Performance Assessments of Two-Way, Free-Form, Speech-to-Speech Translation Systems for Tactical Use[1]

Brian A. Weiss
National Institute of Standards and Technology
100 Bureau Drive MS 8230
Gaithersburg, Maryland 20899 USA
brian.weiss@nist.gov
301.975.4373

Craig I. Schlenoff
National Institute of Standards and Technology
100 Bureau Drive MS 8230
Gaithersburg, Maryland 20899 USA
craig.schlenoff@nist.gov
301.975.3456

[1] *Approved for Public Release, Distribution Unlimited*

Performance Assessments of Two-Way, Free-Form, Speech-to-Speech Translation Systems for Tactical Use

Brian A. Weiss
National Institute of Standards and Technology
brian.weiss@nist.gov

Craig I. Schlenoff
National Institute of Standards and Technology
craig.schlenoff@nist.gov

Abstract

A critical challenge for military personnel when operating in foreign countries is effective communication with the local population. To address this issue, the Defense Advanced Research Projects Agency (DARPA) created the Spoken Language Communication and **Trans**lation Systems for **Tac**tical Use (**TRANSTAC**) program. The program's goal is to develop speech-to-speech translation technologies enabling English-speakers to quickly communicate with the local population without an interpreter. DARPA has funded the National Institutes of Standards and Technology to lead the design and implementation of the TRANSTAC performance evaluations. This paper will present these evaluations which enabled the collection of rich quantitative and qualitative data to apply metrics.

1. Overview

The Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program is a Defense Advanced Research Projects Agency (DARPA[2]) advanced technology research and development program aimed at demonstrating 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 (Weiss et al., 2008; Schlenoff et al., 2009). To date, several prototype systems have been developed for traffic control points, facilities inspection, civil affairs, medical screening, combined training, and combined operations domains in Iraqi Arabic, Mandarin, Farsi, Pashto, Dari, and Thai. Systems have been demonstrated on various size platforms ranging from Personal Digital Assistants (PDAs) to laptop-grade platforms. The primary use cases of these technologies involve US military personnel and local foreign language speakers.

Personnel from the National Institute of Standards and Technology (NIST) have served as the Independent Evaluation Team (IET) of the TRANSTAC Program since 2006. As the IET, NIST is responsible for analyzing the performance of the TRANSTAC systems by designing and executing multiple technology evaluations and analyzing the results of these efforts. This paper presents the evaluation methodology that was employed in the April 2010 technology evaluations. This also happens to be the first live evaluation that focused on English to/from Pashto and the first that required the system developers to use smartphones. Detailed results of the evaluations cannot be presented due to restrictions on releasing the data.

[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.

2. System Description

There were a total of four English-to-Pashto/Pashto-to-English translation systems developed by separate teams that were evaluated in April 2010. Each team's system architecture is similar in that they feature three principal components: (1) Automated Speech Recognition (ASR), (2) Machine Translation (MT), and (3) Text-to-Speech (TTS). When a person speaks, the ASR turns the spoken input into source text. Next, the MT translates the source text into the output target language text. The final step is where the TTS produces spoken output of the target language text. The process occurs in reverse allowing the technology to translate in both directions (to and from English) enabling English and Pashto speakers to converse with one another.

Evaluations prior to the April 2010 test event featured the TRANSTAC technologies operating on laptop-based systems and rugged mobile computer platforms. The April 2010 test event marked the first evaluation where the translation software solely operated onboard Nexus One[3] smartphones. Since the translation software was packaged entirely on the phone, these systems functioned without the need for any wireless or cellphone connectivity. Even though the smartphones featured visual interfaces, test subjects interacted in an eyes-free mode where they could operate the technology using buttons that were either built into the device or connected through its external ports. Since the technologies were tested in both the heavily-controlled lab and more-realistic field-like environments, the teams provided the test subjects with various system configurations that included numerous microphones and headphone options. Each system incorporated the use of the Nexus One's internal microphones to capture speech. In addition, one of the teams featured a configuration with a headset microphone. While some of the teams used the system's built-in speaker to output speech, some of them added on an external speaker for speech output.

