Crew Role-Players Enabled by Automated Technology Enhancements

Beth F. Wheeler Atkinson
Naval Air Warfare Center Training Systems Division
Orlando, FL
beth.atkinson@navy.mil

John P. Killilea
Stracon Services Group
Orlando, FL
john.killilea.ctr@navy.mil

Brian Stensrud, Bob Marinier, Paul Schermerhorn, Chad Dettmering, Sohaib Saadat
Soar Technology, Inc.
Orlando, FL
stensrud@soartech.com, bob.marinier@soartech.com, paul.schermerhorn@soartech.com, chad.dettmering@soartech.com, sohaib.saadat@soartech.com

Emily C. Anania
Embry Riddle Aeronautical University
Daytona Beach, FL
ananiae1@my.erau.edu

ABSTRACT

U.S. Naval aviation, similar to units in its sister services, uses the *family of simulators* approach to training that enables trainees to build on skills progressively throughout the training pipeline. The progression begins with system skills (e.g., *buttonology*), continues to individual tasks (e.g., understanding radar data), and concludes with aircrew coordination for tactical proficiency (e.g., prosecuting an anti-submarine warfare mission). However, this approach requires workarounds (e.g., instructor role-players) or a tradeoff in fidelity when trainees reach a point in skills training that requires communication from other crewmembers while still conducting standalone training tasks. With recent technological advances in speech recognition (Stensrud, Newton, Atkinson & Killilea, 2015), the feasibility of incorporating synthetic role-playing crewmembers into a dynamic training event has increased. This paper highlights the need for this technology within the target transition community, the P-8A Poseidon, as part of its part-task training simulator. Successful integration will promote efficient use of resources (e.g., manpower), increased fidelity through the availability of realistic crew communication and coordination, and flexibility in crew composition availability. The prototype architecture is discussed, including the integration of speech capabilities (e.g., recognition, dialog, understanding, synthesis) and behavior modeling to yield an interactive model for P-8A crewmember agents. Next, the authors provide lessons learned and challenges to the technological implementation, as well as the sustainment, given the rapid pace of tactic and protocol changes that will impact the underlying technologies. Additionally, the authors provide results of a preliminary usability analysis of the system, including primary stakeholder fleet evaluations regarding system reliability and synthetic voice analysis. Finally, the authors highlight the importance of performance testing, offer suggestions for adapting the technology to other use cases, and discuss future directions for interactive system research and development.

ABOUT THE AUTHORS

Ms. Beth F. Wheeler Atkinson is a Senior Research Psychologist at the Naval Air Warfare Center Training Systems Division (NAWCTSD), and lead of the Basic & Applied Training & Technologies for Learning & Evaluation (BATTLE) Laboratory. She manages several Research and Development (R&D) efforts devoted to investigating capability enhancements for training and operational environments. Since 2007, she has supported several R & D efforts regarding novel technology enhancements to provide virtual role-players in individual and team training systems to optimize scenario-based training requiring a crew composition. Her research interests include instructional technologies (e.g., performance measurement, post-mission reporting/review), Human Computer Interaction (HCI)/user interface design and analysis, and aviation safety training and operations. She holds an M.A. in Psychology, Applied Experimental Concentration, from the University of West Florida.

Brian Stensrud, Ph.D., is a Senior Scientist at Soar Technology, Inc. (SoarTech). He is the principal investigator of the ongoing Science and Technology (S&T) effort described in this paper, and is co-lead of the Interactive Systems Technology division at SoarTech. Brian received a Ph.D. in Computer Engineering from the University of Central Florida (2005), and B.S. degrees in Mathematics and Electrical Engineering from the University of Florida (2001). Dr. Stensrud has over 16 years of experience in artificial intelligence, behavior modeling, and simulation.

John P. Killilea is a Research Psychologist supporting the Naval Air Warfare Center Training Systems Division in the BATTLE Laboratory. He holds a Masters in Modeling & Simulation, and is currently a Doctoral Candidate in the same field at the University of Central Florida.

Robert Marinier, Ph.D., graduated from the University of Michigan in 2008. His thesis focused on the integration of emotion and learning in the Soar Cognitive Architecture. He has been doing Department of Defense research for over 8 years. His work at SoarTech has included exploring applications of Soar to a variety of domains, including ground robot navigation, and intelligent training. In many projects he designed and implemented mission-level behaviors in Soar.

