Section I Program Overview

Microphones in modern electronics are ubiquitous and constantly listening for speech signals - using voice driven technologies to interact with people – making it likely that our speech is recorded nearly everywhere. Speech is often transmitted to the cloud for processing, analysis, and storage, where analytics could reveal patterns of life. Once speech has been collected, existing technologies, such as Speaker Identification (SID) and forensic speech analysis (FSA), pose significant threats to speaker privacy. The goal of the Anonymous Real-Time Speech (ARTS) program is to develop novel systems that will modify spontaneous speech in real-time to protect an individual's privacy.

Speaker identification determines an unknown individual's identity based on samples of their speech. SID tools rely on models derived from speech segments taken from known individuals. With many individuals enrolled in a SID system, speaker identification is a 1:N comparison, where an utterance is compared against multiple templates in order to answer the question, "Who is speaking". The ARTS program is seeking novel methods for near real-time Speaker De-Identification (SDID) capabilities to prevent SID systems from attributing the recordings to the actual speaker, even with the individual enrolled in the SID database.

An individual's speech also has characteristics that do not change often. Some of these features can reveal aspects about the speaker, such as dialect, gender, and age. Although such properties can change over time – for instance, some people will gradually affect a new dialect after living abroad for many years – in general, these speech characteristics remain static for many individuals. The ARTS program is looking for novel methods to modify speech so that three existing static traits of dialect, gender, and age, will be removed and replaced with a profile of pre-selected traits.

There are also methods to analyze speech signals that can assess a temporary state of a speaker, like a short-term emotional or physical state. For example, statistical features extracted from speech segments can be used to predict if a person is angry, sick, or intoxicated. Many of these approaches leverage AI/ML techniques to discriminate between different classes of temporary states. The ARTS program is looking for approaches to defeat statistical classification of these short-term states.

With these different threats to privacy, the ARTS program is looking for innovations to anonymize speech in near real-time by addressing three Technical Areas (TA).

- **Technical Area 1 (TA1) – Speaker De-Identification (SDID):** Modify speech signals so that SID systems are unable to recognize the speaker.
- **Technical Area 2 (TA2) – Modification of static traits in speech:** Remove existing traits in speech signals related to dialect, gender, and age, and replace them with pre-selected traits.
- **Technical Area 3 (TA3) – Removal of dynamic traits in speech:** Modify speech signals so that extracted features can't be used to differentiate short term states such as emotion and psychological state.

There have been recent advances in speech synthesis that produce promising results in areas adjacent to the goals of ARTS. These capabilities rely on new technologies that could likely play a role in ARTS systems. However, to satisfy the privacy goals of the ARTS program, successful innovations should also be useful; they should work in real-time and the quality of the speech should not noticeably degrade. Thus, solutions will be constrained by three utility requirements: latency, understandability, and naturalness.

With three TAs and three utility constraints, the ARTS program is driving to find methods to anonymize speech that will address six different and competing dimensions simultaneously. **Figure 1** highlights the six different dimensions of the problem.

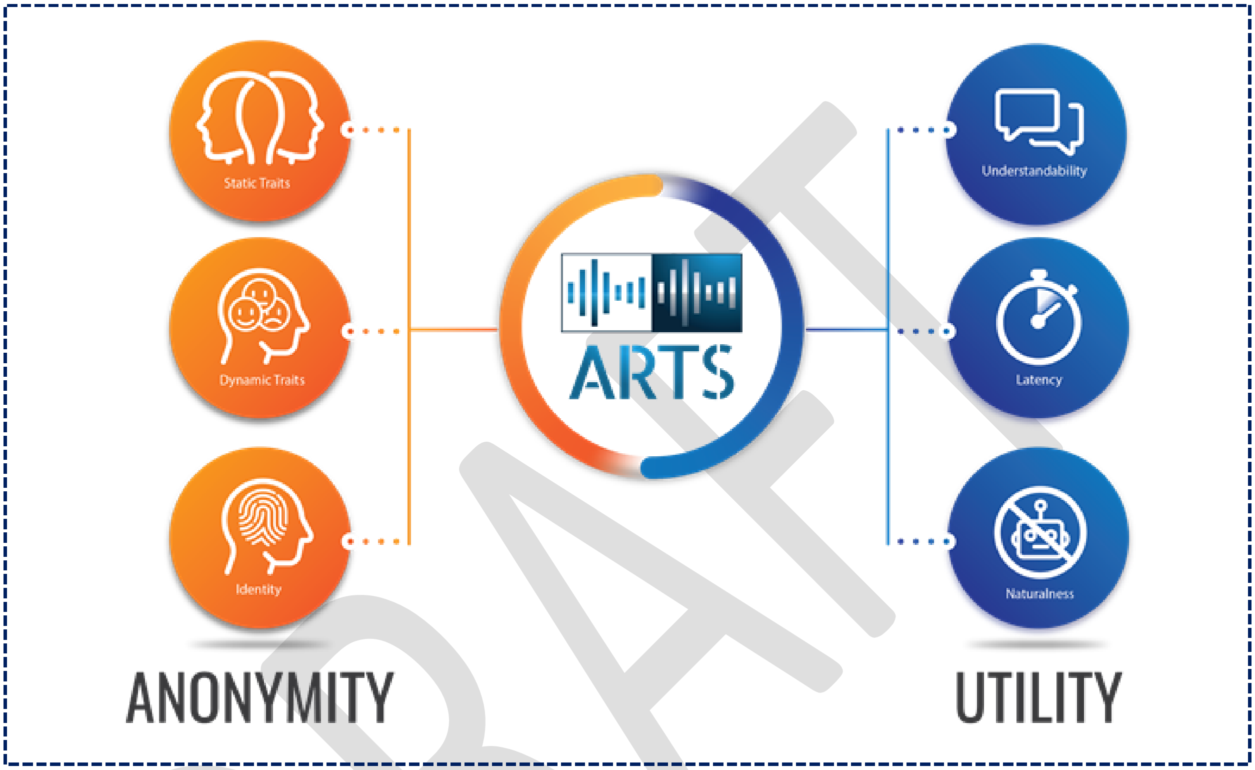

Figure 1 The ARTS program will simultaneously address six dimensions: three TAs for anonymity and competing forces of three areas of utility.

Performers are expected to conduct research against all three technical areas, producing robust software systems that can operate on standalone computers. Performer systems will be required to develop modules for each TA operating individually and a single solution that includes all three TA modules operating together to meet program goals. With the distinct and competitive objectives of the program, potential Performer teams are encouraged to pursue collaborative efforts and teaming opportunities. It is anticipated that teams will be multidisciplinary and may include expertise in one or more of the disciplines listed in **Section I.B, Team Expertise**.