3. Evaluation Design

An experimental method was designed to evaluate the TRANSTAC technologies given their expected state of maturity. The IET created an evaluation approach that would scale well with the technologies as they evolved, which allows for valid assessments of system performance improvements over time. The evaluation design is highlighted by the development of the scalable testing approach, devising the scenarios for training and evaluation, and identifying subjects for the evaluation.

3.1 Developing a Scalable Testing Approach

The April 2010 Pashto evaluation incorporated many elements from previous test events conducted by NIST including the June 2009 Dari evaluation and the November 2008 Iraqi Arabic evaluation. It also featured some new procedures and evaluation scenarios that were not previously used. Each testing approach was specifically created to scale alongside the technologies' maturing capabilities.

Per the Broad Agency Announcement, the following two metrics were the focus for the TRANSTAC evaluation: (1) System usability testing - providing overall scores and assessments

[3] Certain commercial companies, products and software are identified in this paper in order 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.

to the capabilities of the whole system and (2) Software component testing – evaluating individual components of the system to see how well they performed in isolation. The IET employed the System, Component, and Operationally-Relevant Evaluation (SCORE) framework to attain the two TRANSTAC evaluation goals (Schlenoff, 2010; Weiss and Schlenoff, 2009). The SCORE framework states that in order to get a comprehensive picture of how a technology is expected to perform in its intended operating environments, it must be evaluated at the component level, capability level, system level, and in operationally-relevant environments. Each of the above evaluation types yield insight into the various areas of the performance of the technologies being tested. Examining the full results of all of these evaluations provides a multi-faceted perspective of how the technology will perform within its intended environment. Although there is no substitute to testing the system in its actual operating environment, it's often more informative to evaluate the technologies in controlled venues until they are mature enough to move into more challenging environments.

The IET utilized the SCORE framework to evaluate the TRANSTAC technologies by designing system level evaluations with live, operationally-relevant dialogues where both quantitative technical performance and qualitative utility assessment data were captured. The live evaluations occurred in two venues, the *lab* and the *field*, which will be discussed in later subsections of this paper. Additionally, the individual software components were quantitatively evaluated by using pre-recorded audible utterances and pre-defined textual utterances. These software component tests became known as the *offline* evaluations.

3.2 Evaluation Approaches

For both the live *lab* and *field* evaluations, the IET developed tactically-relevant scenarios to gauge the test subjects' perceived utility of the TRANSTAC system. The first 17 scenarios were performed during the *lab* evaluations within controlled conference room environments across three days of testing. The remaining eight evaluation scenarios were performed during the *field* evaluations outdoors on NIST campus for a day following the *lab*. Both the *lab* and *field* evaluations featured Marines, who played the role of the English-speaking test subjects, and Pashto speakers conducting conversations in their native languages using the TRANSTAC technologies. The goal of each conversation was for the speaker pair to accurately convey as many concepts as possible, relevant to their motivations, to one another in their allotted time. Each conversation was inspired by scenarios which provided each speaker with a relevant motivation within one of six tactical domains (Weiss and Menzel, 2010). At the conclusion of both evaluation types, each speaker filled out questionnaires and participated in interview sessions with evaluation team personnel enabling qualitative assessments of the TRANSTAC systems. Likewise, the quantitative technical performance data was captured by having bilingual human analysts review the detailed conversations between the English and Pashto speakers after the test event. Specifically, the analysts focused on what the human speakers said compared to what the technologies translated. Based upon the bilingual's analysis, the IET calculated numerous quantitative metrics (Schlenoff et al., 2009).

The IET conducted the *offline* evaluations by testing each of the technologies against identical pre-recorded audio and pre-defined textual inputs so comparisons among the systems would be "apples-to-apples." Each system processed identical utterances from audio recordings produced according to the same audio data collection procedures that were used for creating the training data (Weiss and Menzel, 2010). As in prior test events, the *offline* evaluation was conducted during the live evaluations. The systems processed inputs in audio format, logging both the recognition output to test the systems' ASR capabilities and the MT output. Transcriptions of the audio were also processed to test the systems' MT capabilities independent of speech recognition. Using both of these evaluation approaches enables the IET to measure the progressive development of the TRANSTAC technologies and to predict the impact these systems will have on end-user performance within an array of tactical scenarios.

