2.2 Proposal Road Map

SAIC's approach to Technical Area 3 (Evaluation) for the DARPA Robust Automatic Transcription of Speech (RATS) Program is based on applying and adapting best practices developed over previous IPTO and Intelligence Community (IC) evaluations. This approach combines methodology—a collaborative specification process based on a comprehensive experimental design, targeted refinement of metrics, and a streamlined process for moving to classified environments—with a scenario-driven approach to the selection and partition of data and an open evaluation framework that provides continuous testing to researchers and multi-level reporting of results to DARPA. SAIC's team—including PI Richard La Valley, Dr. William Hardy, Dr. Henry Goldberg and consultant Professor Paul Cohen—brings together experts in metrics and experimental design, speech analysis technology, and statistics with extensive knowledge of the operational missions that RATS technology is intended to address. **Figure 2.2-1** illustrates the SAIC approach and its relationship to performers in Technical Areas 1 and 2 while providing a roadmap of the sections of the proposal that address each area.

Figure 2.2-2 provides a detailed roadmap to all major elements of the SAIC proposal.

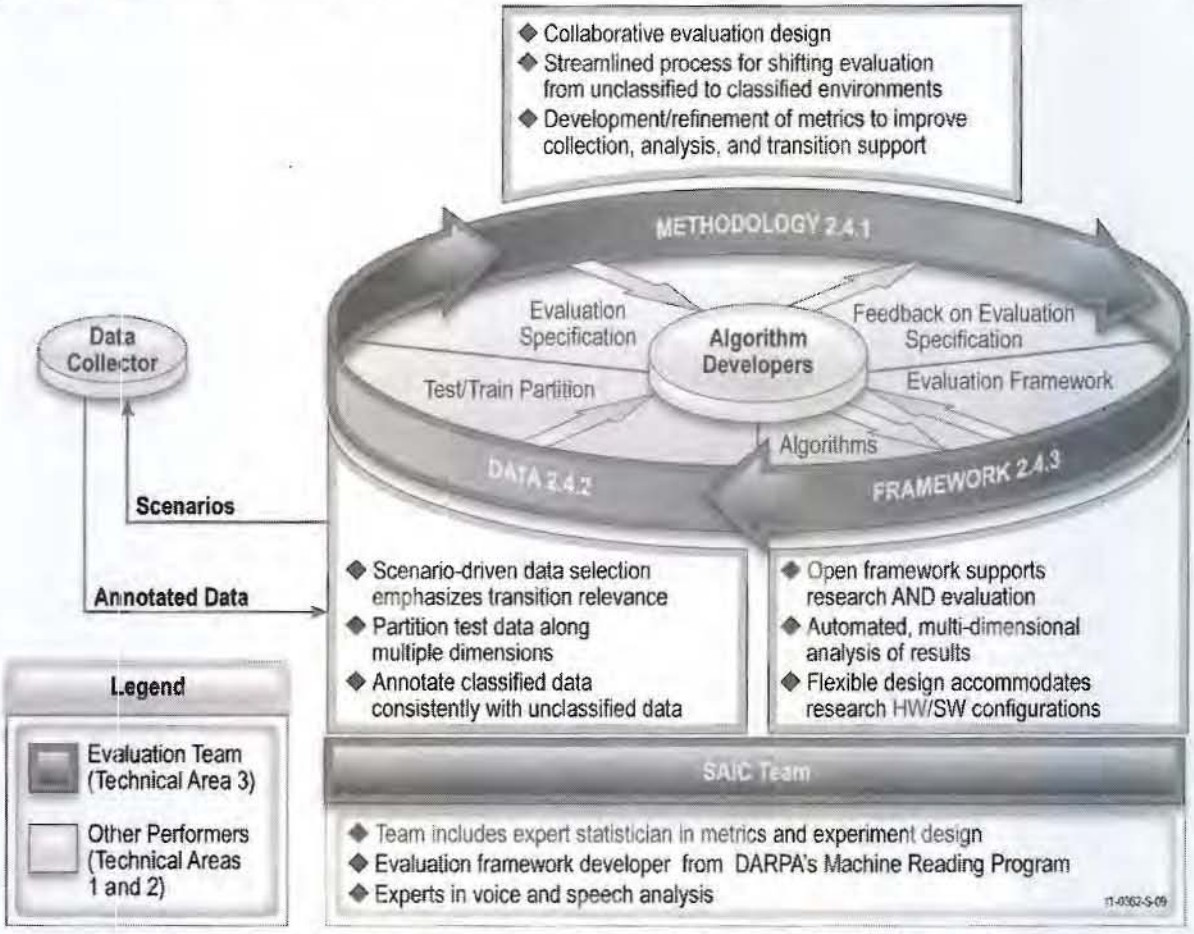

Figure 2.2-1. SAIC's Approach Combines Methodology, Data, and Framework to Provide a Comprehensive Approach to RATS Evaluation. *SAIC's approach includes extensive interaction with both algorithm developers (Technical Area 1) and data collection (Technical Area 2), assuring a collaborative approach that accelerates research progress while effectively measuring that progress. Section references point to sections of the proposal that highlight each area.*

Topic	Synopsis	Page #
Main Goals of Proposed Research (Evaluation)	◆ Create an evaluation specification and methodology that defines a fair and comprehensive approach to measuring and evaluating research progress against RATS goals	2.4-1
	◆ Perform comprehensive evaluation of speech analysis algorithms that shifts smoothly from unclassified environments to classified environments and data in Phases 2 and 3	2.4-1, 2.4-5
	◆ Develop an Evaluation Framework that provides continuous testing during development and layered performance measurement during evaluation	2.4-4
	◆ Partition data sets to support evaluation of performance in realistic operational scenarios and across multiple relevant dimensions (noise level, # speakers, etc.) to emphasize transition relevance of research and support estimation of MOEs from MOPs	2.4-6 2.4-1, 2.4-2, 2.4-4
Tangible Benefits to End Users	◆ Provide clear, comprehensive evaluation results to DARPA that support programmatic decisions	2.4-2, 2.4-4
	◆ Maximize research progress for algorithm developers by tailoring evaluation specifications, offering flexible evaluation schedules to accommodate rapid technology advances, and providing detailed evaluation results	2.4-4
	◆ Direct research toward areas of relevance to potential operational users and transition partners through realistic, scenario-based evaluations	2.4-4
	◆ Provide opportunities for exposure of RATS technologies to potential users and transition partners	2.4-4
Critical Technical Barriers	◆ Difficulty in integrating with research hardware and software platforms rapidly prior to limited end of phase evaluation period	2.4-4, 2.4-7
	◆ Lack of methods to generalize measures of performance (MOPs) into measures of effectiveness (MOEs) that directly impact operational users	2.4-4
	◆ Measurement of performance within required confidence levels to support comparison to phase-specific goals	2.4-7
	◆ Refinement of metrics to address technical limitations in annotator performance	2.4-2
	◆ Challenge in annotating classified speech data consistently with unclassified annotations to ensure utility of resulting classified evaluation results	2.4-5
Main Elements of Technical Approach	◆ Provide experienced team combining experts in metrics and experimental design, speech technology, and statistics	2.4-6, 2.9-1
	◆ Leverage existing evaluation artifacts (specifications and software) from similar programs to maximize effectiveness and minimize risk and cost to the government	2.4-1, 2.4-5
	◆ Collaborate with RATS algorithm developers in design of evaluation specification	2.4-1
	◆ Develop Evaluation Framework that provides robust, automated testing of algorithm performance on all RATS metrics across multiple dimensions	2.4-1
	◆ Incentivize continuous integration by providing Evaluation Framework to algorithm developers, providing continuous testing to facilitate research progress	2.4-4
	◆ Define operational scenarios based on understanding of real IC and DOD missions that emphasize envisioned RATS CONOPS	2.4-6
	◆ Partner with Technical Area 2 team to assure that data collection and resulting partition support planned evaluation protocol and scenarios	2.4-6
Basis of Confidence	◆ Extensive record of successful evaluations for IPTO, the IC, and DOD	2.4-5, 2.4-6, 2.9-1
	◆ Detailed knowledge of state of the art approaches for measuring speech analysis technologies	2.4-6
	◆ Cleared staff and facilities to support evaluations at the TS/SCI level	2.4-5
	◆ Successful previous evaluation of speech transcription technology leading to formal transition to operational environment	2.4-5, 2.4-6, 2.9-1, 2.10-6
Nature and Description of End Results	◆ Thorough evaluation of research progress against defined, phase-specific goals for all RATS metrics	2.4-10
	◆ Multi-level analysis that provides detailed breakdown of algorithm performance across multiple dimensions, identifying strengths, weaknesses, and opportunities for research progress	2.4-1-2.4.4

Figure 2.2-2. Detailed Roadmap to All Major Elements of the SAIC Proposal

Topic	Synopsis	Page #
Nature and Description of End Results (Continued)	◆ Projected MOEs based on MOPs and operational scenarios to support possible transition	2.4-4
	◆ Extensive evaluation framework for speech technology that can be distributed to the wider research community at the discretion of DARPA	2.4-4
Cost and Schedule	◆ Phase 1: 18 months - $717K ◆ Phase 2: 12 months - $487K ◆ Phase 3: 12 months - $392K	2.11-1-2.11-3
Plans and Capability to Accomplish Technology Transition	◆ Develop a framework to support both standard and non-standard interfaces	2.4-7
	◆ Include detailed specifications documentation	2.4-1
	◆ Establish a mirrored testbed for unclassified and classified evaluations to assure that the algorithms can operate effectively in both environments	2.4-5, 2.2-3, 2.10-5
	◆ Develop realistic operational scenarios that can be used to estimate MOEs for potential transition partners	2.4-1,2.4-3, 2.4-4
	◆ Leverage experience from previous successful transitions of speech analysis technology into the IC	2.4-5

Figure 2.2-2. Detailed Roadmap to All Major Elements of the SAIC Proposal (continued)

2.3 Reserved

Note: The outline instructions included in DARPA-BAA-10-34 as Amended on 16 March 2010 did not include Section 2.3. This section is reserved to align with the BAA outline instructions.