Paul Schermerhorn, Ph.D., is a Senior AI Engineer at Soar Technology. He has over 10 years of experience in human-robot interaction, robotic architectures, and agent-based modeling. His robotics work focused on enabling people to interact with robots using natural spoken language. He has experience with integrating speech recognition components and natural language understanding systems into robotic architectures, and with dialogue systems that enable productive natural human-computer interactions. As Vice President of Thinking Robots, he was responsible for integrated robotic architectures that combine low-level navigation and manipulation principles with high-level cognitive abilities (such as goal and plan representations, reasoning mechanisms, and most importantly natural language capabilities). He was an Assistant Research Scientist in the School of Informatics at Indiana University, Bloomington, where he served as Assistant Director of the Human-Robot Interaction Laboratory. Paul holds a BA in Psychology from Goshen College, a MA in Philosophy from Northern Illinois University, and MS and PhD degrees in Computer Science from the University of Notre Dame.

Chad Dettmering earned his Bachelor's degree in Computer Science from the University of Central Florida in 2011. He is a Software Engineer at Soar Technology. He has 6 years of experience in designing and implementing high quality software in the simulations and training field. Throughout his career he has worked with researchers to fully realize their ideas in the form of service-based architecture software. The focus of most of these projects has been to improve the learning process in military and government applications.

Sohaib Saadat is an undergraduate Computer Science student at the University of Michigan's College of Engineering. He is a Software Engineering Intern at Soar Technology. Sohaib has experience working with biomechanical prostheses and cardiac devices. He has also worked on class projects involving AI, including a Bayesian machine learning forum post classifier and an interactive Euchre card game simulator.

Emily Anania is currently a Doctoral Candidate in Human Factors at Embry-Riddle Aeronautical University. At the time of this work, Emily Anania was an intern in the Naval Research Enterprise Internship Program (NREIP) at the NAWCTSD in Orlando, FL.

Crew Role-Players Enabled by Automated Technology Enhancements

Beth F. Wheeler Atkinson
Naval Air Warfare Center Training Systems Division
Orlando, FL
beth.atkinson@navy.mil

John P. Killilea
Stracon Services Group
Orlando, FL
john.killilea.ctr@navy.mil

Brian Stensrud, Bob Marinier, Paul Schermerhorn, Chad Dettmering, Sohaib Saadat
Soar Technology, Inc.
Orlando, FL
stensrud@soartech.com, bob.marinier@soartech.com, paul.schermerhorn@soartech.com, chad.dettmering@soartech.com, sohaib.saadat@soartech.com

Emily C. Anania
Embry Riddle Aeronautical University
Daytona Beach, FL
ananiae1@my.erau.edu

INTRODUCTION

A long-standing training challenge that has been compounded by increased reliance on simulation-based training is how to train a single trainee in a task that requires a crew or group to complete it. Typical solutions include rotating trainees (e.g., having others act as training aids for individual training), or using instructors as role-players. However, these solutions can be inefficient and costly to implement. Further, in the case of using instructors as role-players, this distracts from their primary objective, which is to facilitate learning through performance monitoring and providing appropriate and timely feedback. Fortunately, with recent technological advances in speech recognition and continued advancements in behavior modeling, the feasibility of incorporating synthetic role-playing crewmembers into dynamic training has increased.

Within naval aviation training, the P-8A Poseidon serves as an appropriate and willing testbed to further explore this problem set. The P-8A is a maritime patrol aircraft that is responsible for Anti-Submarine Warfare (ASW), Anti-Surface Warfare (SUW), and Intelligence, Surveillance, and Reconnaissance (ISR) missions. The crew, flying in a modified Boeing 737, includes two pilots, a complement of approximately four sensor operators, and a Tactical Coordinator (TACCO) and Co-TACCO to lead the tactical decision-making. The crew, working in unison to bring the full capabilities of each of their roles to the mission, requires close communication and coordination. As crucial as this is, the specialization of each position, number of crew members, and need to coordinate with personnel outside the aircrew increases the importance of inter- and intra- crew communication.