3.2.1 *Lab Evaluation*

The *Lab* evaluations are created to assess the TRANSTAC systems in a heavily-controlled and ideal environment that features no background noise and stationary participants (both sitting and standing). This type of venue enables the IET and the technology developers to gauge the best the systems can perform at their current state of maturity. *Lab* evaluations have been important to carry on throughout the life of the program because previous *Lab* evaluations provide a means to better understand the technologies' long term progress.

The IET produced 17 spontaneous scenarios for the *Lab* evaluation that the test subjects performed in 15 to 25 minute timeframes. Depending upon the scenario, the speakers were seated across from one another at a table or stood across from one another with the English-speaker holding and controlling the Nexus One. One team presented a configuration that enabled both the English and Pashto speakers to have separate phones that they used to communicate as opposed to a single phone between them. The English-speaking Marines were assigned specific scenarios based upon their deployment experiences. All of the Pashto speakers had experience as interpreters in Afghanistan and/or as role-players in training exercises on US military bases so these personnel were competent in developing their individual dialogues.

Another Marine acted as a scribe during the conversation where they were responsible for noting the information that the speaking Marine received from the Pashto speaker during their conversation with the TRANSTAC technology. The scribe did not interact with the TRANSTAC technology or the Pashto speaker during the evaluation. The use of a scribe, not done in previous evaluations, not only added more realism to the conversation, but also allowed the IET to collect additional qualitative data since the scribe filled out a survey questionnaire at the conclusion of each conversation. It should be noted that the test Marines took turns being the speaker and the scribe during the test week. The same procedure was repeated during the *Field* evaluations.

3.2.2 *Field Evaluation*

The goal of the *Field* evaluations was to assess the TRANSTAC systems in a more realistic environment. Purposely, the *field* evaluations introduced uncontrolled ambient background noise, sunlight, and wind. The Marines carried the TRANSTAC technologies some of which featured external, human-attachable speakers and were allowed to move around within their scenario station. Three unique scenario stations were simulated including a white box truck to support a vehicle checkpoint and forward operating base entry control point scenarios, an area to simulate a local national's home to support census and medical conversations, and another area to simulate a facility inspection and combined operations planning dialogues. Although this environment was not realistic compared to intended operating conditions, it introduced numerous factors that were not present within the *Lab*. For example, the vehicle checkpoint scenario that occurred at the box truck station allowed the speakers the opportunity to move in and around a vehicle including opening doors and other compartments. Figure 1 provides an image of the outdoor *field* setup. Note that the individual scenario stations are on the right of the image while a large tent is shown on the left that supported team setup and staging.

Figure 1 – *Field* Evaluation Outdoor Setup

The Marines and Pashto speakers performed eight spontaneous scenarios in the *field* where a scribe was employed in the same manner as was done in the *Lab* evaluations. Likewise, at the conclusion of each conversation, the English and Pashto speakers completed survey questionnaires and participated in semi-structured interviews at the conclusion of each block of four scenarios. This enabled the IET to capture end-user utility and perceived value of the technology. Quantitative technical performance metrics were not assessed from the *Field* evaluation.

3.2.3 *Live Evaluation Constraints*

The live exercises also allowed the speakers to interact with the TRANSTAC systems in a pseudo hands-free, eyes-free manner. The January 2007 Iraqi Arabic/English evaluation was the first time that this constraint was placed on the technology users. During the testing, neither speaker saw the TRANSTAC screen. The only feedback they were provided from the TRANSTAC system was audio, and their physical interaction was limited to push-to-talk capabilities that were either on the touch-screen of the phone or using a button on the side of the phone. The concept behind this was that the Marines needed to keep their attention on their surroundings, so the TRANSTAC system should minimally disrupt their situational awareness.

Unlike previous evaluations, noise-masking was not used in this evaluation. Noise-masking is a solution that was developed and applied in previous evaluations to selectively mask English utterances so that the foreign language speaker, who is bilingual since they also understand English, cannot hear them. Under the noise-masking solution, bilinguals wore headphones enabling them to hear the translated foreign speech, but when English is spoken, they hear white noise that inhibits their understanding of the English speech.