2.4 Technical Approach

Summary. SAIC's approach to the evaluation task for RATS focuses on applying and adapting best practices developed in previous NIST, IPTO and IC evaluations of speech analysis and other language technologies. SAIC will leverage our extensive experience implementing a research framework and evaluation process that supports research progress, provides fair and objective evaluation, transitions seamlessly from unclassified to classified environments, and maximizes relevance to possible transition partners in the DOD and IC communities. SAIC will utilize its experience and knowledge of likely operational scenarios motivating the RATS performance tasks, as well as its extensive experience transitioning tools into DOD and the Intelligence Community (IC), to focus evaluation on problems of likely interest and to support DARPA in transitioning RATS technologies to operational users.

2.4.1 Methodology

2.4.1.1 Evaluation Specification

A key element to successful evaluation of research programs is clear communication of evaluation methodologies and protocols as early as possible, allowing researchers to focus their efforts and providing the opportunity to recommend improvements that minimize impacts on research and improve the utility and accuracy of performance data collection.

The SAIC Team will develop and publish a draft of the Evaluation Specification Document describing data, tasks, test protocols, and metrics during the first six months of Phase 1. Research teams will be able to provide feedback on a variety of issues that will be incorporated (as appropriate and approved by the government) into a series of releases that are consistent with each phase. This collaborative approach will lead to a comprehensive and fair evaluation specification that will serve as the basis for successful evaluations of RATS.

2.4.1.2 Experimental Design

The method proposed to evaluate RATS technologies at the end of each phase has two distinct approaches. The first approach will be the calculation of the measures of performance as outlined over a set of audio files as outlined by the BAA as well as the calculation over each audio instance within the audio file. The second method proposed uses analysis of variance tests to investigate the performance sensitivity over different dimensions of interest to DARPA's customers. Note that both analyses will be generated from the same testing process, allowing more in-depth analysis with minimal additional effort or cost.

Technical Area 1 systems will be presented with a series of trials which will be geared to evaluate the performance of a selected technology (SAD, LID, SID or KWS). Each trial will consist of a series of audio files that will be scored on whether the condition of interest (Speech, Language, Speaker, Key Word) was present (yes or no) along with certainty of that call.

The measures of performance (MOPs) outlined in the RATS BAA will be calculated for each event within the audio file, as well as cumulatively over the events within the file and overall for the audio files.

Based on our understanding of the overall population of possible events and making the assumption that voice activity will comprise of at least 75% of the activity in the audio provided by the Task 2 team, we have computed the total sample size needed for 95% confidence estimates to be at least 800 events. The training and evaluation data split provided in the BAA provides more than enough of a population of 3 second events necessary for each evaluation phase and will be sufficient to provide estimates of the true performance of the particular algorithm with 95% level of confidence.

SAIC proposes an experimental design approach to evaluate each technology and investigate the performance over multiple dimensions as in **figure 2.4-1**. Experiments will be run on various combinations of these dimensions and the results will be analyzed and examined for statistical differences (**figure 2.4-2**).

In order to obtain a sufficiently large number of samples for variance analysis to achieve a 95% level of confidence, each Technical Area 1 system will be evaluated on a set of audio files which fit the characteristics of the experimental

Technology	Dimension of Interest
SAD	Level of SNR, Noise with and without Music, Gender
LID	# of Languages, Open and Closed Sets of Languages, Gender, Type of Speaker (Native, Non-Native)
SID	# of Speakers, Level of SNR, Open and Closed Sets of Speakers, Gender, Type of Speaker (Native, Non-Native)
KWS	# of Words, Level of SNR, Open and Closed Sets of Words

Figure 2.4-1. Possible Experimental Design Dimensions

design as outlined; the measures of performance (MOPs) will be calculated, and analyzed and reported. Since each of the set of audio files within one of the cells will have been collected by the same Technical Area 2 performer under similar conditions, it is anticipated that the results should provide a good basis for decision making by DARPA at the end of each phase.

2.4.1.2.1 Annotation, Trials, and Analysis. In order to be able to compare RATS to previous testing such as NIST's Rich Transcription (RT), Language Recognition Evaluation (LRE), and Spoken Term Detection (STD) programs, SAIC proposes a similar approach for the RATS evaluation and the handling of the data.

2.4.1.2.1.1 Annotation. A key assumption for this evaluation is the level of annotation of the audio files from the Technical Area 2 team. It is assumed that each segment will be annotated whether speech is present or not. When speech is present, the annotation will provide which language and speaker are being used for the segment as well as whether a key word is spoken during the segment. This will be the basis for scoring the performance of the systems provided by Technical Area 1 development teams.

2.4.1.2.1.2 Trials. SAIC proposes presenting the Technical Area 1 development system with a set of audio files which will be as the basis for the performance evaluation for the four RATS technologies. The RATS BAA defines the segmentation for SAD, LID, SID and KWS as 3 seconds. Each instance of voice will be segmented and scored. This should allow for direct comparisons to previous NIST and industry technology evaluations which use 3 second, 30 second, and 45 second segmentation. Technical Area 1 development systems will be asked to make a decision specific to the evaluated RATS technology (SAD, LID, SID and KWS) for each segment and provide a level of confidence on its decision such that more positive scores indicate greater confidence. MOPs will be calculated on each annotated voice activity and cumulatively over all activities.

2.4.1.2.2 Analysis & Reporting. The three performance measures defined by the BAA will be calculated at the voice activity level cumulatively over each activity within the entire audio file and overall for the entire audio file. SAIC will calculate overall performance achieved for each technology separately, which will be the basis for reporting the performance of the various technologies to DARPA and Technical Area 1 teams as defined by the BAA for each phase. Each MOP will be tested against the desired performance level as outlined in the RATS BAA for each phase at a 95 % level of confidence and reported to DARPA. The results of the analysis of variance will be analyzed using non-parametric statistical techniques such as Friedman's test at the 95% level of confidence.

2.4.1.3 Metrics

2.4.1.3.1 SAD. Speech Activity Detection is a key technology which has been worked on for many years. The major difficulty of the speech detection task in RATS is the variability of the speech and background noise patterns. The noise and voice inconsistency often leads to inaccurate detection of the speech endpoints, by cutting phonemes or passing non-speech events to the speech processing system such as Language Identification detection (LID), Speaker Identification detection (SID) and Key Word Spotting (KWS) when used in the end-to-end configuration envisioned in the BAA.

An important problem in the evaluation of SAD technology is the absence of an accurate

S/N ratio	0–5 dB				5–10 dB			
Gender	M	F	M	F	M	F	M	F
Environment Noise	Music	Music	No Music	No Music	Music	Music	No Music	No Music

Figure 2.4-2. Proposed Experimental Design for RATS

method of checking the correctness of a speech detection algorithm, or for comparing two or more candidate methods. The location of the speech endpoints in the presence of high-level noise becomes a complex task even for experienced phoneticians. The commonly used manual method consists of a rough approximation, followed by a more precise endpoint location with acoustical and visual assistance. The BAA specifies that the Technical Area 2 team collect and annotate audio files which will be annotated at each 200 ms segment of the audio. This represents a significant research challenge as the current state of the art of manual or semi-automatic annotation of speech segments is at the 500 ms level. This challenge has a significant impact on the development of the experimental design for evaluating the algorithms developed by Technical Area 1 teams for RATS. SAIC proposes to work with the Technical Area 2 team to determine the level of segmentation that will be delivered and adapt the evaluation design as required if the BAA level of segmentation is not realized.

Previous Evaluations of SAD technologies typically were evaluated over various dimensions of interest but very few have the SNR level specified in the BAA. The RATS BAA specifically asks for the data to be collected in environments where the signal to noise ratio (SNR) is less that 10 dB. SAIC believes that it is important to also measure the effect of the noise over previously investigated dimensions such as gender and type of noise. The experimental design proposed will allow for an understanding of the possible interactions of these dimensions and the overall performance.

2.4.1.3.2 LID. Language identification detection is the process of determining if a language is spoken in voice stream and determining which language is spoken from a set of given languages. The techniques typically used in LID algorithms are based one or a combination of the acoustic or acoustic-phonotactic or lexical or prosodic information.

Some of the known problems with past LID evaluations occur in testing across genders, similar languages, collection environments, and in selecting language when non-native speakers are in the audio. Previous evaluations have also experimented with both open and closed sets of languages. Previous Evaluations of LID technologies typically have not been at SNR level specified in the BAA.

As was proposed for SAD, SAIC believes that it is important to also measure the effect of the noise on LID over previously investigated dimensions such as gender and type of noise. The experimental design proposed will allow for an understanding of the possible interactions of these dimensions and the overall performance of the LID algorithms and allow more accurate estimation of effectiveness in operational environments and scenarios.

2.4.1.3.3 SID. Speaker identification detection is the procedure of capturing and processing a speech signal and automatically recognizing the speaker. The dominant technique used in speaker identification is based on the use of Mel Frequency Cepstral Coefficients (MFCC) extracted from the power spectrum as representation of the vocal track and GMM for modeling and classification.

There are many factors or dimensions when considering SID evaluation. These include speech quality, speech modality, speech duration, and speaker population. The presence of background noise severely degrades the performance of speaker identification detection algorithms.

As was proposed for SAD and LID, SAIC believes that it is important to also measure the effect of the noise on SID over previously investigated dimensions such as gender and type of noise. The experimental design proposed will allow for an understanding of the possible interactions of these dimensions and the overall performance of the SID algorithms.

2.4.1.3.4 KWS. Keyword (or word) spotting refers to a proper detecting of any occurrence of a limited number of keywords that would most likely express the intent of a speaker, rather than attempting to recognize every word in an utterance. A critical issue in keyword spotting is the modeling of the non-keyword portions.

As was proposed for SAD, LID and SID, SAIC believes that it is important to also meas-

ure the effect of the noise on KWS over previously investigated dimensions such as gender and type of noise. The experimental design proposed will allow for an understanding of the possible interactions of these dimensions and the overall performance of the KWS algorithms.