Naval aviation uses the family of simulators approach to training that enables trainees to build on skills progressively throughout the training pipeline. The crawl-walk-run progression begins with system skills (e.g., *buttonology*), continues to individual tasks (e.g., understanding radar data), and concludes with aircrew coordination for tactical proficiency (e.g., prosecuting an ASW mission). After classroom-based lectures, initial system skills are first learned using table-top trainers or computer-based training. Early training in these types of environments provides opportunities to interact with specific systems to learn required knowledge, skills, and tasks, such as how to use the aircraft radio or interacting with a sensor. Next, the crewmember's individual tasks are trained using a Part-Task Trainer (PTT), a full fidelity replica of their mission crew workstation. Within this environment, trainees have an opportunity to learn complex skills and practice tasks that leverage all available aircraft systems. Lastly, aircrew coordination tasks are trained in a high-fidelity whole-task, whole-crew trainer, which is essentially a fuselage of the aircraft.

Although the interim training step in the PTT is necessary, it does result in some challenges. While this training style is sufficient when practicing simple procedural tasks (e.g., deploying a sonobuoy), maritime missions increasingly involve communication among onboard crew and coordination with other platforms (e.g., organizing ASW with nearby submarines and ships). This *walk* phase in the training approach requires workarounds (e.g., instructor role-players) or a significant tradeoff in fidelity when trainees reach a point in skills training that requires communication from other crewmembers while still conducting standalone training tasks within a PTT. Opportunities to practice

mission-critical soft behavioral skills like Crew Resource Management (CRM) are paramount to warfighter readiness and success. Hence, the P-8A crew in general, and TACCOs in particular, are hampered by these limitations of existing training capabilities to support communication and coordination activities. Further challenges are experienced in the *run* phases of training that involve a single aircrew. Optimally, P-8A flight crews would train alongside other trainees or human role-players to gain skills for interacting within a full mission scenario; however, this is often a cost- and time-prohibitive route. As such, a key requirement for affordable, persistent, on-demand Live, Virtual and Constructive (LVC) training is the integration of interactive synthetic role-players to support flight crew trainees anytime, anywhere.

With recent technological advances in speech recognition (Stensrud, Newton, Atkinson, & Killilea, 2015), the feasibility of incorporating synthetic role-playing crewmembers into a dynamic training event has increased. This paper highlights the need for this technology within the target transition community, the P-8A, as part of its PTT simulator. Successful integration will promote efficient use of resources, increased fidelity through the availability of realistic crew communication and coordination, and flexibility in crew composition availability. The prototype system is discussed, including the integration of speech capabilities (e.g., recognition, dialog, understanding, synthesis) and behavior modeling to yield an interactive model for P-8A crewmember agents. Next, the authors provide lessons learned and challenges that exist related to the implementation of the technology, as well as the sustainment, given the rapid pace of tactic and protocol changes that will impact the underlying technologies. Additionally, the authors provide results of preliminary usability analysis of the system, to include primary stakeholder fleet evaluations regarding system reliability and synthetic voice analysis. Finally, the authors offer suggestions on ways to increase the flexibility of the system to provide a common and modular capability that is adaptable to other use cases.

CURRENT PRACTICE

The operational need for the aforementioned technology lies in the intricacy of a single instructor controlling multiple PTT scenarios while role-playing. That is, within this specific platform, instructors may be responsible for leading training for up to three trainees, each interacting with their own individual training event. Given the complexity of the training system itself, the instructor is responsible for monitoring multiple systems (e.g., instructor operator station, semi-automated forces). This task alone requires significant attention, but instructors must also multi-task/task switch to monitor trainee performance and interject role-playing communications as required. For this reason, trainees receive limited crew communication and coordination training within this environment. Due to this training gap, the fleet has limited opportunities to enhance communication skills until later in the training pipeline. In order for aircrew to receive this training, the current training paradigm would necessitate the assembly of an entire crew, or the use of Subject Matter Experts (SMEs) to support crew training through role-playing. Although the quality of training crew coordination, communication, and teamwork improves when full crews are brought together to train, it is also costly in terms of resources. Thus, a core challenge is determining how individual training could benefit from the added realism, albeit synthetic, provided by crew interaction. Specifically, when attaining skills associated with crew roles, the emphasis is on the crew member's individual skills. However, many tasks associated with their role rely on inputs from other crewmembers. This is especially true for the TACCO, who acts as the information nexus during the mission, relying heavily on communication and coordination with his or her sensor operators and flight crew. While this use case and on-going research efforts discussed in this paper focuses on the P-8A, the present training problem, and use of automatic speech recognition as a solution, has been examined before, albeit with various levels of success.

