2011 NDIA GROUND VEHICLE SYSTEMS ENGINEERING AND TECHNOLOGY SYMPOSIUM
ROBOTIC SYSTEMS (RS) MINI-SYMPOSIUM
AUGUST 9-11 DEARBORN, MICHIGAN

HANDS-FREE, HEADS-UP CONTROL SYSTEM FOR UNMANNED GROUND VEHICLES

Jonathan Brown
Think-A-Move, Ltd.
Beachwood, OH

Jeremy P. Gray
U.S. Army TARDEC
Warren, MI

Chris Blanco
Amit Juneja, Ph. D.
Think-A-Move, Ltd.
Beachwood, OH

Joel Alberts
Autonomous Solutions, Inc.
Petersboro, UT

Lauren Reinerman
University of Central Florida
Orlando, FL

ABSTRACT

This paper describes work to develop a hands-free, heads-up control system for Unmanned Ground Vehicles (UGVs) under an SBIR Phase I contract. Industry is building upon pioneering work that it has done in creating a speech recognition system that works well in noisy environments, by developing a robust key word spotting algorithm enabling UGV Operators to give speech commands to the UGV completely hands-free. Industry will also research and develop two sub-vocal control modes: whisper speech and teeth clicks. Industry is also developing a system that will enable the Operator to drive a UGV, with a high level of fidelity, to a location selected by the Operator using hands-free commands in conjunction with image segmentation and video overlays. This Phase I effort will culminate in a proof-of-concept demonstration of a hands-free, heads-up system, implemented on a small UGV, that will enable the Operator have a high level of fidelity for control of the system.

INTRODUCTION

A key problem limiting the use of Unmanned Ground Vehicles (UGVs) by Warfighters conducting dismounted operations is that current systems require Operators to keep their hands on a UGV controller of some kind (e.g. video game style), and to be heads-down while looking at a video display. This significantly reduces the Operator's Situational Awareness (SA), as well as reducing the effective strength of a squad because the Operator cannot keep his hands on his weapon while controlling the UGV, and also needs security to provide protection.

To address this issue, Industry, through a Phase I SBIR contract from TARDEC, developed proof-of-concept Hands-Free control methods for a UGV, including vocalized speech, whispered speech, and teeth clicks. Industry built upon previous work with speech control systems to enable

an operator to use speech commands without needing a Push-to-Talk (PTT) button.

Industry also examined color segmentation of video images and video overlays, such as grids, and then using speech commands to direct the UGV to a grid coordinate on the video display.

Industry also modified an existing navigation and manipulation systems to incorporate obstacle avoidance and visual servoing control techniques. This system were tested in simulation with voice selection and simple goto actions were found to be feasible and less operator intensive than teleoperation.

Industry also reviewed a number of different wrist-worn and helmet mounted displays that could be used in such a system. And, Industry worked with Academia to develop a means of evaluating a Heads-Up, Hands-Free UGV Control System to determine its impact on human factors.

The work completed in Phase I was very encouraging, and through a proposed Phase II effort, Industry hopes to develop a prototype Heads-Up, Hands-Free UGV Control System built upon the work performed in Phase I.

PHASE I OBJECTIVES

The overall objectives, as listed in the Phase I proposal and how they were met, are described in outline with more detail below:

Objective 1

Evaluate existing technologies and research both potential enhancements to those technologies as well as new technologies to determine which components will be most effective in developing a system for hands free control of an unmanned ground vehicle (UGV).

- Industry completed a market survey of different soldier worn displays. Several types of displays, including helmet mounted and wrist mounted displays were considered. Overall the Parvus system, with a wrist mounted display from Trident Systems seems to be the most feasible, but helmet mounted displays from Rockwell Collins will also be evaluated under the Phase I contract.
- Brain computer interfaces were investigated and found to be impractical for control of ground robotics.
- Whisper commands show great promise for quiet hands free operation.
- Teeth clicks were shown to be easier to use and more accurate than tongue clicks, which were initially considered.
- A proof-of-concept speech command spotting algorithm was integrated into the SPEAR™ Speech Control System.
- Industry determined that a segmentation/grid overlay system is an effective means to providing labeled goal locations for an autonomy system

Objective 2

Develop a comprehensive system design to be implemented in Phase II.

A system with a wrist worn or helmet mounted display, mobile computer, implementing vocalized speech, whisper, teeth click command modes, and using video servoing with a grid overlay will be the basis for a full prototype to be developed in a proposed Phase II effort.

Objective 3

Assess the feasibility of the system design through a limited proof of concept demonstration.

- Industry implemented the use of a grid overlay with speech commands.
- A video of the system in operation will be submitted with the final technical report.

PHASE I RESULTS

Sub-vocal commands

Industry developed two sub-vocal modes of giving commands to a UGV in Phase I – whisper commands and teeth click sequences. The whisper command mode was designed for UGV control in environments where the operator needs to conceal his location from enemy personnel. The teeth click command mode involves making Morse code-like sequences of teeth clicks that map to different commands for the UGV. The teeth click mode was designed to enable UGV control in both low and high noise environments while concealing the location of the operator in both of the environments. The results of the Phase 1 work on whisper and teeth click systems are presented below:

Whisper recognition results

The goal of the whisper command project was to recognize 10 whispered speech commands recorded through the SPEAR in-ear earpiece with a the threshold accuracy of 90% for 5 whisper commands, and with the objective accuracy goal of 97% with 10 commands in quiet environment. These commands are single and multiple word commands recorded in the push-to-talk (PTT) mode. A database of whispered

speech was recorded by Industry where 4 speakers were asked to speak 10 commands, 5 times each, in succession. The commands used are listed below:

1. Stop
2. Forward
3. Backward
4. Turn left
5. Turn right
6. Faster
7. Slower
8. Flippers forward
9. Flippers back
10. Light toggle

Industry conducted research to find an appropriate method for recognition of whispered speech, and found that Dynamic Time Warping (DTW) [1] is the most suitable approach. The procedure involves a profile training module asking the user to speak each command one or more times. The procedure proposed is then used to create a single normalized template for each command for every speaker. Using one normalized template improves the speed during testing. An alternative training procedure where templates for all instances of each command are stored during training is used to get better accuracy.