Speech data will be provided by the Government for use in the ARTS program. This data may be modified by performers to enhance their capabilities. Collection of other data is permitted with Government approval. Speech data will initially focus on the English language, although other widely spoken languages such as Spanish will be used later in the program. Additional details on Program data can be found in **Section I.D, Program Data**.

The ARTS Test and Evaluation (T&E) teams will compile robust sets of diverse and relevant speech corpora to support research goals. In addition to speech collected from conversations of actual speakers, synthetic data may potentially be created to assist training. The T&E teams will be conducting several field data collections, studio data collections and data simulation exercises throughout the life of the program. While a portion of data will be made available to Performers for R&D, T&E will withhold sets of data to facilitate the testing and validating of performance through a series of challenges. The independent test and evaluation performed by T&E on Performer software deliverables will inform Government stakeholders on ARTS research progress but will also serve as valuable feedback to the Performers to improve their research approaches, algorithm training practices, and system development. The ARTS program will work closely with Government leaders in speech processing and speaker ID to continually refine and improve T&E methodologies. Additional information on T&E is described in **Section I.E, Test and Evaluation**.

Developed capabilities must be containerized and compatible with a government furnished Application Programming Interface (API) to facilitate assessment by independent test and evaluation (T&E) according to program metrics described in **Section I.F, Program Metrics**.

Performer systems will be evaluated on an interim basis to track progress and ensure compliance with system requirements. Over the course of the ARTS program, the evaluation challenges will become progressively more difficult. The ARTS program will adhere to a *fixed* training condition, where Performers can use a consistent set of training data provided by the Government. Performers will need to continually improve their systems aggressively to meet these challenges.

The ARTS program is envisioned as a 36-month effort, comprised of two (2) Phases, each lasting 18 months. Proposals shall include a solution for Phases 1 and 2, inclusive of all Technical Areas. Proposals that do not include a solution for both phases or do not address all Technical Areas will be considered non-compliant and will not be evaluated. A schedule of program waypoints, milestones, and deliverables are provided in **Section I.G**.

A. Technical Areas

Proposed solutions for the ARTS program should be a single complete system that includes three software-based modules, one for each TA. The complete system should be configured so the modules can be executed alone, as well as in combination with other modules at the same time. When the system is configured to execute a single module by itself, only the goals of that TA are required to be addressed. However, when the entire system is configured to execute two or three modules at the same time, the system must address the goals of the TAs for each module selected. When the system is operating in any mode, it must address all three utility requirements.

The Government will furnish all data to be used by Performers. All software solutions must adhere to a framework or API developed by T&E Teams, which will be provided to Performers at kickoff. The solutions must run on a stand-alone computer, called the ARTS Processing System (APS), with specifications (operating system, hardware configurations) to be provided at program kickoff. **Figure 2** provides an illustration of this concept.

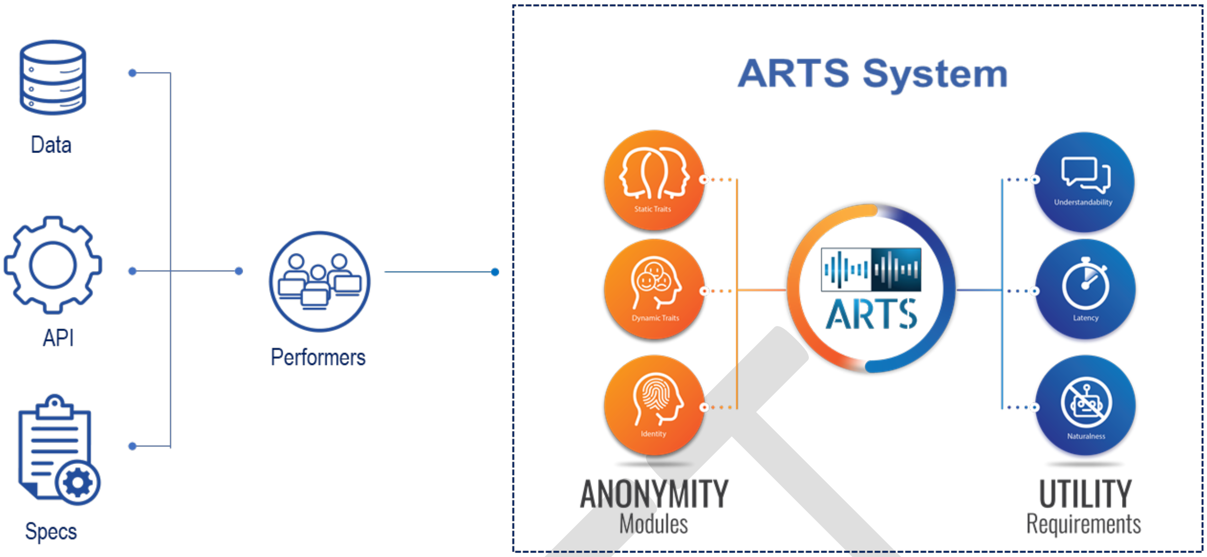

Figure 2 Given the data, API, and specifications of the APS, performers will develop a system that addresses the three TAs and all utility requirements.

T&E will process speech on the APS using sequestered data, transforming speech from an individual, Alice, into speech that sounds like it comes from a new, different person, Pseudo Alice. Pseudo Alice should not sound like a specific target individual. However, T&E must be able to recreate results for transformed speech from original speaker; Pseudo Alice should always sound like Pseudo Alice. See **Figure 3**.

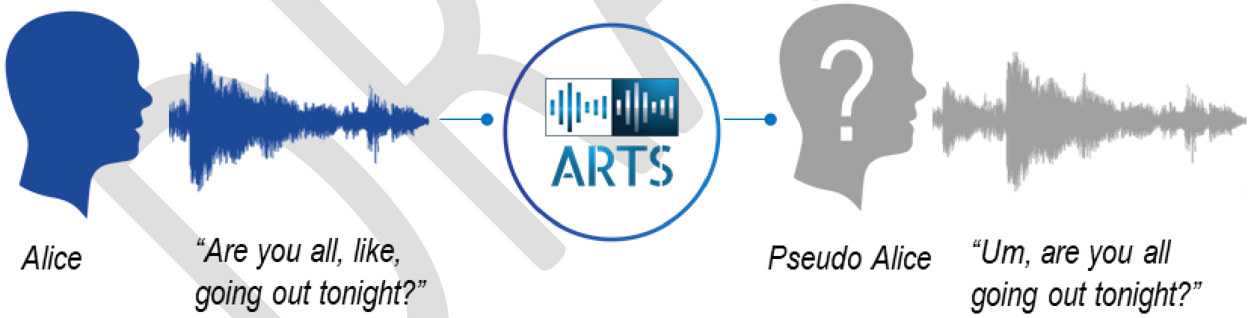

Figure 3 The ARTS system should transform speech from an individual into speech that sounds like it comes from a new, different person.