Previous Automatic Speech Recognition (ASR) Efforts

Previous work related to Automatic Speech Recognition (ASR) has had notable potential military applications for decades, such as: surveillance, data entry, command and control systems, security, and communication (Beek, Neuberg, & Hodge, 1977; Griol, Herrero, & Molina, 2016). Another potential application currently being investigated and developed is the use of ASR for training and simulation use. ASR can conserve many resources, such as personnel, time, and ultimately money. By using ASR systems, trainees can interact with systems through speech, instead of interacting with confederates or instructors.

As an example, the Institute of Creative Technologies (ICT) is currently developing what they term "virtual humans," agents who ideally will interact with individuals in an organic way, with both speech and gesture, to provide users with a comprehensive training environment (Kenny et al., 2007). Similarly, other efforts have been made to integrate speech recognition technology in a "tactical language training system" which functions more like trainees playing a video game, learning foreign language and culture (Johnson, Marsella, & Vilhjalmsson, 2004). These virtual agents use speech recognition to understand and respond to trainees, although Johnson, and colleagues (2004) note that speech recognition capabilities are especially limited when it comes to language learners. Other domains, such as Air Traffic Control (ATC) technologies use ASR, both in a training and operational context. The use of an ASR tool has been shown to reduce controller workload (Helmke, Ohneiser, Mühlhausen, & Wies, 2016), as well as have useful applications for measuring workload after the fact, through collection of what they term ATC "events" (Cordero, Rodrıguez, Miguel, & Dorado, 2013). However, Cordero and colleagues (2013) do recognize that the ASR system is time-consuming to train.

Speech recognition and speech-to-text technologies have also been part of attempts to evaluate team performance, particularly for after-action reviews (Foltz, LaVoie, Oberbreckling, Chatham, & Psotka, 2008; Foltz, LaVoie, Oberbreckling, & Rosenstein, 2007). However, Foltz and colleagues (2008) cite several issues in the application of ASR for team performance assessment. Specifically, they identified a word-error rate of over 30% and found difficulty in assessing performance in real time. While ASR technologies in this context are used to monitor and assess in near-real time or after the fact, the word-error rate may make real-time monitoring and assessment not only difficult, but also yield data that is incorrect.

Addressing the Capability Gap

Although previous research on using ASR in training systems has yielded mixed results, the promise of this technology persists as successful development and implementation can solve long-standing training challenges in various domains. As previously noted, the current solutions for enabling aircrew training without a full aircrew are inefficient and yield mixed effectiveness. Developing a software suite that provides a synthetic role-playing capability serves to enhance the training pipeline and potentially avoid costs by providing value added. Specifically, the resulting benefit is the ability to provide the trainee with the realistic communication and coordination required for training, without the need for an entire crew or complement of SME role-players. Additionally, instructors who also role-play will be free to focus on assessing the trainee's performance and providing quality and timely feedback to enhance training. Although this effort seeks to address the training gap identified by the P-8A platform, other platforms have existing unmet requirements for virtual crewmembers or wingmen that this effort can provide guidance to inform future development.

CREW ROLE-PLAYERS ENABLED BY AUTOMATED TECHNOLOGY ENHANCEMENTS (CREATE)

CREATE is a research and development (R&D) effort focused on designing a synthetic role-playing capability that is anticipated to fill the aforementioned gap. The current scope of the effort involves the development of synthetic, interactive P-8A crew member agents (including the acoustic warfare officer, electronic warfare officer, Co-TACCO, pilot, and ordnance support personnel) to support TACCO-centric training. To successfully implement a synthetic role-player capability requires a technology solution that integrates (a) speech capabilities (i.e., recognition, understanding, synthesis), (b) SME-level tactical domain information, (c) reaction to multitasking and high stress situations, and (d) relay of information via means other than speech communication (e.g., software inputs).

TACCOs are responsible for synthesizing information from multiple sensor operators, as well as their flight crew, to make tactical decisions about how to pursue targets of interest. For this reason, most TACCO activities involve interacting verbally or non-verbally with other members of the crew. For example, a TACCO might need to request data from a sensor operator, coordinate or direct activities with the flight deck, or maintain shared situation awareness with the rest of the crew or command and control organizations outside the on-board aircrew. The novelty of this technology is effectively integrating speech recognition into training tasks, allowing the interactive synthetic agents to simulate the actions and communication of other crewmembers. However, emerging technological advances in science and technology often encounter challenges when translating to applied settings.