2.4.1.3.5 MOPs vs. MOEs. The RATS program seeks to develop technologies in which there will be significant interest among possible transition partners. To facilitate the transition process, SAIC proposes to supplement the Measures of Performance (MOPs) that are the fundamental metrics of the program with Measures of Effectiveness (MOEs) that will more clearly demonstrate the potential of RATS technologies in operational environments. Complete, end-to-end measurement of MOEs is out of scope for Technical Area 3, so SAIC proposes a low-risk and low-cost alternative in which we will combine relevant MOP results with specific operational scenarios to estimate MOEs. Such estimates will be based on clear assumptions about missions and resources and will include confidences levels of predicted performance that will allow potential transition partners to understand the likely impact of RATS technologies.

An example scenario could be based on the HPCP intercept described in Section 2.4.2.1. The full scenario would include a number of operationally relevant factors such as the number of hours of traffic recorded per day, the number of linguists available, the rate of translation of raw audio, and the performance of linguists on large audio streams (which can be estimated from the data collection task). It would include a mission—for example detection of a series of targeted individuals and keywords relevant to a counter-IED mission. These scenarios would align with those used during evaluation, so that data and results would be directly applicable.

Overall MOE performance would be calculated based on the end-to-end relevant MOPs. The overall MOPs would in turn be estimated by joint probabilities, with limited end-to-end testing to validate the assumptions. For example, the end-to-end probability of detecting speech and speaker can be expressed as:

$$P(SID \cap SAD) = P(SAD) * P(SID \mid SAD)$$

Where P(SID) and P(SAD) are the probabilities of correctly identifying speaker and speech, respectively, for a given audio sample. We can estimate P(SID | SAD) from the measured results on P(SID), but this estimate will be inaccurate. By doing limited end-to-end testing, we can provide both a more accurate performance estimate and a confidence interval around that estimate. Certain tipping-and-queuing scenarios would require correct identification of speech, language, and speaker, introducing additional challenges to the overall performance estimates that should be addressed through limited end-to-end testing.

From estimates of end-to-end MOPs, MOEs of operational interest such as % reduction in translator effort/unit of audio and % of total targets correctly identified can be estimated from simple workflow models and scenario parameters. Similar approaches have been demonstrated successfully in RDEC and other related IC programs to estimate mission impacts from performance measures.

2.4.1.4 Conducting Evaluations

2.4.1.4.1 Phase 1 Evaluation. The Evaluation team proposes a flexible design for Phase 1 evaluation that will adapt to the progress achieved by algorithm designers, in the event their systems are not fully implemented by the BAA schedule ("6 weeks before the end of each phase"). This flexible approach, along with our planned early implementation of the Evaluation Framework and Evaluation Specification, will prepare us to execute phase 1 evaluations at any time after month 12, allowing algorithm developers capable of meeting phase milestones early to be tested whenever they are ready. Previous integration with the Evaluation Framework (section 2.4.3), and successful completion of dry run evaluations, will be used to increase the likelihood that the end of phase integration will proceed smoothly, and support from our integration team will be provided to deal with any unanticipated challenges during this critical period. SAIC will confirm successful integration by running a limited set of tests, automatically verifying results against developer-reported results on a subset of the

training data and hand-verification of results on a limited subset of evaluation data.

The evaluation itself will consist of a series of runs based on our experimental design (section 2.4.1.2).Using the results from each run, SAIC's evaluation framework will automatically compute both overall and cross-dimensional results for the program metrics, including ROC/DET curves and confidence intervals. These detailed results will serve as the basis for our analysis and results reporting. As part of the automated analysis, SAIC will create confusion matrices for languages, speakers, and keywords that will help identify strengths and weaknesses of different algorithmic approaches and identify opportunities for improvement in phase 2 and beyond.

SAIC will conduct separate tests of LID and KWS using data from speakers that were not part of the training set. We will either include or exclude these results from aggregate measures of LID and KWS performance based on guidance from DARPA. Results from these open set tests will allow us to determine the dependence of LID and SID performance on previous exposure to specific speakers—a critical consideration in future classified testing as well as potential operational use.

2.4.1.4.2 Phase 2 and 3 Unclassified Evaluation. The Evaluation team will conduct the Phase 2 and Phase 3 unclassified evaluations at the same unclassified laboratory facility used in Phase 1 testing. Algorithm developers will be expected to train SAIC evaluators on the use of their systems, especially in the areas of training on speakers, and entry of keyword lists, no later than eight weeks prior to the end of phase 2 in order to prepare for testing on classified data.. We anticipate a 2-day training period per system, though more or less may be necessary depending on the degree of familiarity gained during Phase 1 and the number of changes since those tests. SAIC will repeat the testing conducted in previous phases with new data not previously encountered by Technical Area 1 systems. In addition to testing SAD, LID, SID, and KWS as in Phase 1, SAIC will train algorithms to recognize new speakers and keywords and evaluate performance on SID and KWS in such scenarios. This will determine the "trainability" of algorithms—critical for understanding results from classified evaluations where speakers and keywords cannot have been part of the training data.

2.4.1.4.3 Process for Shifting from Unclassified to Classified Evaluation. During Phases 2 and 3, SAIC will conduct classified testing using the same methodology for evaluation as the unclassified testing, executing testing in a sensitive compartmented information facility (SCIF) using classified data provided by the government and previously annotated by SAIC translators (see section 2.4.2.4 for details on annotation).

SAIC has extensive experience in shifting evaluations from unclassified to classified environments. During the RDEC program, SAIC conducted evaluations of over 40 different technologies that moved from unclassified to classified environments and data. We have developed a streamlined process to facilitate the transition—a process we propose to leverage on RATS.

Our process is based on the right combination of staff, policies, and control of the environment. For smooth transitions, it is critical to have all key staff members fully cleared and able to work on the high as well as the low side, as we do. This eliminates potential down time for training that can be very costly in classified environments. Policies must be in place to satisfy all security requirements so that there are no disruptions due to security violations or other problems. Most importantly, the unclassified environment must mirror the classified environment as closely as possible, minimizing the risk of unexpected integration problems inside the SCIF. The Evaluation Team will ensure consistency of the environment by using a single configuration of the Evaluation Framework in both unclassified and classified environments. The Evaluation Framework will reside on paired high-end server systems, one for the high-volume Speech Activity Detection (SAD) metric, while the second server system will focus on the smaller data-sets associated with Language Identification (LID), Speaker Identification (SID), and Key Word Spotting (KWS) testing. There will be separate but identically configured pairs for unclassified and classified testing. The

configuration of these systems will include removable hard-drives in order to support multiple possible operating systems (e.g. Linux and Windows) and transfer of pre-loaded software into classified environments. If algorithm developers' solutions include hardware, their systems will connect to the Evaluation Framework servers through a LAN connection.

SAIC has a wide variety of SCIF spaces available for use in RATS evaluations. We have specifically identified our own classified laboratory facility at the 4001 North Fairfax Drive, Arlington VA (CAGE Code 0PSG0) as the preferred location for RATS due to its proximity to DARPA, our familiarity with the security protocols and procedures, and the availability of storage space for RATS equipment and data. However, in case of difficulties establishing a Co-Utilization Agreement (CUA) with the current certifying authority for this SCIF (Air Force Research Laboratory), SAIC has identified multiple potential backup sites. Further details on our security approach are described in our draft security plan in Appendix B.

2.4.2 Data

Equally critical to evaluation methodology is the proper treatment and handling of data. SAIC will work closely with the Data Collection team to assure that collection and annotation protocols support planned research and evaluation goals while maximizing the operational relevance of RATS technical progress.

2.4.2.1 Scenario-driven evaluation design

SAIC believes the most effective way to focus evaluation is to develop a set of operationally relevant scenarios based on notional concepts of operations (CONOPS), using these scenarios as the basis for the selection of evaluation data, performance tasks, and subsequent analysis. SAIC will develop several scenarios at the start of Phase 1, drawing on a broad set of experts with experience in the operational use of speech analysis technology within DOD and the IC from across the company. The completed scenarios will be presented to DARPA, and based on their recommendation scenarios will be selected for use.

SAIC will work closely with the Data Collection team during their design process to communicate these primary scenarios and encourage them to tailor the collection design to emphasize the scenarios. One potential issue is that it may not be possible for the Data Collection team to collect all data in a manner consistent with the scenarios. In this case, SAIC will recommend that the test data of the test/training partition include sufficient relevant data to support scenario-based testing and analysis. SAIC will further recommend that a subset of the training data with similar characteristics to the test partition be created and labeled for the algorithm developers, allowing them the opportunity to learn specific and interesting characteristics of the scenario-based data while using the remainder of the training data for overall task learning and as background material.

Figure 2.4-3 illustrates potential operationally relevant scenarios and some of the considerations for data collection related to their use.

2.4.2.2 Data Partitioning

The Evaluation Team proposes that it work closely with the Technical Area 2 team to characterize all of the data produced and ensure that the training data characterization is similar to the evaluation data. This should allow for consistent evaluation to be conducted during devel-

Scenario	Network Characteristics	Length	# Languages / Speakers	Noise Sources / Recording considerations
High power cordless phone intercept	Full duplex, analog	Variable, short	1–2 / 2 - 4	Interference, variable signal, background voice, background noise
Cell phone intercept	Full duplex, digital	Variable, short	1–2 / 2 - 4	Interference, variable signal, background voice, background noise (including vehicle)
Radio intercept	Half duplex, analog	Variable, short	1–2 / 1 - 6	Interference, variable signal, background voice, background noise (including vehicle)
Covert microphone	Simplex	Variable, long	1–5 / 2–20	Varying distance from recording, background voice, background noise
Radio broadcast monitoring	Simplex, analog	Regular, long	1 / 1	Interference, interruptions

Figure 2.4-3. Potential Operationally-Relevant Scenarios

opment by the Technical Area 1 teams and during independent evaluation by the Technical Area 3 team. This characterization will include but is not limited to the dimensions outlined in section 2.4.1.2. The Evaluation Team will use this characterization of the audio to create evaluation audio files that have certain characteristics to support the factorial experimental design. This process will allow the Evaluation Team to provide a comprehensive set of performance and the analysis of variance testing on the Technical Area 1 systems and provide DARPA with criteria for decision making at the end of each phase.