Noise-masking was not used in this test exercise since the Nexus One hardware did not lend itself well to the noise-masking system that was used previously. Software-based noise-masking was briefly explored, but not implemented due to the lack of time necessary to design it.

The disadvantage of not using noise-masking was that the Pashto speakers could hear the English speech and could be jaded as to how much they understand from the Pashto translation by understanding the Marine's spoken English. The Pashto speakers were instructed to ignore the English speech, though experience has shown from past evaluations that this is very difficult to do.

3.2.4 *Offline Evaluation*

The *Offline* evaluations were setup in a manner where the selected audio and text utterances were input into each team's Nexus One TRANSTAC system where the output text and speech was captured and analyzed by IET members. Specific to this most recent test event, the *Offline* evaluation featured a total of 1245 Pashto and English utterances that were treated as a sequestered data set. The corresponding audio files were input into the TRANSTAC systems which performed ASR, then executed MT to generate text output files. Likewise, the corresponding transcriptions of these same original audio files were fed into the technologies where MT was executed, only, to produce text output files.

Analysis of the offline evaluation focused on component level analysis of the TRANSTAC systems using automated metrics and human judgments. The following metrics were used to analyze the offline data:

- ASR (Automatic Speech Recognition)
    - Word Error Rate (WER)
- ASR and MT (Machine Translation) together
    - BLEU
    - Fine grained concept transfer, performed by bilingual human judges
    - Likert judgment at utterance level, performed by bilingual human judges

These metrics are discussed further in Section 7.0 and can be found in greater detail in (Schlenoff et al., 2009 and Weiss et al., 2008).

3.3 Evaluation Participants

The main participants that interacted with the TRANSTAC systems were the Marines and Pashto speakers. Eight Marines and one Navy surgeon were present at the evaluation, which were identified and provided by the U.S. Marine Corps Forces, Pacific Experimentation Center (MEC). Six Pashto speakers were present for the evaluation, which were identified and provided by a Middle East cultural advisor. Some of the Pashto speakers had served as translators to support the US military in Afghanistan. Detailed demographics about these participants can be found in Section 4.

In addition to the Marines and Pashto speakers, there were members of the IET that served various roles during the evaluation including station coordinators, interviewers/observers, quality assurors, audio/visual experts, and data collection specialists (Schlenoff et al., 2009, Weiss et al., 2008).

4. Demographics

Demographic information was self-reported by each participant via survey instruments. It was collected during the testing period. Participants were asked to provide basic demographic information such as age and gender, some information on their speech and language influences, e.g., languages they speak, places where they have lived, language(s) spoken at home as children, and how often and comfortable they are with using computers. Additionally, the Marines were asked to provide demographic information related to their military experience, such as rank, length of service, Military Occupation Specialties, and Operation Iraqi Freedom and Operation Enduring Freedom deployment duration(s) and locations.

To summarize the Marine demographics, all six English speakers were male and had an average length of military service of 8.67 years ranging from three to 13 years. Their ranks include two Captains, one Gunnery Sergeant, one Staff Sergeant and two Sergeants where three are currently on active duty while the other three are reservists. Six Pashto speakers participated in this evaluation with all being male. All of them had immigrated to the United States where five of them grew up in Afghanistan while one lived in Pakistan. One participant had obtained a bachelors' degree, one reported attending some college, and four reported having a high school degree.

5. Participant Preparation

The English and Pashto speaker training was conducted based upon specific sets of rules provided to each speaker group. These rules were emphasized as IET members explained their roles within the evaluation.

5.1 English Speaker Rules

This training began with each speaker being given specific rules to abide by when they were using the technologies in the evaluation scenarios. The most significant one for the English speakers was:

- Your conversation should stay reasonably within the bounds of the scenario's motivation, but you should not feel confined to the talking points specified and are free to reasonably expand upon the motivation. For example, a vehicle checkpoint scenario could reasonably turn into a medical assessment if the driver claims to need medical attention.