Technical Challenges and Solutions

A robust speech capability is especially important in the P-8A training context, as interactions between crewmembers are frequent and rarely follow doctrinal phraseology. In our work developing interactive agents for training systems, we divide this problem into three sub-problems: *speech recognition*, *speech understanding*, and *dialogue management*. The first two are highly interdependent. Speech recognition is the first phase, in which the audio of the spoken utterance is analyzed and words are extracted. At this point, assuming nothing has gone wrong in the speech recognition phase, the agent knows *what was said*, but nothing about *what was meant*. Speech understanding attempts to glean meaning from the words, whether they denote, for example, a new directive, a response to a previous query, or a request for new information.

Many factors influence the performance of speech recognition systems (Shneiderman, 2000). Interference from noisy environments (particularly noise in "spurts") can render speech unintelligible to a speech recognizer; people speaking in the background and the noise from a passing truck are examples of this kind of interference. Variation between speakers also has a substantial impact on speech recognizer performance, given that recognizers rely heavily on pronunciation models, regional accents, enunciation differences (e.g., mumbling), and even voice pitch (e.g., whether the speaker is male or female, old or young) and the rate of speech are examples of speaker variation that affect recognition performance. Many of these examples could also pose problems for human listeners, but people are still much better at ignoring the noise than computer speech recognizers. Likewise, speech understanding performance is affected by a variety of factors (Jurafsky & Martin, 2008). Speech recognizer performance has a considerable influence, as incorrectly recognized utterances are often hard to correctly understand. Moreover, people often do not speak in "canonical" sentences, and it can be hard to anticipate the many ways in which people might express the same meaning. People often do not fully express the meaning of an utterance, or use pronouns and other referents from earlier in the dialogue, forcing the listener to "fill in the blanks" based on context. Effective speech understanding must find ways to address these challenges.

Our team is currently developing and testing a multi-pronged speech recognition and dialog strategy to provide the necessary capability to support flexible, non-doctrinal interactions between the human trainee and the automated support roles. This approach includes the following activities:

- **A robust P-8A interaction grammar**. While P-8A phraseology amongst the crew is not standardized, there are known best practices and patterns for issuing/requesting information, tasking, and reports that can be encoded into our grammar. This ensures that utterances fitting known patterns will be recognized fully by the system and parsed into a machine-readable format for ingestion. With support from SMEs and the fleet, this grammar can be extended as necessary to support alternate patterns (different ways to say the same things) as those alternatives are identified.
- **A text pattern-recognition parser.** We additionally encode less structured patterns (expressed as *regular expressions)* that can match incoming phrases that are not matched by the grammar above. This parser, when run in parallel with the full grammar, can supplement the primary recognition pathway with either full or *partial* matches on incoming utterances. In cases of a partial match, the software agent can potentially make a high-confidence guess about the message based on known context. Regular expressions can be used to detect full phrase patterns or even simple patterns, such as known keywords.
- **Agent-based dialog management**. Any artifacts from the speech recognition pipeline will also need to be interpreted correctly by the software role-players to complete a dialogue. Our approach to handling the task of fusing these artifacts and placing utterances in the correct conversational context, determining to whom the trainee is speaking, etc., leverages existing work on the Smart Interaction Device (SID). SID will be responsible for both understanding incoming speech and also managing dialogue between each P-8A agent and the trainee. As recognition artifacts are generated and sent to the agent using the Aria infrastructure, SID will process and make sense of them in terms of the agent's knowledge base and understanding of the current dialogue state. In this sense, SID serves as a fusion engine for the various speech recognition artifacts, using the agent's knowledge base and known context to recognize and react to trainee utterances: (1) Is the agent waiting for a particular cue or piece of information? (2) Is a question from the trainee an appropriate thing to expect at the current moment? (3) Does executing the requested task make sense given current activity? (4) Does the message have sufficient information, or do I need clarification? By framing all interactions alongside the rest of the agent's tactical knowledge base, agents can use dialogue to both advance the conversation and also respond to unclear statements. For instance, if communication fails for some reason (e.g., the user gives ambiguous

inputs or omits some information), the system is designed to ask the user for clarification or for the missing information.