Figure 4 provides an illustration of a compliant ARTS system that addresses the three TAs, along with three utility requirements. Performer teams will be provided speech corpora, target profiles, a description of an SID approach, and system requirements for the hardware on which solutions will be executed.

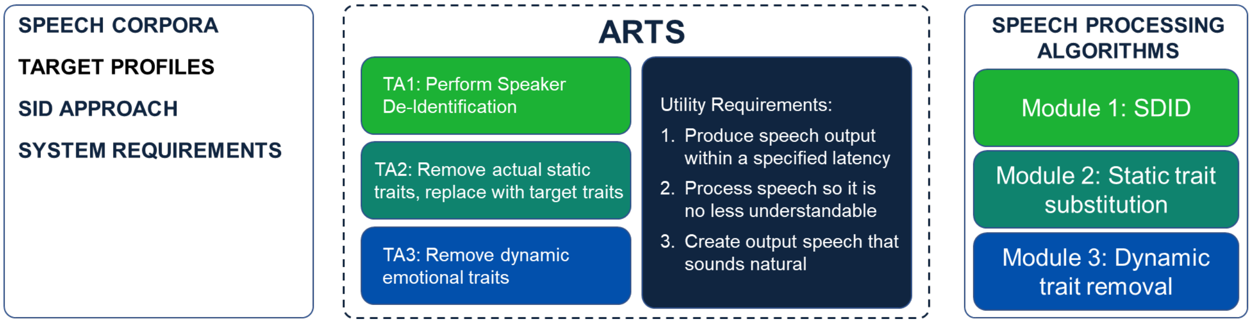
SPEECH CORPORA
TARGET PROFILES
SID APPROACH
SYSTEM REQUIREMENTS
ARTS
TA1: Perform Speaker De-Identification
TA2: Remove actual static traits, replace with target traits
TA3: Remove dynamic emotional traits
Utility Requirements:
1. Produce speech output within a specified latency
2. Process speech so it is no less understandable
3. Create output speech that sounds natural
SPEECH PROCESSING ALGORITHMS
Module 1: SDID
Module 2: Static trait substitution
Module 3: Dynamic trait removal

Figure 4 Illustration of an ARTS system that has components to address each TA, while adhering to three utility constraints.

The following subsections describe each TA in more detail, along with the innovations sought by the ARTS program.

A.1. Technical Area 1: Speaker De-Identification

The goal of TA1 is to develop novel methods to modify speech so that it will not be attributed to the original speaker or any other individual. The ARTS program is looking for innovations that include novel low-level representations of speech, fast speech synthesis models, and text-free speech synthesis.

Solutions for TA1 should have the following properties:

- The TA1 module shall take an input segment of speech from an original speaker, Alice, and output a new speech signal segment, called pseudo-Alice. The output speech segment should have the same linguistic content (lexical phrases) with the exception of speech disfluencies such as discourse markers, restarts, stutters, and other sounds that do not convey content. Linguistic content is defined as the meaning conveyed through vocabulary, syntax, morphology, and semantics, and speech disfluencies are disruptions that occur in the flow of talking.
- If T&E is performing an SID evaluation on a segment from pseudo-Alice using a system that has been trained on a set of original speakers, the segment shall not be attributed to any speakers enrolled in the SID system. This includes the scenario where the original speaker, Alice, is enrolled in the database.
- The TA1 module shall have the ability to transform an input segment in different ways, so that it is possible to obtain different output speech from the same speaker. The outputs should differ so that SID would indicate they were spoken by different individuals. The minimum number of different pseudo-speakers that can be obtained by a single original speaker is eight (8).
- The TA1 module shall produce consistent transformation(s). That is, if desired by the user, the modification of any speech segment from Alice can always appear to come from the same (non-existent) individual, pseudo-Alice.
- For speech from a different original speaker, Bob, which is transformed by the same module, the output should be different from any other pseudo-speaker.

In plain language, the SDID module should change a speaker’s voice in many ways, so that a user can choose to have pseudo-speech consistent with a specific, non-existent person. Also, no two different pseudo-speakers should sound like the same person.

The SID threat model assumed under the ARTS program is one of an *informed attacker*. This means that as part of the evaluation of TA1, T&E Teams performing SID attacks on processed speech will have full knowledge of the de-identification algorithms. The T&E Teams will use these algorithms to help build models of speech modified by the Performers' systems. The intent of this threat model is to have the strength of the de-identification method rely on the algorithm instead of hiding the fact that this processing is potentially taking place.

SID tests consist of *trials*, in which an utterance (test segment) is compared against enrollment data (target segments) in the SID system. The SID system processes trials independently to produce output log-likelihood ratio (LLR) scores. The test conditions for TA1 are as follows:

- The speech duration of the test segments will be uniformly sampled ranging approximately from 10 seconds to 60 seconds.
- Trials will be conducted under the scenario where unprocessed, original speakers are used for enrollment segment(s), to assess how output segments are attributed to real speakers.
- Trials will be conducted under the scenario where transformed, pseudo-speakers are used for enrollment segment(s), to assess whether the solutions can produce output that sounds like it comes from different speakers.
- Trials will be conducted under the scenario where consistent pseudo-speakers are used for enrollment segment(s), to assess whether the solutions can consistently change a speaker to always sound like the same (non-existent) person.
- Trails will include cross-gender comparisons.

A.2. Technical Area 2: Static Trait Replacement

The goal of TA2 is to research and develop speech processing methods that remove existing traits in speech signals related to dialect, gender, and age, and replace them with pre-selected traits. The ARTS program is looking for innovations in speech processing that may include phonetic (segmental and suprasegmental) and non-phonetic (linguistic, non-linguistic, and lexical choices) aspects.

For TA2, an individual speaker will have a known *profile* consisting of three long-term states: dialect, gender, and cohort (age groups). Languages and dialects will differ in Phase1 and Phase 2. The ARTS program will limit dialects to a small set of well-studied groups, acknowledging that regional dialects and accents are far too numerous to include in the scope of this program. One common type of dialect will be called *accented*, in which the speaker is fluent in two languages. For the purpose of this program, an accented language will be from a speaker who is fluent in that language as a second language; the speaker's first language, L1, is different. For instance, a person with an accented Spanish could be someone who has a first language of English and second language Spanish (L1 English/L2 Spanish).

Gender and age will be self-reported by the individual speakers. Cohorts will consist of three sets of non-adjacent ranges: Young (18-24), Adult (35-44), and Senior (55-64).

Solutions for TA2 should have the following properties:

- The TA2 module shall take an input segment of speech from an original speaker, Alice, who has a specific profile (Dialect, Gender, Cohort), along with a randomly chosen profile, and produce new speech with the same linguistic content (excluding speech disfluencies) in a manner so that a panel of human evaluators would believe the speaker has the randomly chosen profile.

