, advanced military technology, speech translation, soldier-worn sensors

1. INTRODUCTION

Over the past six years, the National Institute of Standards and Technology has served as the Independent Evaluation Team (IET) for two 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 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 enabling 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 thirteen live evaluations involving over 100 military personnel and foreign language speakers at locations varying from Military Operations in Urban Terrains (MOUT) sites to hotel conference rooms.

In this paper, we will give a brief description of each of these two DARPA efforts and describe some of the overall lessons learned from our experiences. Section 2 describes the DARPA ASSIST and TRANSTAC efforts at a high level and the evaluation approach that was developed to assess the performance of the technologies being developed. Section 3 describes 11 lessons that were learned during the evaluations and give brief background about each. Section 4 concludes the paper.

2. DARPA ASSIST AND TRANSTAC EFFORTS

This section gives a brief overview of the DARPA ASSIST and TRANSTAC efforts as well at the SCORE (System, Component, and Operationally Relevant Evaluations) evaluation approach.

2.1 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, apprehending suspected insurgents, etc. 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 and/or grid coordinates that were collected during the mission, provided these tools are available to the Soldier. These missions are often very stressful for the Soldier 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 [1] addressed this challenge by instrumenting soldiers with sensors that they can wear directly on their uniform. These sensors include still cameras, video cameras, Global Positioning Systems (GPS), Inertial Navigation Systems (INS), microphones, and accelerometers. These sensors continuously record what is going on around the Soldier while on a mission. When Soldiers return from their mission, the sensor data is run through a series of software systems which index the data and create an electronic chronicle of the events that happen throughout the time that the ASSIST system was recording (as shown in Figure 1). The electronic chronicle includes times that certain sounds or keywords were heard, the 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 Soldier's memory on activities that happened that he did not recall during the reporting period, or possibly even make him aware of an important activity that he 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 Soldier to include in the report, which will provide substantially more information to the recipient of the report than the text alone.

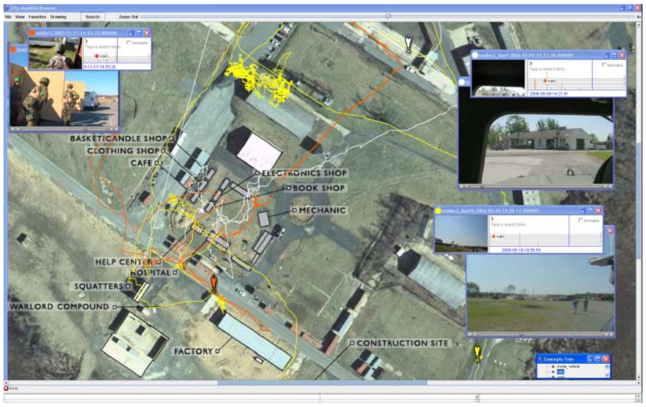

Figure 1: User Interface for ASSIST System

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, vehicles, etc.) and recognize speech
- Shooter Localization / Shooter Classification – the ability to identify gunshots in the environment
- 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

2.2 TRANSTAC

The goal of the TRANSTAC program [2] is to demonstrate capabilities to rapidly develop and field 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 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 (ASR) 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 (MT) technology. The resulting text file is then spoken to the foreign language speaker using Text-To-Speech (TTS) technology. This same process then happens in reverse when the foreign language speaker speaks. This is shown in Figure 2.