PRELIMINARY ANALYSIS

Speech data were collected with representatives from the Patrol and Reconnaissance Wing Eleven (CPRW-11) to support an engineering evaluation of the speech recognition capability that underlies the CREATE technology. The purpose of this analysis was to establish a baseline for the speech recognition performance in the domain early in the development process to identify areas for improvement. In this analysis, we used the P-8A aircrew's audio samples against a fixed, hand-generated speech grammar developed during early prototyping. During the analysis, researchers identified specific gaps and opportunities for improvements that can be addressed during on-going development to advance the state-of-the-practice while increasing the likelihood of transition success.

Two collection events yielded data for analysis from skilled aircrew from the P-8A: the first event at Naval Air Station Jacksonville and a second event at NAWCTSD in Orlando. Additional data collection was conducted internally at SoarTech with seven untrained civilians (three of which had some familiarity with the domain). In total, 2,350 audio samples were collected across 17 participants.

During data collection, participants were asked to record themselves speaking a sequence of utterances randomly selected from the set of in-grammar utterances. Recording took place using a push-to-talk button, so users controlled when the recording started and stopped. Roughly half of the recordings were made using a laptop's built-in microphone, while the other half were made using a noise-cancelling headset microphone.

These recordings were then fed through a speech pipeline to measure its performance. This pipeline starts with speech recognition, which provides a text hypothesis of what was said. The next stage of the evaluation runs the recognition hypotheses through the understanding phase of the speech pipeline, to see whether it can produce correct semantics (i.e., whether the semantics for the recognition hypothesis were identical to those for the expected utterance), even in cases where the hypothesis was not completely correct. In some cases, speech recognition may fail (i.e., not exactly match the expected text), but understanding still succeeds. This is because the understanding phase is typically forgiving about simple variants (e.g., the word "the" may be optional in some places, and the word "AWO" may be substituted for "Jez", etc.). In this context, the most appropriate metric of speech performance and success is correct understanding.

Comparisons were made between recognition and understanding rates with and without a headset. To give a sense of how "close" the failed recognitions were, the word edit distance was calculated for each one (i.e., Levenshtein distance; Levenshtein, 1966). This distance represents the number of edits—word replacements, deletions, or insertions—required to transform the recognition hypothesis into the expected utterance.

Finally, we analyzed the audio itself in order to better understand the causes of failures. Specifically, we identified and quantified utterances that were cutoff at the beginning or end (which is due to user error), utterances in which the user did not say the correct thing (e.g., added, dropped, or substituted words), and high degrees of noise. Due to resource limitations, we were unable to listen to and categorize every recording. Instead, we sampled recordings across all datasets, selecting approximately proportionally from each dataset (i.e., larger datasets contributed more samples). We then extrapolated from these samples to estimate the total number of failures due to each issue. Of course, multiple issues may apply to any given sample.

Results of Testing

Baseline performance results are shown in Table 1. The length of utterances in each condition was essentially the same, as we would expect from a sufficient random sample. Clearly, a noise-cancelling headset made a sizable difference in recognition and understanding accuracy. Additionally, the understanding phase was robust enough to some misrecognitions to improve performance by a few points in both cases.

Table 1. Speech performance

	Avg. Length	Recognition	Understanding
No Headset	10.9	39.7%	44.8%
Headset	10.8	51.8%	55.0%

Analysis of the audio itself is shown in Table 2. While some utterances were correctly recognized despite the errors, the error rates were generally higher for incorrect utterances, as expected. The exception was noise in the no headset case; we expect that this was lower because there were other failure causes in the incorrect cases, but in the correct case those were lower, leaving a higher proportion of noise. Additionally, high noise rate only affected the no headset condition, which explains why overall recognition and understanding rates were lower in that case. Cutoffs were also high, implying that some participants did not understand how to use the push-to-talk function (i.e., they started recording too late or ended too early). Finally, utterance variations were also high in the incorrect cases. Examples of these include incorrect pronunciation (e.g., saying "E-W-O" as separate letters, or saying "four zero zero" instead of "four hundred"), repeating "TACCO" at the start of the utterance, and word addition or omission (e.g., leaving out "and", or adding "the"). There were also occasionally mis-reads, where the wrong word was said entirely, and stutters. Finally, sometimes different terminology was substituted (e.g., "three" instead of "EWO").