- The profile of the TA2 output speech shall match the randomly selected profile regardless of whether the original speaker has a static trait that already matches a target trait. That is, if an original speaker speaks with an SAE dialect, and the SAE dialect is randomly chosen as part of the output profile, the TA2 module should produce speech that has an SAE dialect.
- Solutions that add or remove non-linguistic content that may be inherent to a specific static trait are permitted. For instance, certain discourse markers, slang words, and other utterances or non-vocal sounds that are present/absent may be modified, removed, or added to assist with the transformation of the profile. In addition to these approaches, the ARTS program seeks additional novel methods to modify static traits.

Results from the TA2 module will be assessed by a panel of human evaluators. Evaluators will be asked to assess traits for original speakers in order to obtain a baseline of the panel’s ability to differentiate between dialects, ages, and genders.

A.3. Technical Area 3: Dynamic Trait Removal

The goal of TA3 is to modify speech signals so that extracted features can’t be used to differentiate short term states such as emotion and psychological state. A common method to assess these dynamic traits is to extract high-dimensional features and use deep neural networks to classify the speech. These features can vary but are usually be found implemented in speech analysis software packages like Praat and the open-source Speech and Music Interpretation by Large-space Extraction (openSMILE) toolkit. Innovations to address this area may include source modeling of emotional speech, generative adversarial networks, and universal adversarial perturbations.

The ARTS program will categorize emotions into two sets: positive traits, which consist of happy and excited, and negative traits, which include angry and sad. Under this construction, the goal of TA3 is to defeat binary classifiers which are designed to differentiate between these classes.

Solutions for TA3 should have the following properties:

- The TA3 module shall take an input segment of speech from an original speaker, Alice, and produce new speech with the same linguistic content (excluding speech disfluencies) in a manner so that positive and negative traits present in the speech are removed.

T&E will assess the output from the TA3 module by evaluating a set of low level descriptor (LLD) features extracted from modified speech signals. These LLD features will be evaluated by three different deep convolutional neural networks (CNN) previously trained on unprocessed emotional data sets.

B. Utility Constraints

In addition to the three modules described above, a complete solution must also perform under three utility constraints: latency, understandability, and naturalness. Systems are required to operate under each of these constraints for each single TA module and the final solution containing all three TA modules.

B.1. Utility Constraint 1: Latency

Latency is a key component to the ARTS program, and goals will be for delays on the order of common communication channels such as satellite phones or voice over IP (VoIP). Latency is perceived as the time difference between when a specific word is spoken as input and when the output of that word is produced. The ARTS program allows for the synthesis of sounds that might not be present in original speech. Spontaneous speech often contains linguistic, non-linguistic, and lexical choices that don't change the meaning of what is being said. This includes fillers ("uh", "um", "like", "you know", "I mean"), restarts (repetitions, corrections, restarts), pauses (silent, breathed, filled), and keywords/slang. Much like the human brain uses these disfluencies to buy time as we formulate our thoughts and words, ARTS systems may introduce, remove or modify to help reduce the impact of processing time. An example is shown in **Figure 5**, where the synthesis of the word "um", along with a breathed pause, allows for more processing time, since latency is determined by when the utterance "um" is produced.

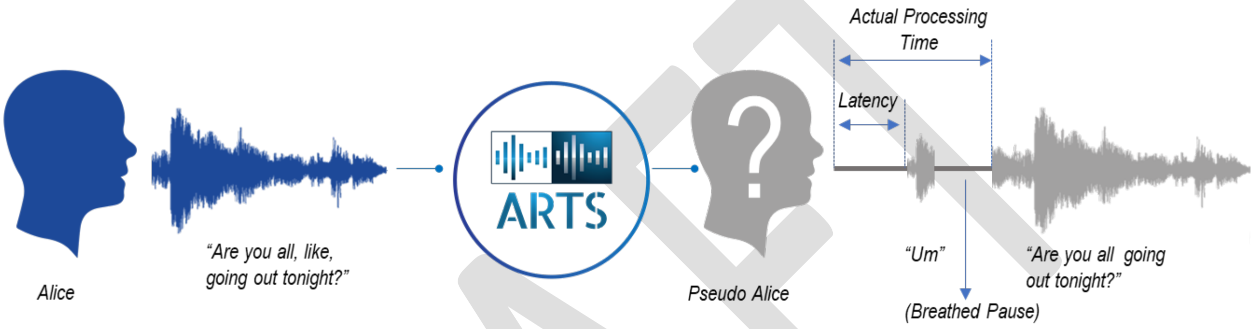
Actual Processing Time
Latency
ARTS
Alice
"Are you all, like, going out tonight?"
Pseudo Alice
"Um"
(Breathed Pause)
"Are you all going out tonight?"

Figure 5 The utility constraint for latency is determined by the detection of the first utterance, rather than the difference in time between exact words.

With the understanding that the output speech might not be a word-for-word replication, latency will be measured as the duration between speech input and speech activity detection (SAD) on the output. Latency will be measured by T&E Teams on the APS.

B.2. Utility Constraint 2: Understandability

Many approaches to anonymizing speech will reduce the understandability. T&E will evaluate output speech to ensure that the understandability of transformed speech is within reason. Since program data will include annotated transcripts, this constraint will be measured by automatic speech recognition (ASR) system such as Kaldi Speech-to-Text. With the ASR evaluation system, discourse markers will be ignored, and the word error rate will be measured by comparing ground truth transcripts with ASR output and computing the number of wrong words (substitutions), missing words (deletions), and addition of words (insertions). Because ASR systems are not perfect and quality of data sets can vary, T&E Teams will compute the baseline word error rates (WER) on original, unprocessed data.

B.3. Utility Constraint 3: Naturalness

The ARTS program aims to produce solutions that can transform speech and continue to sound natural. Listeners often find many speech synthesis tools create speech that sounds artificial, and especially with SDID systems, the output can sound robotic. T&E will assess naturalness with a panel of ~200 unique listeners, balanced in gender, with individuals assessing a small set of output speech. This approach has been successfully used in events such as the Voice Privacy Challenge.

For this evaluation, listeners will be informed that they will hear test segments that may be of high quality, but some may sound artificial due to deterioration caused by computer processing. Participants will then evaluate segments on a five-point scale (Bad, Poor, Fair, Good, and Excellent). The mean opinion score (MOS) will be computed for both original, unprocessed data, as well as output from the solution systems.

C. Program Phases

All three Technical Areas and all three utility constraints must be addressed by performers in both Phase 1 and Phase 2. The key difference between the two phases will be the languages involved. Phase 1 will concentrate on speech segments in English. Phase 2 will initially focus on variants of Spanish, before including other languages that are spoken by a large portion of the global population.