SAIC proposes that the Technical Area 2 team isolate a set of data that Technical Area 1 team will not have access prior to Phase 2 evaluation. In addition to the planned 10% of data for evaluation, this data will include all data from one of the 15 languages, all data from small set of speakers, and a set of keywords not provided for training to the Technical Area 1 teams. This will allow the Evaluation Team to conduct measurement on the time and skills needed for independent analysts to train Technical Area 1 algorithms in language, speaker and key words. This experience will also enable the Evaluation Team to conduct rehearsals for the evaluation of the classified audio where it is anticipated none of the speakers and key words will be previously available to the Technical Area 1 or Technical Area 2 teams.

2.4.2.3 Annotating Classified Data Consistently with Unclassified Data

A critical aspect of successful classified evaluation is the availability of classified data that is annotated in a manner consistent with that of the unclassified data used for algorithm development. Without such consistent annotation, it will be impossible to generate evaluation results that can effectively demonstrate and measure the capabilities of RATS technologies.

SAIC has identified trained, cleared linguists available to annotate the classified data sets. To achieve consistency with the unclassified data, SAIC proposes to have these linguists perform limited annotation of unclassified data using the tools and processes of the Technical Area 2 performer. SAIC recommends limited testing of inter-rater consistency of our annotators with those of the Technical Area 2 performer to identify any potential biases or other issues with their annotations. After producing limited amounts of unclassified training data, SAIC's annotators will work in our classified facilities to annotate the classified data in preparation for Phase 2 and 3 evaluations.

2.4.3 Evaluation Framework

The Evaluation Framework was initially developed under the DARPA Machine Reading program. SAIC proposes to re-use the code base already developed as the core of the RATS Evaluation Framework. The framework provides basic capabilities common to many evaluation tasks such as serving test data, scoring, recording results and generating reports. On top of the core services, RATS specific components such as a custom scoring engine, appropriate GUIs and reporting will be built.

The framework has a component based architecture which supports development of customized data retrieval and scoring engine modules. For RATS the scoring engine will be modified to score results using the relevant metrics by comparing system output to ground truth and the data retrieval modules will be customized to serve the audio samples for each test. The Visualization component will be used by evaluators to aid analysis and reporting and by researchers during the training phase to gauge system progress.

2.4.3.1 Concept of Operations for an Evaluation

Evaluations will occur using the RATS evaluation platform. The evaluation platform is a SOA solution providing test data sets, scoring, metric calculation, results logging and visualization as detailed in **figure 2.4.3.1-1**.

The RATS Evaluation Framework exposes several interfaces with which RATS systems will be required to interact during the evaluation, including audio sample retrieval and metric scoring. All interfaces are available over both SOAP and REST protocols to allow the widest possible range of clients to connect and make use of the framework. XML Schemas defining request and response formats will be provided early in the first phase to allow system developers ample time to integrate with the framework.

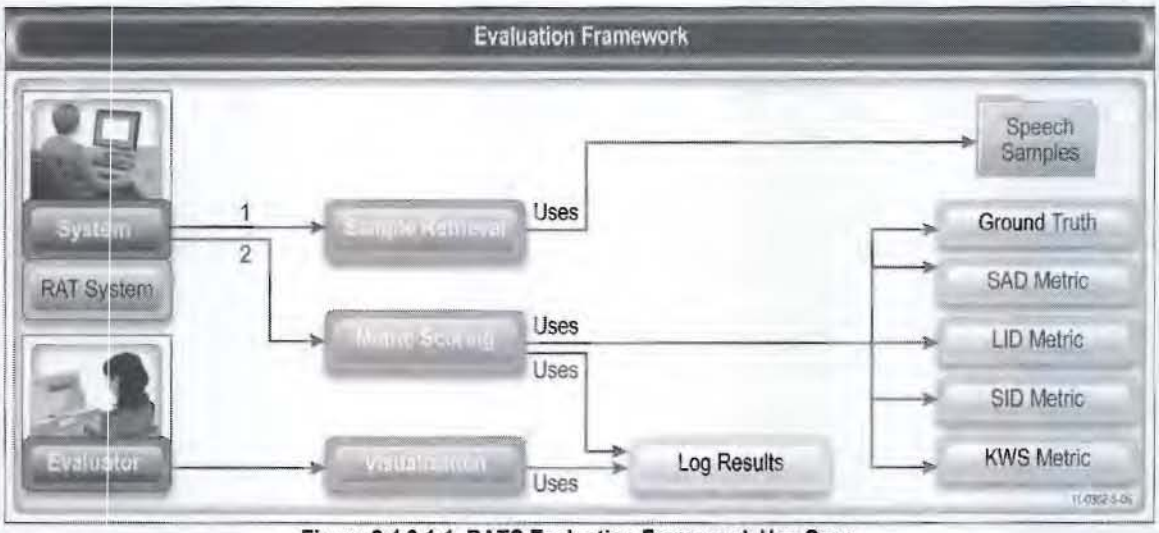

Figure 2.4.3.1-1. RATS Evaluation Framework Use Case

Evaluation begins with the system under test (SUT) making connection with the evaluation platform via an HTTP request to retrieve a test plan. The evaluation platform will reply with a set of URIs used to retrieve audio sample files for that test along with speaker sets for SID and a set of keywords for KWS.

The SUT will process the samples included in the test manifest, generating a result matrix for each task which it will submit to the scoring engine. A scoring report will be generated and returned to the SUT.

Once scoring is complete a results log is generated and stored in a database of results. The visualization component uses the results log to provide a GUI (**figure 2.4.3.1-2**) for navigating a system's performance over an audio sample. The GUI allows any user running the platform, evaluator or researcher, to view performance over arbitrary segments of the test sample including all metrics and a graphical representation of P_{miss} and P_{fa} over each speech event.

2.4.3.2 Evaluation Framework Architecture

The evaluation system components interact with each other and the system under test (SUT) to execute, record, and score evaluations of unclassified or classified audio data (**figure 2.4.3.2-2**).

Evaluation System GUI. When results are provided by the system under test, the Progress Display functionality will show the audio being processed, the results of the algorithm processing, ground truth value and incremental scoring. The Results Display functionality provides an interface for reports summarizing scoring of algorithm results across multiple tests, types of tests, algorithm team, and historical results.

Audio File Server. Control of test execution is managed through this component, in response to requests from the system under test. Data will be supplied to system under test as URIs to the files containing the samples in the test. Files supplied for test will be in the same format as those supplied as training data at the beginning of the phase. This component responds to requests from the system under test to control test execution, for example, requests for the next test to be initiated. Once the system has processed all data in the test it can then send its complete result set to the scoring engine for processing.

Scoring Engine. The scoring engine is activated by a call from the system under test when it has completed processing the samples provided by the audio file server. The SUT will send with the request a matrix for each test metric with results by segment, along with time stamps indicating start and end of processing time. This will support testing of latency requirements described within the FAQs for RATS. The engine will calculate the results of each metric over the sample set by comparing system results to annotated ground truth, then save results into a database which is used by the progress display GUI.

Data Preparation. SAIC will develop several tools for processing the data sets supplied by the Data

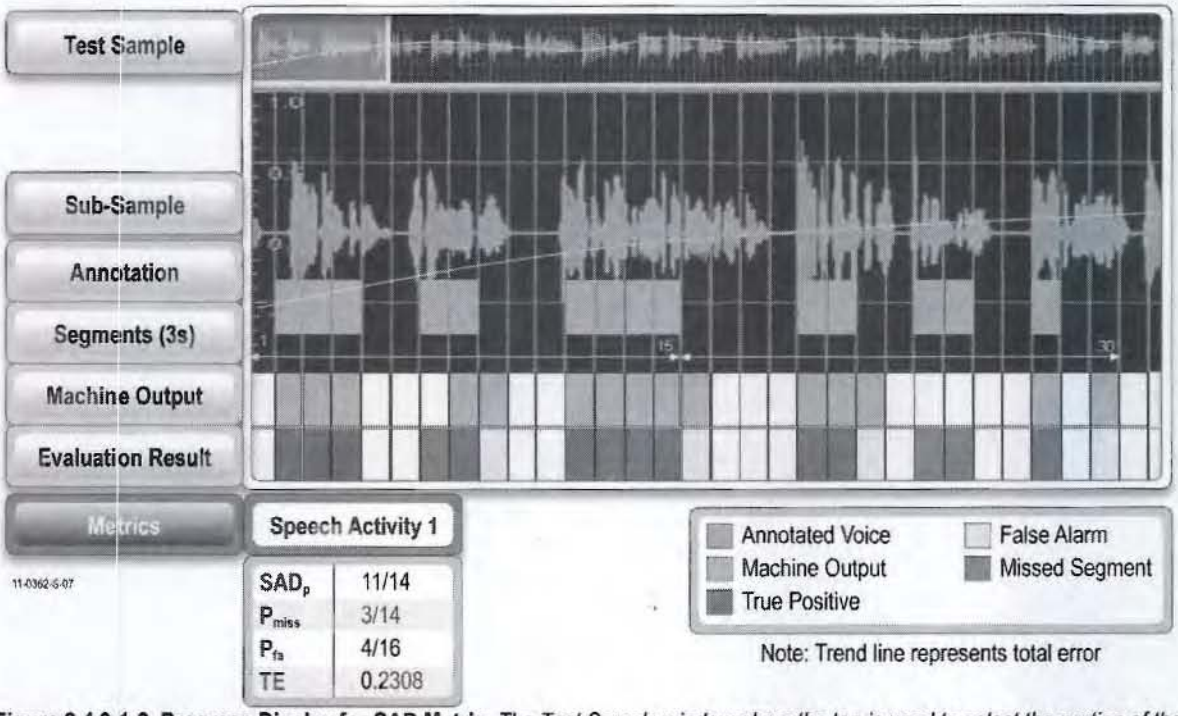

Figure 2.4.3.1-2. Progress Display for SAD Metric. *The Test Sample window along the top is used to select the portion of the sample to be viewed. Note that all metrics can be selected by a user for display on a similar GUI*