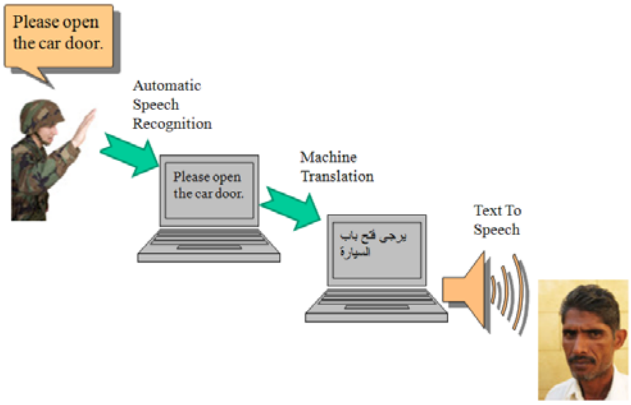
Please open the car door.
Automatic Speech Recognition
Please open the car door.
Machine Translation
يرجى فتح باب السيارة
Text To Speech

Figure 2: How Speech Translation Works

2.3 SCORE

While designing the ASSIST and TRANSTAC evaluations, the IET formulated an evaluation approach to comprehensively assess the performance of the systems. The resulting effort is known as the SCORE (System, Component, and Operationally Relevant Evaluations) [3].

SCORE is a unified set of criteria and software tools for defining a performance evaluation approach for complex intelligent systems. It provides a comprehensive evaluation blueprint 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, in order 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 [2].

The SCORE 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, what type of testing environment is needed to execute the test, what mechanisms are needed to collect the data, data analysis considerations, e.g., if time and resources exist to code many hours of video data.

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, this is performed for each relevant component with the system. At the next level, the overall system is tested in a highly structured environment to understand the performance of individual variables on the system as a whole. Then, 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 unique in that:

- It is applicable to a wide range of technologies, from manufacturing to defense systems
- Elements of SCORE can be decoupled and customized based upon evaluation goals
- It has the ability to evaluate a technology at various stages of development, from conceptual to full maturation
- It combines the results of targeted evaluations to produce an extensive picture of a systems' capabilities and utility

3. LESSONS LEARNED

The rest of this paper will focus on the overall lessons learned while implementing the evaluations on the technologies described above. Listed below are 11 lessons, each with brief explanatory text.

3.1 Designing an effective evaluation can be as much of a research issue as the technology development

To truly design and implement a comprehensive evaluation plan, one must have a deep understanding of the details of the technology under test, including:

- How the technology works
- What the variables are that affect the technologies' performance
- How the technology is expected to be used by the target users including how it will be physically interacted with, in which scenarios it is most appropriate, how it will be carried around, etc.

Understandably, these are the same issues that the developers of the technology are wrestling with. However, in addition to knowing all of these factors, the people that are assessing the capabilities of the technology must also understand:

- How to develop a testing environment that can exercise the full capabilities of the system and understand the shortcomings
- How to ensure that the results obtained are statistically significant and indicative of the performance that will be experienced in the field
- How to identify and train test subjects that are representative of the targeted end users
- How to identify and instrument an environment that is representative of where the technology is expected to be used
- How to determine the metrics and measures that should be used to evaluate the systems and how to properly analyze the results of the evaluations.

These last items are key research challenges that the evaluation team has to face but the development teams rarely have to examine. Any of these factors, if not strongly considered and addressed appropriately, can detrimentally affect the validity of the evaluation results. Many of these factors are described in further detail throughout the remainder of this paper.

For the reasons stated above, it the firm belief of the authors the design of successful evaluation can be as much of a research challenge as is the design of the technology itself. This is primarily due to the number of additional factors and design constraints that must be considered to truly get a comprehensive and accurate assessment of the capabilities of the systems under test.

3.2 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 minutia 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 trying to be accomplished.

As an example, in the DARPA TRANSTAC program, there was much discussion regarding the Soldiers' and Marines' ability to look at the screen of the TRANSTAC system while it was being used. The screens of the TRANSTAC systems contain the textual version of spoken translations. One camp felt that by looking at the screen, the Soldier/Marine was losing situation awareness of what was going on around them which could be dangerous. The other camp felt that the information was available and the Soldier/Marine should be able to look at it if they so desired. After much discussion, it was determined that the Soldiers/Marines would often be protected when using the system so they should be permitted to look at the screen.