C.1. Phase 1: English

The goal of Phase 1 is to develop anonymization systems for speech in the English language. Dialects will include Standard American English (SAE), Standard Southern British English (SSBE), and accented English where the speaker's L1 is Spanish and L2 is English.

Phase 1 shall have a duration of 18 months, with interim deliveries of interim containerized software solutions at months 5 and 11, and final deliverables at month 16. Interim deliverables will be evaluated against subsets of the Technical Areas to measure progress and provide performers feedback, and Final Phase 1 deliverables will be evaluated against all three Technical Areas.

Table 1 provides a list of software deliverables for Phase 1, along with the challenge areas that must be addressed and language/dialect information.

Table 1 Software deliverables for Phase 1, along with the challenge areas that will be evaluated and the languages/dialects present in the data.

Month	Deliverable Type	Challenge Areas	Languages/Dialects
5	Interim	TA1, Latency, Understandability	Regions US Dialects
11	Interim	TA1, TA3, Latency, Understandability	SAE, SSB, L1 Spanish
16	Phase 1 Final	All TAs and utility constraints	SAE, SSB, L1 Spanish/L2 English

C.2. Phase 2: Non-English

The goal of Phase 2 is to develop anonymization systems for speech in non-English languages. The first 12 months of Phase 2 will focus on the Spanish language and will likely include the very distinct and well-studied dialects such as North Central Peninsular Spanish (NCPS), Eastern Andalusian Spanish (EAS), and accented Spanish where the speaker's L1 is English and L2 is Spanish. Stretch goals for the end of Phase 2 will include a variety of other languages that are commonly spoken throughout the world. These languages include Korean, Russian, Modern Standard Arabic, and Mandarin Chinese (MC), where data for each language will consist of native speakers and L1/L2 speakers of that language. For example, with Korean the data will include native speakers of Korean and speakers with L1 English/L2 Korean.

Phase 2 shall have a duration of 18 months. Phase 2 consists of advanced algorithm development to address moderate to high challenges. Throughout the phase, quarterly deliveries of containerized software are anticipated and will be tested and evaluated through challenge activities.

Table 2 provides a list of software deliverables for Phase 1, along with the challenge areas that must be addressed and language/dialect information.

Table 2 Software deliverables for Phase 2, along with the challenge areas that will be evaluated and the languages/dialects present in the data.

Month	Deliverable Type	Challenge Areas	Languages/Dialects
23	Interim	TA1, Latency, Understandability	NCPS, EAS
29	Interim	TA1, TA3, Latency, Understandability	NCPS, EAS, L1 English/L2 Spanish
34	Phase 2 Final	All TAs and utility constraints	Spanish, Russian, MSA, MC (All with native and L1 English speakers)

IARPA will continue to use ARTS API developed by the ARTS T&E Team in Phase 1 and 2 for all Technical Areas. The T&E teams will provide datasets for the purpose of constructing the evaluation challenges. Following each challenge, performance analysis results and challenge data will be provided to performers for review and methodology improvement. The sharing of this data with Performers after the challenges is to facilitate communication and internal Performer error analyses. More details on the datasets are available in **Section I.D, Program Data** and more details on the API are available in **Section I.G.2.1, Program Application Programming Interface**.

Section II Team Expertise

Collaborative efforts and teaming among Offerors are highly encouraged. It is anticipated that teams will be multidisciplinary and may include expertise in one or more of the disciplines listed below. This list is included only to provide guidance for Offerors; satisfying all the areas of technical expertise below is not a requirement for selection and unconventional or innovative team expertise may be needed based on the proposed research. Proposals should include a description and the mix of skills and staffing that the Offeror determines will be necessary to carry out the proposed research and achieve metrics.

- Speech and signal processing
- Speech synthesis
- Speech and speaker recognition
- Forensic speech science
- Natural language processing
- Linguistics
- Phonetics

- Acoustics
- Audio engineering
- Speech pathology
- Modeling and simulation
- Machine learning, deep learning, or hierarchical modeling
- Artificial intelligence
- Systems integration
- Systems engineering
- Software engineering
- Data reduction and analysis

A. Program Scope and Limitations

Proposals shall explicitly address all of the following:

- **Underlying Theory:** Proposed strategies to meet program-specified metrics must have firm theoretical bases that are described with enough detail that reviewers will be able to assess the viability of the approaches. Proposals shall properly describe and reference previous work upon which their approach is founded.
- **Research & Development approach:** Proposals shall describe the technical approach to meeting program metrics.
- **Technical Risks:** Proposals shall identify technical risks and proposed mitigation strategies for each.
- **Software Development:** Proposals shall describe the approach to software architecture and integration.

The following areas of research and approaches are **out of scope** for the ARTS program:

- Research that does not have strong theoretical and experimental foundations.
- Development of hardware solutions or methods that require special hardware.
- Development of voice conversion or voice cloning systems, in which the aim is to convert speech to sound like a specific target individual.
- Approaches that rely on external (non-Government provided) data sources.
- Research that utilizes proprietary data.
- Methods that require a human-in-the-loop as part of the integrated end-to-end system.
- Approaches that consist merely of integrating currently existing software.

Delivered software will be evaluated by an independent T&E team on sequestered and shared evaluation datasets. Performers will build prototype algorithms and subcomponent modules that will be run and evaluated by the T&E Team. Testing protocols do not allow for expert operators, human-in-the-loop operation, or any operations not deemed "turnkey." However, systems or algorithms that have been trained using human-in-the-loop methods may be submitted, provided they run autonomously.

Section III Program Data

For ARTS to facilitate innovative R&D and achieve program metrics, diverse program data in sufficient quantities are needed for development and statistically reliable evaluations. As a result, the program will include robust and explicit data provision and collection by the ARTS T&E Team.

Performer teams will not be permitted to record or collect speech data from human subjects. Teams may synthesize artificial speech data, where there is no actual speaker, and they may acquire licenses to use publically available datasets. However, any external data sets used in research and development must be made available to the Government upon system delivery.

The program will leverage existing data sets as well as collect and potentially simulate evaluation data from approved sites. The data will consist of speech segments in durations 10s-60s, collected with a range of conditions, sensors, and environments. Speech segments will reflect a range of languages, dialects, genders, age groups, and short term emotional states. Evaluation data will be explicitly excluded from any algorithm training approaches and be withheld from Performers until the completion of evaluation events (challenges).

For a corpus to be of value for speaker recognition, it must have the following properties:

1) It must contain multiple speakers;
2) The distinction between the speakers must have been verified through some reliable method independent of the application of speaker recognition technology; and,
3) The corpus must contain at least two recordings of some target speakers collected at different times or under different conditions.