Table 2. Audio analysis

	High noise?	Cutoff?	Utterance variation?
No Headset Correct	41.7%	12.5%	4.2%
No Headset Incorrect	24.5%	22.5%	36.7%
Headset Correct	0%	10.6%	11.8%
Headset Incorrect	0%	31.9%	42.0%

The final analysis was conducted to understand failures for the headset condition. The results, wherein word edit distance is characterized, are shown in Table 3. Also shown in Table 3 are the percentages of utterances that were one edit away from correct, within two edits, and within three edits. This indicates that the majority of understanding failures were actually very close to correct. We also characterized it with the cutoff utterances removed, as we would expect real-world users to not make this mistake.

Table 3. Word edit distance analysis for understanding errors

	1 Edit	1-2 Edits	1-3 Edits
Headset	39.6%	63.9%	81.2%

LESSONS LEARNED

One prominent lesson learned is that noise-cancelling headsets are critical for good speech performance. Most modern speech recognition technologies still struggle with background noise. Additionally, utterances that are cut off by users hitting the push-to-talk button too late or releasing too early, have a detrimental impact on recognition and understanding. It is likely that these issues can be addressed for future data collection through training with the data collection tool, or through use of equipment that more closely replicates the technology used in training. With these changes, we would expect these rates to be much lower among the target training population. There may be other techniques to help with this as well, such as continuous recording in which the push-to-talk signal is treated as a hint as to where utterances start and end, but the system is intelligent about actually looking for gaps in speech.

Many utterance variations can likely be addressed via improved training or increased use of end users (e.g., a real user would never say "E-W-O" as separate letters). Further, improved parsing such as adding support for variants would also likely result in increased speech system performance. Additionally, users were often off by only a small amount in their utterances. This indicates that small improvements in recognition performance can make a substantial difference. Another approach would be to take advantage of small edit distances. Fixing small edit

distances may also help address cutoffs, as typically these affect just the first and/or last word in an utterance. Overall, many of these changes are within relatively easy reach.

Table 4 shows prospective understanding performance with various issues addressed. This assumes a noise-cancelling headset, which already addresses the noise problems. The columns include estimated understanding rates that could be expected from fixing each of the issues individually; the final column identifying the estimated understanding rate that would be achievable if we fixed all issues identified by these analyses. Because many samples have multiple problems, these do not strictly add (i.e., fixing one class of problems may also fix some of another class of problems). If each issue were fixed individually, there would be moderate improvements in performance, with increasing edit distance fixes making the largest difference. If all fixes were combined, for this dataset we would theoretically achieve 98% understanding. While it is probably unrealistic to achieve 100% fixes for each of these areas, this estimation demonstrates that focusing effort in this space of problems is likely to result in performance gains that would increase the likelihood of technology success as part of the CREATE training tool.

Table 4. Speculative performance improvements to understanding from addressing identified speech failure categories

	Base	Address Cutoffs	Address Utterance Variations	Address 1 Edit Failures	Address 1 & 2 Edit Failures	Address 1, 2, & 3 Edit Failures	Cumulative
Headset	55.1%	69.4%	74.0%	72.9%	83.8%	91.6%	98.0%

CONCLUSION

While these analyses highlighted several challenges that exist with this prototype speech recognition and understanding technology, the overall effort is still in the early R & D phase. By collecting this data early, as highlighted by the results above, several achievable methods for improving performance were identified. Further, the analyses and results discussed here were done under a minimal budget without impacts to the development schedule. Through early, frequent, and continued analyses such as this, the team can continue to identify common issues that, when addressed, can potentially provide performance gains. This is critical for an interactive technology such as the CREATE training tool, because human-computer interaction significantly impacts user perceptions of technology. From usability issues that decrease buy-in for using technology, to unintended negative training when the system reacts in ways that do not mirror the real-world environment, poor performance of technologies such as speech tools often result in failed transitions. Or worse, transition of technologies that remain unused due to challenges that users perceive as insurmountable. Further, from an instructor perspective, performance shortfalls of an interactive technology have the potential to increase workload and impact their trust in this or other automated systems.

As this research effort continues in the coming months, the team will continue to collect similar data sets for further analysis and system refinement. In addition, for a better understanding of challenges that may exist within the larger interactive technology, the team proposes to collect usability data from end users when using an interactive speech capability (i.e., full up speech recognition, understanding, and synthesis), as well as a full CREATE prototype. Finally, as a full-scale solution is prototyped, research into instructor workload and displays that provide transparency on automated systems that underpin the technology are necessary to deliver an effective and efficient training tool that facilitates instructional processes.