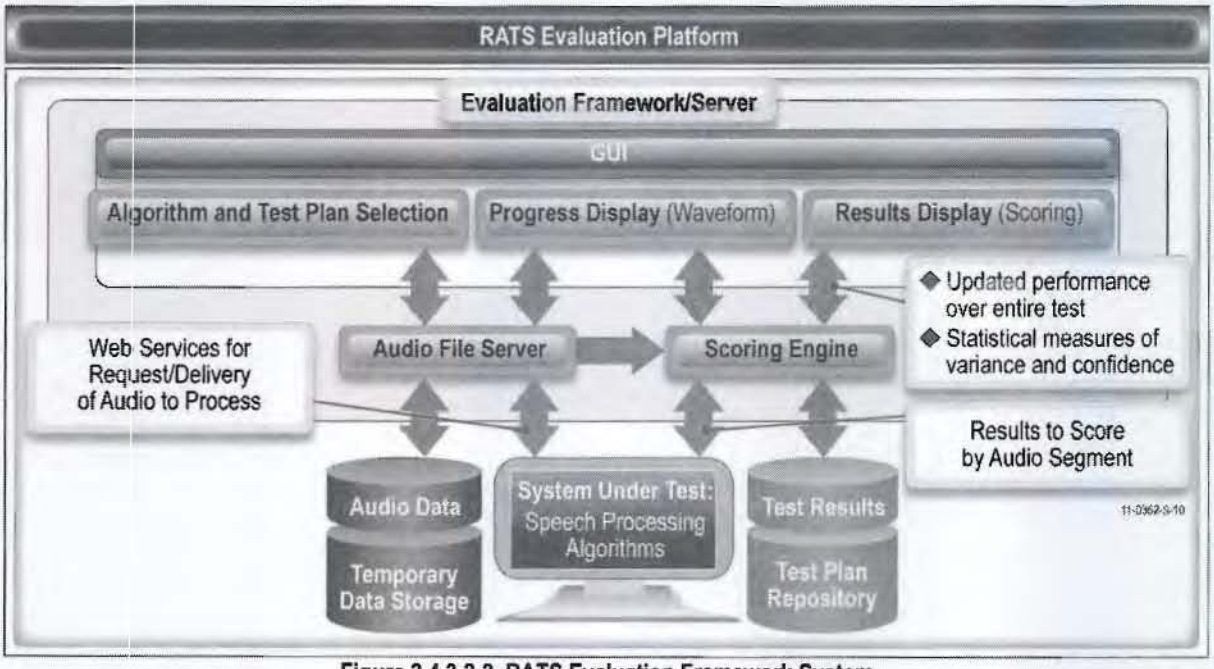

Figure 2.4.3.2-2. RATS Evaluation Framework System

team. These include Test Plan Generation tools to simplify preparation of evaluation test plans and relevant sample sets, audio file conversion tools for preparation of the data for ingestion by the system under test, and ground truth extraction and for determining and building a database of the characteristics of the datasets.

2.4.3.3 Component Development Plan

SAIC brings an existing evaluation framework code base to RATS Evaluation system devel-

opment representing approximately 55% of the capability needed to support the program (**figure 2.4.3.3-1**). This framework already supports a highly automated, loosely coupled, services-based model for providing evaluation data to a set of heterogeneous systems under test. We plan to extend this framework to handle large audio data sets, score the results of speech processing, and provide the graphical display of results.

Figure 2.4.3.3-1. Approximate Percent Completion of the RATS Evaluation Framework

These components will interact to allow the evaluation user to select a test scenario and test data set, execute the test, observe progress during execution, and then report trends and comparisons between algorithms. Required to support these capabilities is integration with third-party code libraries developed for the various audio file representation formats that may be needed by the speech processing algorithms.

The first components to be extended will be the Test Plan Selection GUI and the Audio File Server. The Audio File Server is anticipated to be required to support audio header and data file processing supporting functions as described in **figure 2.4.3.3-2**.

Next will be implementation of the interface for the SUTs to report results to the evaluation system. These results will be scored on an ongoing basis by the Scoring Engine and displayed by the GUI as the Progress Display. Results will be stored and maintained in a database by the Evaluation system both during and after test plan execution, to allow for interruptions and restart of tests by the SUTs. This capability to stop and restart test execution from intermediate steps is made necessary by the quantity of audio data needed to provide statistically significant scoring results, and by the necessarily limited evaluation timeframe at the end of each phase. The final results will be accessible through a Results Display, a tabular presentation comparing algorithm performance against prior evaluations, or against other algorithm result.

Compression / Decompression	
Header Extraction	Formats including NIST SPHERE with self-describing file header specifying number of samples, the sampling rate, the number of channels, and the kind of sample encoding, as well as whether the speech data are compressed or not
Decompression	Utilities for lossless compression /decompression, ratios up to 2:1 on 16-bit audio files
Manipulation and Extraction	
Selection Extraction	Manipulate SPHERE files (compress or uncompress using shorten, read or modify header contents, remove the header, extract portions from waveform files, de-multiplex two-channel files, etc).
Selection of Sampling Rate, Channels	Single or dual channel, 16-bit PCM or 8-bit mu-law, any sampling rate. Options for controlling output include conversion of mu-law data to PCM, selecting one channel from a two-channel input file.
Waveform, Sample Rate Conversion	Provide alternative file formats (AU, AIFF, and more) and changes in sampling rate, etc.

Figure 2.4.3.3-2. Table of Audio Data Processing Algorithms to be Implemented

2.5 Comparison with Current Technology

The RATS program intends to advance the state of the art in speech analysis technology in three distinct areas: performance in noisy environments, individual performance levels for detection, language and speaker identification, and keyword detection, and overall end-to-end system performance. Achieving the RATS BAA performance metrics will represent a significant advancement in the state of the art for each of these technologies, particularly within a noisy environment. The resulting high performance, end-to-end system would allow DARPA's customers to seriously consider utilization of speech analysis technology for tipping and cueing in intelligence analysis. If that performance can be expressed within the language of MOEs relevant to the operational community for the RATS program can have a significant impact on the overall effectiveness of many collection and analysis efforts.

The evaluation approach proposed by SAIC builds on previous evaluations in Speech Detection. RATS Technologies (SAD, LID, SID and KWS) have been evaluated in various forms for the past 50 years. Some of the earliest advancements in voice activity compression and voice activity detection occurred in telephony and was spearheaded by Bell Labs[1]. Evaluation of these technologies in the US continued at Texas Instruments (TI) in 1981 and since 1984 have been spearheaded by NIST[2] in its Rich Transcription[3] (RT) Speech Activity Detection[4] (SAD) series, Language Recognition Evaluation[5] (LRE) series, Speaker Recognition Evaluation[6] (SRE) series and Spoken Term Recognition[7] (STR) series. The Computer Sciences Laboratory for Mechanics and Engineering Sciences[8] (LIMSI) in its project CHIL – "Computers in the Human Interaction Loop" has included both SAD and SID technologies in its CLEAR 06 and 07 Workshops.

NIST's Rich Transcription (RT) evaluation series promotes and gauges advances in the state-of-the-art in several automatic speech recognition technologies including SAD. NIST began working in the area of Automatic Speech Recognition (ASR) in 1984, before DARPA's Speech Recognition program, with the development of quantitative measures of performance for ASR. NIST published its first benchmark tests in the Proceedings of the Speech Recognition Workshop sponsored by DARPA in February 1986. In 2003, NIST implemented the first tests in the DARPA Effective, Affordable, and Reusable Speech-to-text (EARS) Program. The goal of this DARPA Program is "Rich Transcription" – providing not just a text stream, but a rich transcript that includes metadata - as the output of an ASR system. The focus of the EARS efforts are on both Broadcast News and Conversational Telephone-based Speech (CTS), for English, Chinese, and Arabic

The latest SAD efforts at NIST was conducted in 2005 and 2006 Spring Meeting Recognition efforts. The evaluation techniques for SAD in this most recent series include the same performance measures as defined the RATS BAA as well as detection cost functions defined as a weighted sum of the False Alarm and Missed probabilities. The method for analysis has been predominantly through ROC curves and DET curves. The CHIL efforts in SAD were largely in support of the NIST RT effort and used the same evaluation metrics as NIST as well as Mismatch Rate (MR), Speech Detection Error Rate (SDER), and Non-Speech Detection Error Rate (NDER). Audio Segmentation for SAD has been addressed differently over the years but most recently appears to be in the sub second level of granularity and consistent with the segmentation proposed in the RATS BAA. The metrics and segmentation in this proposal are consistent with the NIST past SAD evaluation efforts but build and extend those efforts to emphasize measurements of interest to potential operational users. The proposed experiment design will provide DARPA with a better understanding of the interaction effects which may exist between gender, type of noise and level of noise in detecting speech activity.

NIST's Language Recognition Evaluation (LRE) series was established to baseline the performance capability of language recognition of

conversational telephone speech. It started in 1996 and has been ongoing to the present starting in 2003. The NIST LRE task is defined as given a segment of speech and a language of interest to be detected (i.e., a target language), to decide whether that target language was in fact spoken in the given segment (yes or no), based on an automated analysis of the data contained in the segment. This is essentially the same task as outlined in the LID portion of BAA. The latest evaluation metrics in LRE are the same as outlined in the RATS BAA and they include detection cost functions defined as a weighted sum of the False Alarm and Missed probabilities for LID. The predominant analysis approach for LRE has been both ROC and DET curves like SAD. Speech Segmentation has been addressed by looking at different durations of test conditions of 3, 10 and 30 seconds. The metrics and segmentation in this proposal are consistent with the NIST efforts in LID as well as providing DARPA with a look at the interactions between the dimensions of interest. The proposed evaluation framework in this proposal allows the flexibility of building paired language, open set, and closed set testing of LID based on operationally relevant scenarios using the segments produced for RATS.

SID has been addressed by NIST in its' Speaker Recognition Evaluation (SRE) series, which was started in 1997 and has run continuously since. The goal of the SRE series is to contribute to the direction of research efforts and the calibration of technical capabilities of text independent speaker recognition. In the 2010 Evaluation plan for SRE, NIST defines the performance metrics in a manner similar to the RATS BAA and also includes the detection cost function similar to the one defined for SAD and LID. The CHIL efforts in SID were largely in support of the NIST RT effort and used the same evaluation metrics as NIST as well as Mismatch Rate (MR), Speech Detection Error Rate (SDER), and Non-Speech Detection Error Rate (NDER). The SAIC approach investigates the interactions between the dimensions as well the mix and matching of various speakers supported by the evaluation framework, allowing for realistic operational simulations.