A primary source for data meeting these requirements comes from corpora hosted by the Linguistic Data Consortium (LDC). A recent Government evaluation of data sets suitable for the purpose of forensic speech analysis revealed 146 distinct data sets, including over 20 evaluation sets used by the National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE). The review of these holdings also feature a wide range of speech conditions, to include:

- Conversational
- Different levels of vocal effort (soft/whisper, shout, Lombard)
- Stress
- Physiology (illness, intoxication)
- Emotion
- Language/Dialect
- Non-Native Speech
- Sex/Gender

The T&E Teams will evaluate this collection of speech corpora to create sets for development and evaluation.

A. Development Data

A limited amount of sample data will be provided in advance of evaluation events to performers. Sample data to serve as an example of formatting and facilitate development of ingest tools will be provided, but this data will not be sufficient in volume to facilitate algorithm training and methodology development. The Government will provide sample data upon Program kick-off.

The Government will make additional development data sets available in both Phases of the program. Data relevant for Phase 1 evaluations will be provided at Program months 3 and 12, and data relevant for Phase 2 will be provided at Program months 21 and 30.

B. Evaluation Data

ARTS will utilize distinct test data to evaluate the performance of Performer subcomponents, modules, and systems against program goals, objectives, and metrics. Intended uses of the evaluation datasets include both use by the T&E Team for independent evaluation of program deliverables against target metrics during quarterly challenges and use by Performers after these challenges to refine and improve their algorithms. The evaluation datasets will be provided to Performers to enable internal T&E and exploratory error analysis by Performers and to improve the consistency and communication between Performers and T&E following each challenge. No unreleased evaluation data will be permitted in any aspect of algorithm training or functionality until after it has been used in an evaluation. Additional sequestered or external datasets may be used to supplement performance evaluations at the discretion of the ARTS PM.

Section IV Test and Evaluation (T&E)

T&E will be conducted by an independent team of Government and contractor staff carrying out evaluation and analyses of Performer research Deliverables using program test datasets and protocols. In addition to independent T&E, the program will regularly gauge interim progress of Performer research activities towards ARTS objectives and target metrics using T&E results measured and reported by the Performer teams themselves. The ARTS evaluation data and test protocols will be the primary mechanism by which the T&E Team carries out their evaluations.

The ARTS program will pursue rigorous and comprehensive T&E to ensure that research outcomes are well characterized, deliverables are aligned with program objectives, and that algorithm performance is measured across the full range of architectural, sensor, and environmental conditions. Such T&E activities will not only inform IARPA and Government stakeholders on ARTS research progress but will also serve as invaluable feedback to the Performers to improve their research approaches, algorithm training practices, and system development. The ARTS program will work closely with Government leaders in speech processing, speaker recognition, and applied linguistics to continually refine and improve T&E methodologies. Evaluations will occur quarterly through challenge events that will exercise performer solutions across technical challenges described in section 1.A., independently and in combination.

The Government will provide Performers with an API and container requirements to integrate in a program test harness with relevant scripts to run program test protocols on program datasets. T&E will develop an ARTS Processing System (APS) to execute Performer Deliverables on evaluation datasets. This processed data will be distributed amongst the T&E partners for evaluation against the different TAs and utility constraints.

Performers will have specific Deliverable Milestones when all subcomponent and system algorithms and software will be delivered to IARPA and its designated T&E Team. The T&E Team will then conduct evaluations at the direction of the ARTS PM and with the objective of characterizing the quality, functionality, and performance of the ARTS Deliverables. In addition to quantitative measurements, T&E will be carried out to establish a thorough understanding of the progress, status, and limitations of the Performer's research.

T&E results and feedback will be provided to Performers at regular intervals to keep them abreast of current independent performance measurements and to inform and improve their R&D approaches and methods. T&E results from all Performers will be shared with all teams to establish an understanding of the current state and progress of ARTS research; T&E results will also be shared with USG external stakeholders, including their contractors, for Government purposes. For example, a PI Review Meeting will be held at the phase mid-point and at the end of each phase to share research ideas, progress, and results across the ARTS program (reference 1.H.1. Workshops).

IARPA may conduct other supplemental evaluations or measurements at its sole discretion to evaluate the Performers' research and Deliverables.

Section V Program Metrics

Achievement of metrics is a performance indicator under IARPA research contracts. IARPA has defined the ARTS program metrics to evaluate the effectiveness of the proposed solutions in achieving the stated program goal and objectives, and to determine whether satisfactory progress is being made. The metrics described in this BAA are shared with the intent to scope the effort, while affording maximum flexibility, creativity, and innovation to Offerors proposing solutions to the stated problem. Program metrics may be refined by the Government during the various phases of the ARTS program; if metrics change, revised metrics will be communicated in a timely manner to Performers.

At its core, ARTS is an R&D program focused on speech anonymization. Performance metrics are focused on the ability to defeat speaker characterization threats, balanced with metrics to ensure that solutions can meet the utility goals (practical needs and logistical challenges) of use cases. Metrics were chosen with the following considerations:

1. What is technically achievable but challenging based on current state-of-the-art in the speech processing R&D communities;
2. What is statistically measurable based on the planned program evaluation data; and
3. What is useful to mission partners based on USG stakeholder needs and use cases?

Although Performers must deliver interim solutions at different waypoints in the program, the metrics are for use only with final Phase deliverables. For interim deliverables, the metrics may serve as a measure of progress. The ARTS program may use these measurements to highlight improvements, identify risks, and reassess relevance/realism of metrics.

A. Metrics for TA1

The metric associated with TA1, SDID, is based on LLR scores produced by SID systems. For a set of scores, the equal error rate (EER) is where the false alarm rate and the false miss rate are equal. One situation is where a test segment is from a person enrolled in the SID system. For this case, we say the target segment is known to be present in the test segment. Under this circumstance, we expect very low EER. However, another situation is where the test segment is from a person not enrolled in the SID system, in which case we say the target is not present in the test segment. For this, we expect an EER to on average be around 50%, reflecting random guessing.

T&E will perform three types of SID attacks against output data. As noted above, the SID attack for the final Phase deliverables will be one of an informed attacker, where T&E will use the Performer systems to enroll pseudo-speakers into the SID database. The three attacks performed in the final Phase deliverables are:

- SDID test segments against original target speakers: desired result is to have inconclusive SID results. The EER should be above 45%. The target metric for this test is 0.5± 0.05.
- SDID test segments against non-matching target pseudo-speakers: desired result is to have inconclusive SID results. The EER should be above 45%. The target metric for this test is 0.5± 0.05.
- SDID test segments against matching target pseudo-speakers: desired result is to show target segments are present in test segments. The EER should be below 5%.

B. Metrics for TA2

The metric associated with TA2, static trait replacement, is based on the accuracy of a panel of human evaluators. The listeners will assess samples and predict the category for dialect, gender, and cohort. A correct prediction is one where the assessment matches the randomly selected trait. For each category within a profile, accuracy will be computed as:

$$\frac{\#\ of\ correct\ predictions}{\#\ of\ total\ predictions}$$

T&E Teams will also produce accuracy measures for original (unprocessed) data. The target metric for each category at the end of each Phase will be the panel's average accuracy for original data.