Once this was determined, a member of the research team asked, "If the user can see the screen, why do we need to speak out the translations at all. Let's just let them look at the screen." Even though this was a logical next step, it defeated the goal of the program, namely, to create a speech-to-speech translation system. If we started to go down that path, we would be driving the technology in a direction that was contradictory to its intended purpose. Thus, we would have lost sight of the ultimate goal.

As another 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 between the speakers using the system in a ten minute period. Once the development teams understood this metric, they started making their systems faster at the expense of accuracy. The English and the foreign language speech 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 TRANSTAC systems would progress in a way that was not aligned with the goals of the program as a whole.

3.3 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 non-trivial when it comes to testing emerging technologies whose end-user group(s) has yet to be specifically determined, the technologies' exact use-cases have yet to be finalized, and the precise usage procedures 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 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 either don't show up for work on time, are limited in their translation skills or have ulterior motives when facilitating dialogue between US and foreign forces. Other significant challenges include general unavailability of interpreters. This leads to Soldiers and Marines attempting to have conversations with foreign speakers using extremely limited vocabularies. All of these challenges can lead to misunderstandings, damaged relationships, and in some instances, injuries and/or loss of life.

Besides understanding the current challenges faced without machine translation technologies, it was important to understand Soldiers' and Marines' constraints if provided with this new technology. In order to create a relevant and appropriate TRANSTAC evaluation design, it was critical to gather information from Soldiers and Marines to identify numerous elements that would ultimately feed into the evaluation including:

- The relevant dialogues for which a machine translation technology would be viable and/or most useful. Specifically, Soldiers and Marines identified six tactical domains that would lend well to machine translation either because interpreters were scarce for these tasks and/or the conversations took place in relatively secure areas. These domains were (1) Traffic Control Points/Vehicle Checkpoints, (2) Facilities Inspections, (3) Civil Affairs, (4) Medical, (5) Combined Training, and (6) Combined Operations.
- The potential operating environments that would support the use of machine translation technology. These environments aligned themselves with the above six domains. This area also includes their operating constraints or liberties available to them. For example, a Marine may have the ability to sit down with a local police official within a secure base and have a somewhat relaxed conversation about working together. On the other hand, a Soldier may be conducting census operations in a neutral village where he could encounter some unfriendly citizens.
- Criteria for success. It is important for Soldiers and Marines to accomplish their missions in timely and accurate manners without incident. For example, this correlated into the evaluation team's development of high level concept transfer metrics. These metrics included accuracy scores of the technology's ability to translate the English and foreign language concepts and the time it took each speaker to convey an utterance using the technology.

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

3.4 Utility and technical performance assessments are both very important perspectives. Each requires different means to gather and process assessment data and yield different types of analyses.

Technology evaluations can take many forms yielding varying types and amounts of data. Data output can yield two unique types of information which are 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 can be a non-trivial process. Overall failures can lead to individual component and/or capabilities testing to identify the point of failure and determine which variables and/or parameters are responsible for this 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's too heavy for the user to carry, then they will seldom use it and therefore place a low value on it. Individually, both of these data types paint very contrasting pictures. It is important 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 inform 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 was captured that assessed individual technology components, capabilities, and systems. For example, component level evaluations of the TRANSTAC systems' ASR, MT, and TTS 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, image annotation, etc) 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 a definitive picture to be created of the technologies' current successes, shortcomings, and areas that must be improved.

3.5 There are often multiple approaches to evaluating a technology where it's crucial to identify those which 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. The SCORE 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, what type of testing environment is needed to execute the test, what mechanisms are needed to collect the data, data analysis considerations, e.g., if 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., a MOUT site, but there is a high probably 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.

3.6 System training data and/or extensive background scenario information may be needed to perform some assessments. This must be accounted for within the test plan.

A critical element in technology development is the training data provided to the developer that will be used to 'teach' the technology how to act/behave/function appropriately during a given scenario and/or instance. Systems will only perform as effectively as their input models. Similarly, evaluation scenarios that are representative of the training data will yield performance data more indicative of the technology's actual performance as compared to those scenarios not aligned with the training data.