Example dialogues were then discussed highlighting appropriate interactions (a single speaker talking at a time, the English speaker directing the microphone and/or speaker at the Pashto speakers when appropriate, etc.) along with undesirable interactions (both speakers talking at the same time, long-winded Pashto speakers where their natural responses would be minimal, etc).

5.2 Pashto Speaker Rules

The Pashto speaker training was centered on a list of rules that were provided to these speakers at the onset of their training. These included:

- You should provide consistent and relevant answers (example – if you stated you have 2 children and the technology did not like '2', then you should not change your answer to another number. Rather rephrase your answer or move on, as directed by the English speaker.)
- You should pay attention only to the Pashto speech coming out of the technology. Do NOT respond to the English speech from the Marine or from the technology. However, you should expect that you won't receive perfect translations meaning that a system output of "House

mine" reasonably means that this house is mine if the question asked is "Who owns this house?"

The Pashto speakers were presented with examples of both appropriate and inappropriate interactions.

6. Metrics

The IET intends the metrics to reflect the goal of the TRANSTAC program: the deployed use of speech-to-speech MT technology that enables consistently successful communication between U.S. military users and local civilians whom they encounter. The TRANSTAC Community is in agreement that the two aspects that best identify the ability of TRANSTAC systems to meet that goal are (1) the semantic adequacy of the translations, leading to justified user confidence in the system's translations, and (2) the ability of English and Pashto speakers to successfully carry out a task-oriented dialogue in a narrowly focused domain of known operational need under conditions that reasonably simulate use in the field. The latter of those two aspects is presented in the Sections 7.1. The former is elaborated upon in the rest of this section.

6.1 End-user Feedback from Test Participants

After each live *lab* and *field* scenario, the Marines and the Pashto speakers, filled out a detailed survey asking them about their experiences with the TRANSTAC systems. The surveys explored how easy the system was to use, how well they perceived it worked, and errors that the users encountered when interacting with the system. The Marines and Pashto speakers also participated in semi-structured interviews after each morning and afternoon block of live evaluations. These interviews, led by IET members further explored various questions including "What did you like?, What didn't you like?, What would you change?", etc to obtain more candid and pointed feedback on the technologies.

6.2 High Level Concept Transfer for Live Evaluations

Semantic adequacy of the translations was assessed by six bilingual judges telling us whether the meaning of each utterance came across. The high-level concept metric is the number of utterances that are judged to have succeeded. Thus, failed utterances are not directly scored (other than taking up time). The high-level concept metric is an efficiency metric which shows the number of successful utterances per unit of time, as well as accuracy. This metric is roughly quantitative.

6.3 Low Level Concept Transfer for Offline Evaluations

Low level concept transfer is a quantitative measure of the transfer of the low-level elements of meaning in each utterance. In this context, a low-level concept is a specific content word (or words) in an utterance. For example, the phrase "The school past the bazaar before the clinic." is one high-level concept, but is made up of three low-level concepts (school, past the bazaar, before the clinic).

We had an analyst who is a native speaker of each source language identify the low-level elements of meaning (low level concepts) in representative sets of input utterances from the offline datasets and then asked a panel of five bilingual judges to tell us which low-level concepts were successfully transferred into the target-language output (where failures are deletions, substitutions, or insertions of concepts).

Progress from one evaluation to the next may be presented as a comparison of odds ratio. Odds of successful concept transfer is a more quantitative measure of translation adequacy than the Likert-type judgments of semantic adequacy — the Likert-type judgments give the bilingual judges the opportunity to take into account the relative importance of the various concepts while the low-level concept transfer does not (Sanders et al., 2008).

6.4 Likert Scores for Offline Evaluations

The next metric is a judgment of the semantic adequacy of the translations. The standard is to measure this by having a panel of bilingual judges rate the semantic adequacy of the translations, an utterance at a time. We asked our panel of five bilingual judges to assign a Likert-type score to each utterance, choosing from a seven-point scale.