C. Metrics for TA3

The T&E Teams will use three deep CNNs to evaluate data under TA3. Using the openSMILE 3.0 toolkit, a set of over 4k LLD features will be extracted from unprocessed emotional datasets such as the IEMOCAP data, establishing two classes. One class will contain speech with positive traits, (happy and excited) and the other class will have speech with negative traits (angry and sad). T&E will use supervised training approaches to train CNN-4, ResNet-50, and VGG-16 networks.

T&E will assess the output from the TA3 module by evaluating the same set of LLD features extracted from transformed speech from module 3. The performance will be measured by unweighted average recall (UAR), with a target metric of .48.

D. Metrics for Latency

In order to obtain independent, consistent, repeatable test and evaluation results, T&E will use the APS to execute all interim and final Phase deliverables. The motivation for this is to ensure that latency can be processed reliably as speech is processed. Once test segments are transformed, the output files can be evaluations against the TAs and other utility constraints on other T&E test systems. Latency is the only metric which must be processed on the APS.

The APS will take an input test segment in the form of an uncompressed Waveform Audio File (WAV file) in the linear pulse-code modulation (LPCM) format. The output of a Performer's system shall produce a WAV file with the modified speech. Software will mark the time in which the output speech triggers SAD, and the difference between the time of input and SAD will be considered the duration.

The event that triggers SAD does not necessarily have to be the same event that would trigger SAD on an input signal. However, Performers should consider impact that introducing synthesized artifacts could have on metrics for other aspects of this program.

Latency will be measured in millisecond (ms), and target metrics will be on the order of common communication channels, such as satellite or VoIP phones. The target metric for final Phase 1 deliverables is 350ms, and the target metric for final Phase 2 deliverables is 150ms.

E. Metrics for Understandability

T&E will evaluate output speech to ensure that the understandability of transformed speech is within reason. This will be accomplished by using Kaldi ASR on processed speech and comparing with ground truth transcripts of the segments. With this approach, discourse markers will be ignored, and the word error rate will be measured by comparing ground truth transcripts with ASR output and computing the number of wrong words (substitutions), missing words (deletions), and addition of words (insertions). Thus, ignoring discourse markers, we have the following:

- S is the number of substitutions (wrong word)
- D is the number of deletions (missing word)
- I is the number of insertions (added word)
- N is the number of words
- The word error rate is computed $WER = \frac{S+D+I}{N}$

Because ASR systems are not perfect and quality of data sets can vary, T&E Teams will compute the baseline WER on original, unprocessed data. In addition, T&E will compute WER for two baseline anonymization systems. The target metric in Phase 1 for understandability of output speech is the mean of WER for original data and WER for baseline methods of anonymization. Intuitively, this means the performance loss for understandability is cut in half from existing anonymization systems. The target metric for Phase 2 builds on the figures for Phase 1, using the mean of original WER and Phase 1 WER, which can be viewed as cutting the performance loss in half again.

F. Metrics for Naturalness

The ARTS program aims to produce solutions that can transform speech and continue to sound natural. T&E will assess naturalness with a panel of ~200 unique listeners, balanced in gender, with individuals assessing a small set of output speech. For this evaluation, listeners will be informed that they will hear test segments that may be of high quality, but some may sound artificial due to deterioration caused by computer processing. Participants will then evaluate segments on a Likert scale (1-Bad, 2-Poor, 3-Fair, 4-Good, and 5-Excellent). This range is commonly used in human evaluations of speech and is intended to capture a range of feelings for a given item. The mean opinion score (MOS), which is the arithmetic mean of scores, will be computed for output from the solution systems, with a target metric of 3.5 for Phase 1, which reflects speech that is considered to have a natural quality of Fair to Good. The target metric for Phase 2 is 4.5, reflecting an assessment of Good to Excellent. The MOS will also be computed for original, unprocessed data, but will not be used in the target metrics. The MOS scores for original data will be used for informational uses only, such as a point of reference if targets should need to be adjusted after Phase 1.

G. Summary of Metrics

Table 2, below, summarizes metrics that will be used to assess performance in Technical Area 1 in Phases 1 and 2.

Evaluation	Metric	Phase 1 (English) Target	Phase 2 (Multi-Lingual) Target
TA1 – original target or nonmatching pseudo target not present in test	Average EER	0.5± 0.05	0.5± 0.05
TA1 – matching pseudo target present in test	Average EER	0.05	0.05
TA2 – Dialect	Average Accuracy	Original Accuracy	Original Accuracy
TA2 – Gender	Average Accuracy	Original Accuracy	Original Accuracy
TA2 – Cohort	Average Accuracy	Original Accuracy	Original Accuracy
TA3 – All three CNNs	UAR	0.48	0.48
Latency	Mean duration from input to SAD	350ms	150ms
Understandability	Mean WER	Mean of original and baseline WER	Mean of original and Phase 1 WER
Naturalness	MOS	3.5	4.5

Section VI Program Waypoints, Milestones, and Deliverables

Waypoints, Milestones, and Deliverables are established from the program's onset to ensure alignment with ARTS objectives, organize research activities in a logical and reportable manner, and facilitate consistent and efficient communication among all stakeholders – IARPA, ARTS T&E, USG Stakeholders, and Research Performers.

A. Program Milestone, Waypoint, and Deliverables Timeline

An overview of the schedule for ARTS Milestones, Waypoints, and Deliverables is presented in **Table 3**. Additional details for each of these activities are shown in **Table 4**.

Table 3 ARTS program schedule for Phase 1 and Phase 2

Phase 1 (18 months)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Kickoff Meeting	X																	
Sample Dataset Delivery	X																	
Development Dataset Delivery			X									X						
Site Visits				X						X					X			
Software Delivery					X						X					X		
T&E Software Evaluation and Reporting						X						X					X	
Demos						X						X					X	
PI Meetings									X								X	
Monthly Status Reports	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Phase 1 Final Report																		X
Phase 2 (18 Months)	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Kickoff Meeting	X																	
Sample Dataset Delivery	X																	
Development Dataset Delivery			X									X						
Site Visits				X						X					X			
Software Delivery					X						X					X		
T&E Software Evaluation and Reporting						X						X					X	
Demos						X						X					X	
PI Meetings									X								X	
Monthly Status Reports	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Phase 2 Final Report																		X