KWS is addressed by NIST in its' 2006 Spoken Term Detection (STD) project. This project is an open evaluation series of technologies that search vast, heterogeneous audio archives for occurrences of spoken terms in three languages: Arabic, English, and Mandarin. The evaluation used three different audio sources: Conversational Telephone Speech, Broadcast News and Conference Room Meetings.

In this evaluation, basic detection of performance was characterized using detection error tradeoff (DET) curves of miss probability (P_{Miss}) versus false alarm probability (P_{FA}). In this evaluation plan, they provided a unique definition of trials as a function of the amount of speech in the audio file. As in the other technologies, NIST defined two overall system detection performance metrics (Occurrence and Term) as a cost function of the False Alarm and Missed probabilities. These cost functions were used to discriminate the performance for spurious detection and term specific performance. This proposal provides consistent metrics with prior work as well as diagnostic analysis of the experiment design for improved performance.

SAIC's experience with both the Intelligence Community and DOD customers will allow us to estimate Measures of Effectiveness that will be vital for future transition opportunities and the overall success of the RATS program.

1 G. R. Doddington and T. B. Schalk, "Speech recognition: turning theory into practice", IEEE Spectrum,Vol. 18, #9, pp.26-32, Sept. 1981.

2 D. S. Pallett, "Performance Assessment of Automatic Speech Recognizers", Journal of Research of the National Bureau of Standards, Vol. 90, #5, Sept. – Oct. 1985.

3 National Institute of Standards and Technology, NIST Rich Transcription Evaluations. WWW-site, http://www.itl.nist.gov/iad/mig/tests/rt/

4 National Institute of Standards and Technology, NIST Language Recognition Evaluations. WWW-site, http://www.itl.nist.gov/iad/mig//tests/lre/

5 National Institute of Standards and Technology, NIST Speaker Recognition Evaluations. WWW-site, http://www.itl.nist.gov/iad/mig/tests/sre/

6 National Institute of Standards and Technology, NIST Speaker Recognition Evaluations. WWW-site, http://www.itl.nist.gov/iad/mig/tests/sre/

7 National Institute of Standards and Technology, NIST Spoke Term Recognition Evaluations. WWW-site, http://www.itl.nist.gov/iad/mig//tests/std/

8 Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, Spoken Language Processing Group, WWW site, http://www.limsi.fr/Recherche/TLP/bibsearch.cgi?keyword=speaker

2.6 Statement of Work

The goal of RATS Technical Area 3 is to construct an evaluation framework to measure the performance of the algorithms developed in Technical Area 1 against DARPA-defined metrics.

SAIC will leverage our existing DARPA Machine Reading Program evaluation framework and materials to provide a high quality, efficient solution that supports research progress and fair, comprehensive, accurate evaluation. The SAIC evaluation team will work collaboratively with the Technical Area 1 and 2 teams to define a research environment and develop the evaluation framework to facilitate the independent execution of evaluations. Our integrated approach with the Technical Area 2 data collection task assures smooth integration of data sources into evaluation framework/execution.

The program statement of work is organized into three phases, with Phase 1 as the base task and Phases 2 and 3 as options. The SAIC Evaluation Team will design and apply an evaluation process that measures the progress of the Technical Area 1 Development Teams during each phase of the program.

Scope

This statement of work lists the tasks identified for the Evaluation Team in Phase 1, with tasks for Option Phases 2 and 3, where appropriate. We further describe the implementation of these tasks in the Technical Approach (Section 2.4), and the Work Breakdown Structure (WBS) with milestones and duration of these tasks in Section 2.8. We describe management tasks in the Project Management and Interaction Plan (Section 2.10).

SAIC is the primary organization responsible for each task defined for Technical Area 3, Evaluation. Our consultant, Dr. Paul Cohen, is a subject matter expert on the design and execution of evaluations and will provide an independent peer review on the evaluation activities described herein.

During each phase, SAIC provides the algorithm developers with the evaluation framework. This approach allows the developers and/or SAIC to support informal test runs to identify progress and issues during the algorithm development stage and allows the SAIC evaluation team to refine evaluation methodology and metrics.

At the conclusion of each phase, SAIC will perform evaluation of each Technical Area 1 system using a data set that has been sequestered from Technical Area 1 teams. The evaluation methods will be the same in each phase, scenario-focused data supplied by the Technical Area 2 Team to represent real-world operational scenarios. SAIC will develop several potential scenarios at the start of Phase 1 and work with DARPA to determine the final set for use in the evaluation activities. As described in Section 2.4, we propose a flexible approach that will adapt the Phase 1 evaluation to the progress achieved by algorithm designers.

During Phase 1, we anticipate initial delivery of data from the Technical Area 2 performer during month 6 and drops at regular intervals until the final evaluation period commences. In addition, starting at the end of Phase 2, the SAIC Evaluation Team will perform secondary evaluations using classified "real world" data provided as Government Furnished Information (GFI). The proposed SAIC personnel and facilities are cleared to the SCI level. Our extensive Intelligence Community (IC)/Classified experience assures relevance and assists DARPA with providing access to required facilities and personnel to conduct evaluations.

2.6.1 Develop Evaluation Specification Document

The development of the Evaluation Specification Document is defined below. **Figure 2.6-1** identifies the task by WBS reference 1.1.1, the dependencies and exit criteria and definition of task deliverables.

Objective. To define the evaluation environment including processes and procedures and assure test integrity, fidelity, and interoperability.

Task	Task	Task Duration	Depends On	Exit Criteria (product, event, milestone)	Deliverable
1.1.1	Develop Evaluation Specification	6 Months	Contract award	Initial draft, reviews and updates as required. Final Document delivered	Draft Evaluation Specification Document Final Evaluation Specification Document

Figure 2.6-1. Develop Evaluation Specification Document

Approach. The Evaluation Team will consult with the Technical Area 1 and 2 Teams as it develops the Evaluation Specification Document. Specifics of the implementation of each evaluation will be defined including data, tasks, test protocols, and metrology. The specification document includes the evaluation methodology and identifies the data and graphical presentation format to be used to measure and display algorithmic performance and interface standards. SAIC will publish initial specifications and planning products for comment and review by the Technical Area 1 and 2 Teams with final review and approval of the Evaluation Specification Document from the DARPA PM. We will refine and publish the specification and metrics, as necessary. For Phases 2 and 3, we will update, review, and publish the Evaluation Specification Document to reflect the goals for the specific phase, revised metrics and lessons learned from previous phases.

Benefit of SAIC's Approach. Collaboration with the Technical Area 1 and 2 Teams and timely publication and updates will assure system interoperability in the final test environment.

Definition of Task Deliverables. Evaluation Specification Document

The evaluation specification document will describe the data, tasks, test protocols, metrology and interface requirements associated with evaluations. The document will include a description of the data and graphical presentation format to be used to measure and display algorithmic performance.

2.6.2 Develop Evaluation Framework

The Evaluation Framework development activity includes design and development of the evaluation framework and development of customized interfaces as described in Section 2.4.1. **Figure 2.6-2** identifies the task by WBS reference 1.2, the dependencies and exit criteria and definition of task deliverables.

Objective. Provide a flexible evaluation framework.

Approach. SAIC will work closely with the algorithm developers to develop an evaluation framework that supports algorithm developers in evaluating their own research progress, allowing efficient utilization of our code base across all program performers and minimizing the risk of integration challenges at end of phase evaluations. The evaluation framework is delivered at Month 6 to support development activities and can be refined to accommodate unique interfaces and changes up through month 11 of Phase 1. Additional capabilities and features can be added in Option Phases 2 and 3 as required.

SAIC will develop an evaluation framework, consulting with the Technical Areas 1 and 2 as we enhance the framework to accommodate additional languages and capabilities anticipated in Phases 2 and 3. We will refine the Evaluation Specification Document, as necessary, to finalize the description of the evaluation framework. For Phases 2 and 3, we will update, review, and publish the Evaluation Specification Document monthly to reflect changes as required.

Benefit of SAIC's Approach. The proposed evaluation framework can be leveraged to support research development and determine progress throughout all phases of the program and facilitate a realistic final evaluation conducted over the largest possible set of evaluation data.

Definition of Task Deliverables. Evaluation Framework (Version Release #).

Task	Task	Task Duration	Depends On	Exit Criteria (product, event, milestone)	Deliverable
1.2.1	Design Evaluation Framework	2 Months	Decomposition of BAA Requirements and Collaboration with Technical Area Teams	Design Baseline complete	Framework Version (Release #)
1.2.2, 1.2.3	Develop and Deliver Evaluation Framework	4 Months	Framework Design	Framework available for Development Teams	
1.2.4	Customize Framework Interfaces	1 Month	Requirement to accommodate unique system interfaces	Customize d interfaces	
1.2.5	Refine Framework to accommodate required changes requested by developers in Technical Area 1	1 Month	New Requirements identified by the development team	Framework revision	
1.2.6	Configure Phase 1 Test bed	3 Days	Configure lab environment to facilitate testing	Test bed configured	Configuration Drawing included in *Evaluation Specification Document*

Figure 2.6-2. Develop Evaluation Framework

2.6.3 Evaluation Test Design

Evaluation Test Design activities include development of scenarios for data collection, collaboration with the Data Collection Team to plan and review data collection requirements; characterization and partition of test data and formal evaluation data. This approach ensures that data collection and data partitioning are aligned with program goals and planned evaluation protocols. **Figure 2.6-3** identifies the task by WBS reference 1.3, the dependencies and exit criteria and definition of task deliverables.

2.6.3.1 Develop Scenarios

Objective. Establish a context for data collection.

Approach. SAIC will develop a set of potential scenarios at the start of Phase 1, drawing on a broad set of experts with experience in the operational use of speech analysis technology within DOD and the IC from across the company. The draft scenarios will be presented to DARPA, and based on their recommendation primary scenarios will be finalized for use in the evaluations.