This lesson is visible within the design of the TRANSTAC evaluations. In addition to being responsible for creating and implementing the tests, NIST personnel were also tasked with gathering appropriate and relevant conversational training audio data. This data, composed of English and foreign language speakers conducting tactically-relevant conversations, was based upon real-world military situations and directly motivated the design of the evaluation scenarios.

Creating both the training data scenarios and the evaluation scenarios relied extensively on knowledge collected from Soldiers and Marines who have a detailed understanding of their operating environments and the types of situations the technologies would be viable (refer back to Section 3.3). The evaluation scenarios were significantly modeled after the training data scenarios to not only include realistic elements gathered from Soldiers and Marines, but to also ensure that the evaluation was representative of the data the technologies were trained.

The collection of training data was a non-trivial, multi-month process that impacted the entire evaluation schedule. This process began with IET personnel meeting with Soldiers and Marines so the necessary information could be gathered to create appropriate and current tactical scenarios. Once the information was collected and the scenarios were developed, the IET conducted weekend data collections at a recording studio where English-speaking Marines (or Soldiers) conducted conversations with foreign language speakers (of the upcoming evaluation's target language) through an interpreter based upon these tactical scenarios. Each weekend produced between 30 to 40 hours of audio data which was transcribed and translated by a separate organization. Once the data was ready for distribution, a majority was sent out to the technology developers so it could be used for training data. Only a small amount of data was held back so it could be used for the evaluations. The technology developers required at least several months to work with the data before their technologies would be ready for testing.

Under ideal conditions where only a single weekend of data is collected, this process can occur in as little as four months. Under normal conditions and uncertainties with multiple data collection events, this process takes between seven to eight months. It was crucial that this time be accounted for in the test plan.

3.7 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, their experience being in scenarios where the technology would be useful and understanding how it can be best applied is also a critical factor.

Apart from the user itself, 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, how close the microphone is to the speaker, glare issues, how dusty the environment is, wind conditions, dialect of the speakers, etc. Almost all of these variables are not true or 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 that its developer. The developer is best prepared to have detailed understanding of what is happening underneath the hood and understand how fundamental evaluation design procedures and variables will affect the performance of the system. However, the developer also has a vested interest in ensuring that their system works as well as possible. There is often a balancing act between setting up the evaluation environment in a way that shows the system in the best possible light vs. having an environment that is as realistic as possible to how it is expected to be used.

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 as to the best way to test the systems and what variables would be most appropriate 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 technology was expected to be used, including variables such as background noise, temperature and 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. Often, the final evaluation procedures and environments could not take all concerns into account, but it is important that both sides understood that there were often competing goals and all parties' opinions had to be considered.

3.8 The background and experience of the test subjects can greatly affect their impression of the systems under test.

Test subjects, referring to those individuals using a technology during an evaluation where qualitative and/or quantitative data is collected, greatly impact data quality by their actions during the test. Their actions are dictated both by the technology training they receive prior to the evaluation and their specific backgrounds and experiences. The latter may include experiences with similar technologies and/or experiences within the operating environments the technologies under test are envisioned to be used within.

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 experiences. 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 each test subject scenarios that 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 given their backgrounds. Allowing these test subjects to use the TRANSTAC systems to facilitate dialogues they are intimately familiar 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 that they have little familiarity then their dialogue struggles have great potential to negatively influence their perception of the technology.

Another background influence on the test subjects' perceptions of the technologies is if they've had prior experiences with comparable or similar systems. TRANSTAC is one of the very first programs to employ two-way, free-form, speech-to-speech translation technologies. However, several one-way speech translation systems have been previously deployed receiving mixed reviews from the Soldiers and Marines using them. During the course of the TRANSTAC test events, the evaluation team has encountered several test subjects who have used these similar, yet different technologies. To avoid their perceptions of these earlier systems from bleeding over to their TRANSTAC feedback, the evaluation team has had to make it very clear that these are two entirely separate technologies and they should ‘forget’ everything they know about the earlier systems.