Table 4: ARTS Milestone, Waypoint, and Deliverable Schedule

Phase	Month	Event	Description	Comment	Deliverable
1-2	All	Waypoint	Monthly Status Report, Technical and Financial	Due on 15th of each month	MSR
1-2	All	Waypoint	Progress and Status Meeting	Biweekly Teleconference with ARTS PM	N/A
1	1	Waypoint	Kickoff Meeting	DC metro area	N/A
1	1	Waypoint	Sample Data	Provided as GFI	NA
1	3	Waypoint	Phase 1 Development Data Set 1	Provided as GFI	N/A
1	4	Waypoint	Site Visit	At performer site	N/A
1	5	Deliverable	Interim Solution Delivery	Containerized solution delivery for interim challenge	Software Container
1	9	Waypoint	PI Review Meeting	DC metro area	N/A
1	10	Waypoint	Site Visit	At performer site	N/A
1	11	Deliverable	Interim Solution Delivery	Containerized solution delivery for interim challenge	Software Container
1	12	Waypoint	Phase 1 Development Data Set 2	Provided as GFI	N/A
1	15	Waypoint	Site Visit	At performer site	N/A
1	16	Deliverable	Phase 1 Solution Delivery	Containerized solution delivery for Phase 1 challenge	Software Container
1	17	Waypoint	End of Phase 1 PI meeting and Demo	Washington, DC metropolitan area	N/A
1	18	Deliverable	Data Evaluation Assessment	Evaluation of data used on Phase 1 to improve data designed for Phase 2	Report
1	18	Deliverable	Phase 2 Risk Reduction Report	Report describing activities to reduce risks in Phase 2	Report
1	18	Deliverable	Phase 1 Final Report	Any updated software or data is also due	Report

Phase	Month	Event	Description	Comment	Deliverable
2	19	Waypoint	Kickoff Meeting	DC metro area	N/A
2	19	Waypoint	Sample Data	Provided as GFI	NA
2	21	Waypoint	Phase 2 Development Data Set 1	Provided as GFI	N/A
2	22	Waypoint	Site Visit	At performer site	N/A
2	23	Deliverable	Interim Solution Delivery	Containerized solution delivery for interim challenge	Software Container
2	27	Waypoint	PI Review Meeting	DC metro area	N/A
2	28	Waypoint	Site Visit	At performer site	N/A
2	29	Deliverable	Interim Solution Delivery	Containerized solution delivery for interim challenge	Software Container
2	30	Waypoint	Phase 2 Development Data Set 2	Provided as GFI	N/A
2	33	Waypoint	Site Visit	At performer site	N/A
2	34	Deliverable	Phase 2 Solution Delivery	Containerized solution delivery for Phase 2 challenge	Software Container
2	35	Waypoint	End of Phase 2 PI meeting and Demo	Washington, DC metropolitan area	N/A
2	36	Deliverable	Lessons Learned Report	Report describing programmatic and technical lessons learned	Report
2	36	Deliverable	Phase 2 Final Report	Any updated software or data is also due	Report

B. Software Deliverable Formatting

Performers will be required to provide algorithm and software deliverables in a manner that conforms to a standardized industrial method or methods that will be provided at program Kickoff. To facilitate planning, Offerors may assume that the standardized configuration will require the use of software containerization technology (e.g., Docker and a REST API). This means that the entirety of a Performer's system, including pre- and post-processing, must be included within the delivered software container. For models that require training, the expectation is for the initial model training to occur on Performer systems, with the ability for the T&E Team to re-train and test the model with the same and/or other data.

Each team is required to include among their Key Personnel a Lead System Integrator (LSI) who shall be responsible for preparing software Deliverable subcomponents, modules, and systems, performing quality control of Deliverable, and integrating key components into the primary ARTS system(s). The LSI will also oversee communication and coordination across a Performer's research teams including subcontractors, if applicable, to ensure research products are functional, integrated and following software coding best practices (e.g., inline comments, documentation). Additional team members and roles are dependent on the proposed research, as such, there is no predetermined or required skill mix.

B.1. Program API

The ARTS program will be utilizing a standardized API for all software deliverables and evaluations. The first version of the ARTS API will be provided to Performers at the Phase 1 Kick-off Meeting and updated periodically thereafter. The API will define function calls, data structures, and gallery creation and management for operating and evaluating ARTS software in a standardized manner.

Section VII Meeting and Travel Requirements

Performers are expected to assume responsibility for administration of their projects and to comply with contractual and program requirements for reporting, attendance at program workshops, and availability for site visits. The following paragraphs describe typical expectations for meetings and travel for IARPA programs as well as the contemplated frequency and locations of such meetings. In addition to ensuring that all necessary details of developed software, algorithm, and operational instructions are clear and complete, each Performer will be required to be available for questions and troubleshooting from the T&E Team in weekly and/or bi-weekly status meetings.

A. Kickoff Meetings and PI Workshops

All Performer teams are expected to attend workshops, to include Key Personnel from prime and subcontractor organizations.

The ARTS program intends to hold a program Kick-off Meeting workshop in the first month of the program and first month of each subsequent program phase. In addition, the program will hold a PI Review Meeting at the end of each phase and at the phase midpoint. Kick-off Meetings and PI Review Meetings may be combined for logistical convenience. The dates and locations of these meetings are to be specified at a later date by the Government, but for planning purposes, Offerors should use the approximate schedule listed in Table 4. Both types of meetings will likely be held in the Washington, D.C. metropolitan area, but IARPA may opt to co-locate the meeting with a relevant external conference or workshop to increase synergy with stakeholders.

Kick-off Meetings will typically be two days in duration and will focus on plans for the coming Phase, Performer planned research, and internal program discussions. PI Review Meetings will typically be two days in duration and will have a greater focus on communicating program progress and plans to USG stakeholders. These meetings will include additional time allocated to presentation and discussion of research accomplishments and interactive demonstrations for Government stakeholders.

In both cases, the workshops will focus on technical aspects of the program and on facilitating open technical exchanges, interaction, and sharing among the various program participants. Program participants will be expected to present the technical status and progress of their projects to other participants and invited guests. Individual sessions for each Performer team with the ARTS PM and T&E Team may be scheduled to coincide with these workshops. Non-proprietary information will be shared by Performers in the open meeting sessions; proprietary information sharing shall occur during individual breakout sessions with the ARTS PM and T&E.

B. Site Visits

Site visits by the Government Team will generally take place semi-annually during the life of the program. These visits will occur at the Performer's facility. Reports on technical progress, details of successes and issues, contributions to the program goals, and technology demonstrations will be expected at such site visits. IARPA reserves the right to conduct additional site visits on an as-needed basis or virtually if desired.

Section VIII Anticipated Period of Performance

Anticipated PoP: 36 Months as follows:

Phase 1: April 1, 2024 - October 31, 2025

Phase 2: November 1, 2025 – March 31, 2027

Note: Proposals shall include a solution for Phases 1 and 2, inclusive of all Technical Areas.

Section IX Place of Performance

Performance will be conducted at the Performers' sites.