The primary use case for prototype development has been TACCO training within the PTT. Existing funding will expand this use case to include training for other operators within the PTT environment and expanded mission sets. However, other opportunities exist to increase training efficiencies for the P-8A as well. For example, pilot training conducted within the Operational Flight Trainers (OFT) currently lack the noise and interaction that would be encountered during a live flight. In this use case, the technology would be expanded to provide background chatter overheard on the radio when monitoring command and control frequencies and provide the ability to interact with air traffic control organizations. Additionally, as noted in the introduction, crew-based training is accomplished in the later phases of training using a Weapons Tactics Trainer. During these training events, a full crew is expected; however, there are occasions when competing priorities or other factors (e.g., illness) result in missing

crewmembers. As the technology matures, the CREATE training tool provides an option to "turn on" desired synthetic crew members to fill out an aircrew for training, without requiring another operator to fill the gap.

While the use case for this specific effort is currently the P-8A, the training challenges described are prevalent across multiple aviation platforms and domains. For example, even though most fighters only have a single individual comprising the crew, formation flying and integrated warfare requires interaction with others in the mission environment. For this reason, expanding the technology to provide options such as virtual wingmen would be follow-on use cases. Further, due to the limited availability of cross-platform training due to training schedules and resources, this technology provides unique opportunities to *train as we fight* even during standalone training events.

Advancements in ASR and component technologies have provided opportunities to develop an integrated capability that provides virtual crewmembers in simulation-based training environments through the development of synthetic, interactive models. The results of this paper highlight that while not insurmountable, developers of emerging technologies must take the steps necessary to identify possible points of failure early during design and development to overcome pitfalls that may lead to failure. Further, this technology does not represent an end to this challenge on its own. Rather, continued R&D is required to address additional training and lifecycle issues that face speech systems such as increased utility for automated performance assessment (e.g., Foltz et al., 2007, 2008) and mechanisms to maintain speech libraries after fielding to overcome outdated technologies as tactics, techniques, and systems change over time.

ACKNOWLEDGEMENTS

The views expressed herein are those of the authors and do not necessarily reflect the official position of the DoD or its components. Sponsors for underlying R&D that have informed the CREATE effort from a Navy perspective have included the Naval Air Systems Command (NAVAIR) PMA-290 and the Small Business Innovative Research/Small Business Technology Transfer (SBIR/STTR) program.

REFERENCES

Beek, B., Neuberg, E., & Hodge, D. (1977). An assessment of the technology of automatic speech recognition for military applications. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *25*(4), 310-322.

Cordero, J. M., Rodrıguez, N., Miguel, J., & Dorado, M. (2013). Automated speech recognition in controller communications applied to workload measurement. *Third SESAR Innovation Days*, *26*, 28.

Foltz, P., LaVoie, N., Oberbreckling, R., Chatham, R., & Psotka, J. (2008, December). DARCAAT: DARPA competence assessment and alarms for teams. In *Proceedings of the 2008 Interservice/Industry Training, Simulation & Education Conference*.

Foltz, P. W., Lavoie, N., Oberbreckling, R., & Rosenstein, M. (2007). Tools for automated analysis of networked verbal communication. *Network Science Report*, *1*, 1.

Griol, D., Herrero, J. G., & Molina, J. M. (2016, November). Military usages of speech and language technologies: A review. In *Meeting Security Challenges Through Data Analytics and Decision Support* (pp. 44-68).

Helmke, H., Ohneiser, O., Mühlhausen, T., & Wies, M. (2016, September). Reducing controller workload with automatic speech recognition. In *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th* (pp. 1-10). IEEE.

Johnson, W. L., Marsella, S., & Vilhjalmsson, H. (2004). The DARWARS tactical language training system. *Interservice/Industry Training, Simulation, and Education Conference*.

Jurafsky, D., & Martin, J. H. (2008). Speech and language processing (prentice hall series in artificial intelligence).

Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsella, S., & Piepol, D. (2007). Building interactive virtual humans for training environments. *Interservice/Industry Training, Simulation, and Education Conference*.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady (Vol. 10*, No. 8, pp. 707-710).

Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM*, *43*(9), 63-65.

Stensrud, B., Newton, C., Atkinson, B., & Killilea, J. (2015). Automatic Speech Recognition in Training Systems: Misconceptions, Challenges, and Paths Forward. *Proceedings of the Interservice/Industry Training, Simulation, & Education Conference,* Orlando, FL.