- +3 Completely_adequate
- +2
- +1 Tending_adequate
- 0
- –1 Tending_inadequate
- –2
- –3 Inadequate

6.5 Automated Metrics

Automated metrics are intended to enable the technology developers to better understand what aspects of performance account for the end-to-end success of their systems. It is the intent to identify the automated metrics that can be run quickly and easily yet will correlate strongly with judgments of semantic adequacy provided by bilingual judges. The automated metrics focus on the core technologies. For speech recognition, we calculated Word-Error-Rate (WER) — using SCTK version 2.2.2 and automated procedures for normalizing the hypothesis and reference texts. For machine translation, we calculated BLEU (Papieni et al., 2002). BLEU was calculated with four reference translations, and is the default version using unigrams through 4-grams.

7. Conclusion

The NIST IET learned numerous lessons from the April 2010 test event that will be explored for the August 2010 Dari test exercise. These included (1) shortening the training time for the speakers since the technologies are very straightforward, (2) allowing the English speakers the ability to look at the Nexus One during the interaction to view the output English ASR and Pashto to English MT, (3) enhancing the observation capabilities of the technology developers so they can better view successful and challenging interactions and (4) targeting English speakers from previous evaluations since they require a smaller learning curve to use the translation systems. Many of these lessons are becoming evaluation improvements which are expected to be deployed in August 2010.

Acknowledgements

This work is funded through the DARPA TRANSTAC program and the authors greatly appreciate the support of the TRANSTAC program manager, Dr. Mari Maeda.

References

Kishore Papieni, Salim Roukos, Todd Ward, and Wei-Jing, Zhu, "BLEU: A method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

Gregory Sanders, Sebastien Bronsart, Sherri Condon, and Craig Schlenoff, "Odds of successful transfer of low-level concepts: A key metric for bidirectional speech-to-speech machine translation in DARPA's TRANSTAC program," *Proceedings of the 6th edition of the Language Resources and Evaluation Conference*, Marrakech, Morocco, 2008.

Craig Schlenoff, "Applying the System, Component, and Operationally-Relevant Evaluation (SCORE) Framework to Evaluate Advanced Military Technologies," *International Test and Evaluation Association (ITEA) Journal*, vol. **31**:1, pp. 112 – 120, March 2010.

Craig Schlenoff, Brian A. Weiss, Michelle P. Steves, Gregory Sanders, Fred Proctor, and Ann Virts, "Evaluating Speech Translation Systems: Applying SCORE to TRANSTAC Technologies," in *Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop*, 2009.

Brian A. Weiss and Marnie Menzel, "Development of Domain-Specific Scenarios for Training and Evaluation of Two-Way, Free Form, Spoken Language Translation Devices," *International Test and Evaluation Association (ITEA) Journal*, vol. **31**:1, pp. 39 - 47, March, 2010.

Brian A. Weiss and Craig Schlenoff, "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.

Brian A. Weiss, Craig Schlenoff, Michelle P. Steves, Sherri Condon, Jon Phillips, and Dan Parvaz, "Performance Evaluation of Speech Translation Systems," *Proceedings of the 6th edition of the Language Resources and Evaluation Conference*, Marrakech, Morocco, 2008.

Biographies

Brian A. Weiss has been a mechanical engineer at the National Institutes of Standards and Technology in Maryland since 2002. His focus is the development and implementation of performance metrics to quantify technical performance and assess end-user utility of intelligent systems throughout various stages of development. His current projects include assessments of soldier-worn sensor systems and spoken language translation devices. He has a Bachelors of Science in Mechanical Engineering from the University of Maryland, a Professional Master of

Engineering from the University of Maryland and is working towards his Doctor of Philosophy in Mechanical Engineering with the University of Maryland.

Craig Schlenoff is the Acting Group Leader of the Knowledge Systems Group in the Intelligent Systems Division at the National Institute of Standards and Technology. His research includes performance evaluation techniques applied to autonomous systems and manufacturing as well as research in knowledge representation/ontologies. He previously served as the program manager for the Process Engineering Program at NIST and the Director of Ontologies at VerticalNet. He leads numerous million-dollar projects, dealing with performance evaluation of advanced military technologies. He received his Bachelors degree from the University of Maryland and his Masters degree from Rensselaer Polytechnic Institute, both in mechanical engineering.