Benefit of SAIC's Approach. SAIC will utilize its experience and knowledge of likely operational scenarios motivating the RATS performance tasks to focus evaluation on problems of likely interest and to support DARPA in transitioning RATS technologies to operational users

Definition of Task Deliverables. The final set of selected scenarios will be included in the Evaluation Specification Document.

2.6.3.2 Collaborate with Data Collection Team [WBS 1.3.2, 1.3.3]

Objective. Establish the requirements for data deliverables

Approach. A Data Collection Planning meeting and periodic reviews will be held with the Data Collection Team to establish the requirements for data deliverables.

Benefit of SAIC's Approach. Establishing the collection and tagging specifications early in the program will assure that end of phase evaluation and Phase 2 and 3 classified evaluations are efficient and effective. An initial data drop and periodic deliveries will facilitate our ability to measure progress and refine the test specification and environment.

Task	Task	Task Duration	Depends On	Exit Criteria (product, event, milestone)	Deliverable
1.3.1	Develop Scenarios	1 month	Start date	Operational Scenarios	Test Artifacts (Final Report)
1.3.2, 1.3.3	Collaborate with Data Collection Team	2 Face-to-face Meetings, 1 day each	Specifications, Operational Scenarios	Collection Requirements	
1.3.4, 1.3.5	Receive drops, Characterize Data, Partition Test and Evaluation Data	Intervals over an 8 month period starting at Month 6 in Phase 1	Receipt of annotated data (incremental deliveries starting in Month 6	Test data segmented from evaluation data. Data characterized.	

Figure 2.6-3. Split Test and Evaluation Data

Definition of task deliverables. The schedule and requirements for the Data Deliverables will be included in the Evaluation Specification Document.

2.6.3.3 Partition Test and Evaluation Data

The Partition Test and Evaluation Data activity includes receipt and verification of data and partition of test data and formal evaluation data.

Objective. Ensure that the partition of data between training and test sets supports scenario-based evaluation while being normalized across important dimensions.

Approach. SAIC will work closely with the Data Collection team, ensuring that the division of data between training and test sets is normalized across important dimensions while sequestering limited data on certain dimensions (speakers, keywords) to supports evaluation of the trainability of research algorithms.

Benefit of SAIC's Approach. This approach ensures that data collection and data partitioning are aligned with program goals and planned evaluation protocols.

Definition of task deliverables. All data artifacts used in the evaluations will be submitted with The Final Report

2.6.4 Conduct Evaluations

Evaluations are performed during each phase. **Figure 2.6-4** identifies the Phase 1 tasks by WBS reference 1.4, the dependencies and exit criteria and definition of task deliverables. **Figure 2.6-5** identifies the additional tasks necessary to conduct Classified Evaluations in Option Phases 2 and 3.

2.6.4.1 Perform (Unclassified) Evaluation

Objective. To assess progress, analyze performance and draw conclusions and recommendations.

Approach. SAIC will perform a final evaluation on each Technical Area 1 System at the end of each phase. The final version of an integrated system is delivered to the SAIC evaluation team as GFI 6 weeks before the end of the phase but 8 weeks is preferable to accommodate required training. The system will be baselined upon receipt "as delivered" through the Configuration Management Process. The SAIC Evaluation Team will work directly with each Algorithm Development Team to perform integration and training prior to final evaluations as required. A preliminary delivery of the system will allow for training before the final system delivery is required. It is expected that the Algorithm Developer Teams will assign resources to support the final training and integration activities. The SAIC Evaluation Team will evaluate the system in accordance with the methodology and technical framework documented in the Evaluation Specification Document and prepare a Final Evaluation Report at the completion of the test event.

Benefit of SAIC's Approach. Training and communication with the development team will promote a solid understanding of the algorithm before the evaluation tests are initiated.

Task	Task	Task Duration	Depends On	Exit Criteria (product, event, milestone)	Deliverable
1.4.1	Perform Phase (n) (Unclassified) Evaluation	40 days	Final Systems are required 6 weeks before end-of-phase. Reports are due 2 weeks before end of Phase.	Final Evaluation Reports complete	
1.4.1.1 1.4.1.2	Receive preliminary and final systems for training	0 Days	Technical Area 1 systems delivered as GFI	Configuration Item number assigned (CM process)	
1.4.1.3	Integrate research algorithms with data and test framework, train system	15 days	Receipt of GFI (Technical Team 1 Systems)	System ready to test	Evaluation Reports
1.4.1.4	Evaluate Integrated Systems	19 days	Integration and training schedule	Test Complete	
1.4.1.5	Analyze Results and Prepare Report	6 days	Completion of evaluation	Results documented	
1.4.1.6	Deliver Report(s)	0 days	BAA Requirement (2 weeks before end of phase)	Report delivered	

Figure 2.6-4. Conduct Evaluations

Task	Task	Task Duration	Depends On	Exit Criteria (product, event, milestone)	Deliverable
1.4.2	Perform Classified Evaluations (N/A for Phase 1)	55 days	Authorization to perform Classified Test	Option Phases 2 and 3 Classified Evaluation Reports Complete	Classified Evaluation Report
1.4.2.1	Perform Security Planning and Documentation	7 days	Required Security documentation received from Technical Area 1 Teams	Security Plan complete	Security Plan
1.4.2.2	Transition equipment from Low to High	3 days	Approved configuration and receipt of classified data	Classified Test Range	N/A
1.4.2.3	Perform Security Audit	1 day	Government Security Compliance Testing Complete	Approval to conduct classified test	N/A
1.4.2.4	Annotate Data	1 month	GFI of Classified Data	Annotated Test Data	Classified Test Report /Test Artifacts
1.4.2.5	Train Systems	12 days	Receipt of systems	Systems ready to Test	N/A
1.4.2.6	Execute tests	18 days	Systems Ready, Data Ready	Test Results	Test Report
1.4.2.7 1.4.2.8	Analyze results and prepare report	12 days	Test Completion	Final Evaluation Report	Classified Test Report
1.4.2.9	Perform Security Sanitization Activities	1 day	Test Completion	Sanitized Lab	N/A

Figure 2.6-5. Conduct Classified Evaluations (Option Phases 2 and 3)

Definition of Task Deliverables. Final Evaluation Report.

The Final Evaluation Report documents progress, results, conclusions and recommendations from the evaluation activities. System

2.7 Intellectual Property

In compliance with DFARS 252.227-7017 Identification and Assertions of Use, Release, or Disclosure Restrictions (June 1995), SAIC provides the following technical data or software rights assertions, including those of our proposed subcontractors.

2.7.1 Noncommercial Technical Data and Software Rights

Figure 2.7-1 indicates that noncommercial technical data with less than unlimited rights is not proposed for use on this contract. In addition, SAIC hereby provides notification that any software created during the life of any resultant contract will be delivered with unlimited rights. Should any of these assertions change, SAIC will inform the Contracting Officer and provide the justification for the restricted rights before delivering the software in question.

During performance of this contract, SAIC intends to modify its Evaluation Framework software that was developed previously under government contract for the DARPA Machine Reading program. In accordance with the data rights clauses of that contract, the government has unlimited rights in the software. Please note that the software does contain commercial third party code subject to various terms and conditions, which are identified in **figure 2.7-2** below.

Technical Data or Computer Software to Be Furnished With Restrictions	Basis for Assertion	Asserted Rights Category	Name of Person Asserting Restriction
None	None	None	N/A

Figure 2.7-1. Noncommercial Technical Data

Technical Data or Computer Software to Be Furnished With Restrictions	Basis for Assertion	Asserted Rights Category	Name of Person Asserting Restriction	Link to License Terms
Sun Java 1.6	Developed at private expense	Open-source license rights(SUN Binary Code License)	SUN Microsystems	http://java.sun.com/javase/6/jre-6u14-license.txt
HP Jena 2.6.0	Developed at private expense	Open-source license rights	Hewlett Packard	http://jena.sourceforge.net/license.html
ARQ	Developed at private expense	Open-source license rights	Hewlett Packard	http://jena.sourceforge.net/ARQ/license.html
ICU 3.4	Developed at private expense	Open-source license rights	IBM	http://source.icu-project.org/repos/icu/icu/trunk/license.html
IRI	Developed at private expense	Open-source license rights	Hewlett Packard	http://jena.sourceforge.net/iri/license.html
xercesImpl.jar	Developed at private expense	Open-source license rights	Apache Foundation	http://www.apache.org/licenses/LICENSE-2.0
Xml-apis.jar	Developed at private expense	Open-source license rights	Apache Foundation	http://www.apache.org/licenses/LICENSE-2.0
Apache Ant	Developed at private expense	Open-source license rights	Apache Foundation	http://ant.apache.org/license.html
NIST Sphere (audio file headers)	Developed at private expense	Open-source license rights	NIST	http://www.itl.nist.gov/iad/mig//tools/sphere_26atarZ.htm
Shorten(decompression)	Developed at private expense	Open-source license rights	Softsound Ltd	http://www.etree.org/shncom.html
SoX(conversion utilities)	Developed at private expense	Open-source license rights		http://sox.sourceforge.net/

Figure 2.7-2. Commercial Computer Software with Less Than Unlimited Rights

2.9 Personnel, Qualifications, and Commitments

2.9.1 Personnel and Commitments

Richard La Valley is selected as the Principal Investigator (PI) of the SAIC RATS Evaluation Team for his experience in designing and executing evaluations for advanced technology programs and analysis of tools in classified environments. Mr. La Valley is proposed as key personnel on this program. He will be supported by Matthew Reardon, as Program Manager (PM) to ensure continuous adherence to schedule and budget in addition to supporting test and evaluation activities as required. Dr. Paul Cohen, Dr. William Hardy, and Dr. Henry Goldberg who are subject matter experts with significant technical expertise in the areas of language transcription technology will support Mr. La Valley as required to develop specifications and analyze test results. Dr. Goldberg and the RATS Evaluation Framework software engineer, Jonathan Herr, bring significant expertise on the development and implementation of the proposed evaluation framework and methodology from the DARPA Machine Reading Program. In addition to the evaluation subject matter experts, we will enlist the services of SAIC linguists specializing in Arabic and Farsi as required to collaborate with the Data Collection Team and help to establish specifications for data annotation. This collaboration in Phase 1 will facilitate the annotation process and use of the annotation tool on classified GFI during the classified test portions of Phases 2 and 3.