3.9 The structure and content of the technology training and the feedback requests of the test subjects greatly influences 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 need to be determined. Complex systems can present additional challenges when 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 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 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 the 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 first-hand how their systems are received and hear subjects’ concerns. Also, as mentioned above, the systems may be sufficiently complex that only the system developer can provide adequate training or are so prototypical in nature that only the developer 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.

While controlling “unofficial” feedback and subject interactions, official feedback requests need to be carefully tailored to collect data that will inform the metrics selected that meet the test objectives. How and when official subject feedback of system performance is requested will also have an impact on what is reported. For example, if subjects perceive a heavy emphasis regarding their perceptions of speed and accuracy, they will tend to focus their attention on those aspects of system performance and possibly under report other aspects of their experience with the system. Feedback solicitation always needs to be tailored to what the test was intended to collect while leaving opportunities for open-ended responses to garner feedback that might not have been anticipated. Overall, it is the test design team’s responsibility to explore all of these options, consider the impacts, and make choices that meet the test and program objectives.

3.10 There are often multiple options available to assess specific metrics so it’s critical to identify those options which are optimal to produce 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, as mention earlier, 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, therefore 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 be often much faster to perform than analysis of free-form responses, however can give a quite different perspectives of the same experience interaction.

A case in point was documented in [4]. In the early stages of the TRANSTAC evaluations, utility data was collected solely via survey instruments. Although a combination of Likert-like response questions and free-form inquiries were used, the free-form responses became repetitive and sparse over the course of the evaluation period. Adding semi-structured 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.

3.11 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, when 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, state of development, and user requirements may impact the initial assessment vision, as it 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 be made and at times how they were assessed also changed. 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 smart phone. 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.

4. DISCUSSION

In this paper, we describe the evaluation approach that has been applied to two DARPA-funded efforts over the past five years and focus on 11 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 represent some of the most critical ones as determined by the authors.

The main lesson described in this paper is that additional effort put into the design and logistic 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 decision in the design can be very difficult to fix later on. This can be compared to the manufacturing product development cycle. Problems that are identified and resolved in the design stage of a product can cost orders of magnitude less to fix that if those same problems are not identified until the manufacturing or distribution phases.

The SCORE effort described in Section 2 of this paper evolved over the past five years to address many of the lessons described in this document. Almost all of the enhancements focused on the design state of the evaluation, including determining more well-defined approaches to characterizing the all of the evaluation participants, characterizing the variables the affected the system performance of identifying ways to control them, and assuring the metrics that were used to assess the performance of the systems under test truly addressed the overall goal of the programs.

5. ACKNOWLEDGMENTS

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

6. 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.

BIBLIOGRAPHY

[1] C. Schlenoff, "ASSIST: Overview of the First Advanced Technology Evaluations," in *Proceedings of the 2006 Performance Metrics for Intelligent Systems (PerMIS) Conference* Gaithersburg, MD: 2006.

[2] C. Schlenoff, B. Weiss, M. Steves, G. Sanders, F. Proctor, and A. Virts, "Evaluating Speech Translation Systems: Applying SCORE to TRANSTAC Technologies," in *Proceedings of the 2009 Performance Metrics for Intelligent Systems (PerMIS) Conference* 2009.

[3] C. Schlenoff, "Applying the Systems, Component and Operationally-Relevant Evaluations (SCORE) Framework to Evaluate Advanced Military Technologies," *ITEA Journal of Test and Evaluation*, vol. 31, no. 1 Feb. 2010.

[4] M. Steves and E. Morse, "Utility Assessment in TRANSTAC: Using a set of complementary methods," in *Proceedings of the 2009 Performance Metrics for Intelligent Systems Conference* Gaithersburg, MD: 2009.