Figure 2.9-1 summarizes the hours and percentage of time proposed for the PI Richard LaValley, the named key personnel. Mr. La Valley will devote 50% of his time to the evaluation tasks across all phases. He is an expert statistician in metrics and experimental design with extensive experience in the design and execution of unclassified and classified evaluations. Non-key personnel who will support Mr. La Valley include a PM, Mathew Reardon, who has extensive program management experience within the IC, DOD, and the telecommunications industry. Dr. William Hardy, an expert in voice and speech analysis and metrics, will serve as a subject matter expert on speech transcription technology. Dr. Henry Goldberg, the PI on DARPA's Machine Reading Program, will contribute his expertise in evaluation of speech recognition systems. Jonathan Herr, who created the evaluation framework for the Machine Reading Program, will develop the RATS Evaluation Framework and interfaces leveraging the framework developed for the Machine Reading. Dr. Paul Cohen from the University of Arizona will serve as a consulting subject matter expert to develop the Evaluation Specification Document, refine metrics, assist in data partitioning, and analyze evaluation test results. The SAIC linguists will support collaboration with Technical Area 2 teams and the classified data annotation in Phases 2 and 3. Detailed breakdown of hours are provided in the Cost Proposal (Volume 2).

Concise summaries of SAIC key and significant supporting personnel are included in the resumes section that follows.

Key Individual	Project	Pending or Current	Phase 1	Phase 2	Phase 3
Richard LaValley (PI) SAIC	RATS Tech Area 3	Proposed	1504 hours (54%)	976 hours (52%)	836 hours (45%)

Figure 2.9-1. Mr. La Valley, Experienced in Evaluating New Technologies, will Lead the RATS Evaluation Program to Ensure Successful Conduct of Evaluations

Mathew Reardon — Program Manager

Summary of Qualifications

- Has 10+ years of experience as a professional analyst and strategic planner across multiple disciplines.
- Has 10+ years of project and business management experience.
- Has extensive experience in operations research and analysis across the military and telecommunications domains.
- 18 years of experience as a Naval Officer (active and reserve components).

Education

- Master of Science (with Distinction), Systems Management, Naval Postgraduate School, Monterey, CA, 1997.
- Bachelor of Science, Ocean Engineering, U.S. Naval Academy, Annapolis, MD, 1991.

Clearance: TS/SCI

Work In Related Research Areas and Previous Accomplishments

Science Applications International Corporation, Director of Strategic Plans — *10/2008–Present*

- Manages IC Collaboration and Technology Evaluation & Experimentation cells across the IC established under the umbrella of IARPA's Research & Development Experimental Collaboration (RDEC) program.
- Manages Joint Capability Technology Demonstration support project in support of USD (AT&L).

AOL, LLC., Vice President, Planning & Partner Management, Dulles, VA — *08/2001–06/2008*

- Led strategy and planning for AOL's Access Division, managed operations across global contact center network and marketing partnerships with network of Retail and OEM providers.
- Led teams responsible for developing the strategy, partner management, and optimized channel mix and operations of co-marketing partnerships with nationwide network of DSL, FiOS, and Cable providers. Managed sales and marketing strategy, program implementation, and operations across global call center network.
- Managed teams responsible for forecasting, operations and marketing analysis, strategic & operational planning, and workforce management.

Military Sealift Command, Director of OIF/OEF Contingency Plans, Washington, DC — *03/2003–01/2004*

- Mobilized reservist in support of Operation Iraqi Freedom/Operation Enduring Freedom, planned and scheduled the sealift movement of over 30 million square feet of combat and support cargo with a surge fleet of 120 government and commercial ships.

Chief of Naval Operations, MPN Strength Planner & Program Analyst, Washington, DC — *07/1997–08/2001*

- Developed and executed personnel models resulting in tactical and strategic plans (recruiting, training, retention, and advancement) and MilPers appropriation programming for Navy's enlisted workforce and over 100 specialty skills. Served as Chief of Naval Operations representative to Military Operations Research Society working groups

William C. Hardy, Ph. D. — Subject Matter Expert

Summary of Qualifications

- Has more than forty years' experience as an operations analyst specializing in decision support based on acquisition, organization, interpretation, and analysis of relevant data to produce credible, scientifically defensible answers to questions posed by decision-makers.
- 13 years in conduct of analyses for military communications and command and control systems, 13 years in analysis of commercial telephony, 6 years dedicated to evaluation and development of modeling aids for intelligence analysts.
- Analytical efforts for commercial telephony focused on the problem of measurement and evaluation of quality of telephone services, with emphasis on the measurement and evaluation of user perception of the quality of telephonic speech. Work in this area resulted in innovations in test technology and measurement and analysis of acoustic speech waveforms that earned more than 23 US patents.

Education

- Ph.D in Mathematics, University of New Mexico, Albuquerque, NM, 1970.

Clearance: TS/SCI

Work in Related Research Areas and Previous Accomplishments

Science Applications International Corporation, Senior Metrics Analyst and Technical Fellow *3/2004–Present*

- Responsible for developing measures, test and protocols for evaluation of effectiveness and utility of automated tools for intelligence analysts, such as group detection algorithms, machine text translators, and document retrieval facilities.
- Developed and applied ad hoc, rapid prototypes of numerous data handling and analysis algorithms; innovations in automatic pattern recognition for knowledge discovery; innovative techniques for analyzing scalability of network analysis algorithms.

MCI/WorldCom, Executive Staff Analyst for Measures and Analysis *1/1989–6/2003*

- Responsible for development of data acquisition and analysis tools for measuring and evaluating quality of commercial telephone services.
- Designed and successfully exploited the Service Attribute Test, which provides a viable test platform for collection of data from which to determine relationships between system manifestations of system performance problems and user assessment of call quality.
- Designed, and directed development of MCIs Telephone Quality Measurement System (TQMS) and Voice Quality Evaluation System (VQES), which create capabilities for automated data collection of telephonic voice samples and analysis of the captured waveforms to predict user perception of quality.
- Earned the following TQMS/VQES-related patents that exploit or are based on innovations in speech detection and speech waveform analysis directly related to this effort:

5,748,876 System and Method for Testing Acoustic Modems with Semantically-Encoded Waveforms	6,556,677 Single-Ended Echo Cancellation System and Method	6,246,978 Method and System for Measurement of Signal Fidelity from Samples of Telephonic Voice Signals	7,085,230 Method and System for Evaluating the Quality of Packet-Switched Voice Signals
6,115,465 System and Method for Modifying Voice Signals to Avoid Triggering Tone Detectors	6,564,181 Method and System for Measurement of Speech Distortion from Samples of Telephonic Voice Signals	6,370,120 Method for Predicting Perceived Quality of a Packet-Switched Telephone Connection	7,099,282 Determining the Effects of New Types of Impairments on Perceived Quality of a Voice Service
6,130,943 Method and Apparatus for Suppressing Echo in Telephony	6,985,559 Method and Apparatus for Estimating Quality in a Telephonic Voice Connection	6,553,061 Method and Apparatus for Detecting a Waveform	7,154,855 Method and System for Determining Dropped Frame Rates over a Packet-Switched Transport

Henry G. Goldberg, Ph.D — Subject Matter Expert

Summary of Qualifications

- Has experience in the design and execution of evaluations for technologies developed under the DARPA Machine Reading Program. Creates the materials, specifications, and processes necessary to maximize advances in machine reading research and to measure these advances clearly.
- Has 17 years of experience in research, design, construction, and operation of knowledge-based systems for detection and discovery of financial crimes, fraud, and other behaviors of interest to intelligence, law enforcement, and regulation for government and the private sector.
- Has experience evaluating AI methods and systems, especially speech recognition and NLP.
- Has knowledge-based systems engineering expertise in the application of data mining and knowledge discovery, and exploitation of temporal and network patterns in large databases.
- Authored and co-authored 20 articles and publications in speech recognition systems, evaluation of pattern recognition performance in AI systems, link analysis, and innovative applications of AI involving natural language processing, temporal pattern recognition, and link analysis.

Education

- Ph.D., Computer Science (AI), Carnegie-Mellon University, 1975
- B.S., Mathematics, Massachusetts Institute of Technology, 1968.

Clearance: Top Secret

Work in Related Research Areas and Previous Accomplishments

Science Applications International Corporation, Chief Scientist and Engineer — *1/2009–Present*

- Serves as PI of the Evaluation Team under the DARPA Machine Reading Program, whose goal is to make information from natural language corpora available to reasoning systems. The Evaluation Team's mission is to prepare materials, design and execute evaluations, and coordinate common technologies among the 3 performing teams.
- Leads experiments and metrics development for intelligent systems evaluations, especially in the areas of natural language, knowledge representation and reasoning, and machine learning.

FINRA (formerly NASD), Rockville, Md., Systems and Knowledge Engineer, Securities Regulation — *10/1996–12/2008*

Special Projects, Business Solutions Department — *9/2007–12/2008*

- As principal technical officer studying alternatives for migration of installed base of detection scenarios, provided consultation and review of technology strategies to meet the challenge of increased volumes, multiple sources of market data, and migration to new hardware platforms.
- Evaluated systems for text mining as part of the NASD Sonar system as senior knowledge engineer and KDD specialist.

Director, KDD Team, Market Regulation Department — *6/2001–9/2007*

- Directed a team of knowledge engineers and programmers engaged in maintenance and ongoing new development of regulatory surveillance programs, patterns and data mining solutions.

Senior KDD Specialist, NASD Technology — *10/1996–6/2001*

- Innovative knowledge-based systems for fraud and violation detection in financial markets.

U.S. Treasury, Financial Crimes Enforcement Network, Vienna, Va., Senior Research Computer Scientist — *7/1991–10/1996*

- KDD and knowledge-based technology for detection of financial crimes and money laundering.

U.S. Courts, Federal Judicial Center, Washington, D.C. — *7/1977–7/1991*

- Application of advanced information processing to court administration.