To process the input whisper signal 13 Mel-frequency cepstral coefficients (MFCCs) with cepstral mean and variance normalization [2] were applied with a window size of 30 ms and a window shift of 10 ms. Before calculating the MFCCs each input signal is normalized to a zero mean unit variance signal (this can be done because of the assumption that the signal for the full word is available for this isolated word recognition scenario). Similarity distance obtained by calculating the correlation coefficient between two vectors of MFCC coefficients was used in the DTW procedure.

The following table shows the results of whisper command recognition in a quiet office environment.

Speaker	Accuracy (%) with single template	Accuracy (%) with 4 templates
1	95.0	97.5
2	90.0	92.5
3	100.0	100.0
4	97.5	100.0
Overall	95.62	97.5

Using the training procedure that stores multiple templates an accuracy of 97.5% is achieved that surpasses the objective goal of 97%. The GUI for the whisper command mode that industry developed is show in Figure 1.

Figure 1 Whisper Control Command GUI

Teeth click detection results

A tongue click recognition system was originally proposed in which a user would click his/her tongue to the upper palate and a transient signal would be detected in the ear canal by the in-ear microphone. Tongue clicks would be done in a certain sequence pattern, for example, a "2-3" pattern would involve 2 clicks followed by a pause and then followed by 3 clicks. Different patterns could then be mapped to different commands for a UGV.

A database of 5 speakers generating tongue clicks was collected by Industry for development and evaluation purposes. It was found that users found it difficult to reliably generate tongue clicks, but they could easily and consistently generate mild teeth clicks. The teeth click signals can be reliably captured by the in-ear microphone in the ear canal. On the basis of this research Industry has developed a teeth click system instead of the tongue click system.

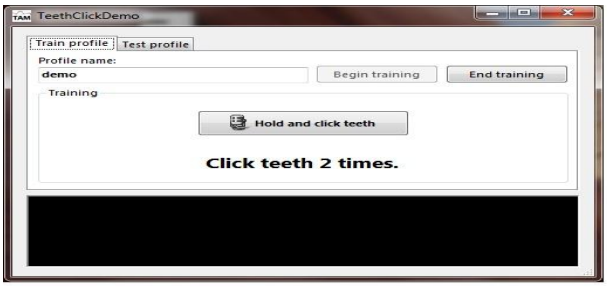


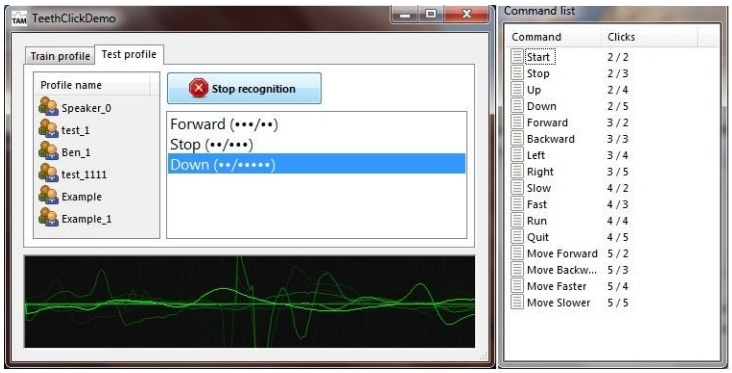
Figure 2 Teeth Click Training GUI

The teeth click system has been developed using a time-domain matched filter algorithm. In the profile training phase the user is asked to generate some teeth clicks. The generated teeth clicks are recorded and a representative form is stored. At the time of decoding of the teeth clicks the signal from the microphone is run through a matched filter that is essentially the representative signal for that speaker.

A teeth click is detected whenever the output of the matched filter is greater than a certain threshold. The teeth clicks are detected continuously and occurrence of the patterns that map to a command is observed. Teeth click detection results were evaluated on data collected from five speakers who made two repetitions of each the following combinations of teeth click sequences - "2-3", "4-5", "2-2", "4-4", "5-3", "4-2", "4-3" and "3-5".

Figure 2 shows a screen shot of the GUI that industry developed for a user to train a profile using teeth clicks. And Figure 3 shows the teeth click command GUI, including the commands that were implemented in Phase I, along with the teeth click combinations used to actuate those commands.

Figure 3 Teeth Click Command GUI



The percentage of combinations that were correctly detected using the matched filter approach is shown in the following table in quiet, 80dBA and 90dBA noise. Greater than 90% accuracy was obtained in quiet and 80dBA ambient noise, and 78% accuracy was obtained in 90dBA noise.

There is a very significant scope of improvement in 90dBA noise by using a matched filter in the Wigner time-frequency domain [3], but this implementation was out of the scope of Phase 1 and is planned for Phase II. A test was also conducted to see if normal speech causes insertions of valid teeth click sequences. The teeth click detector was run through a speech sample for each of the five speakers. No insertions were observed from the speech signal which proves the validity of the teeth click sequence detector for command and control purposes.

Noise	Detection accuracy (%)
Quiet	97
80dBA	91
90dBA	78

Speech command discrimination

The SPEAR speech recognition system has been successfully demonstrated to control UGVs in high noise environments. A major overhaul of SPEAR has been proposed for Phase II in which a discriminative training algorithm and other algorithms shall be developed and integrated into SPEAR to improve discrimination of speech commands from other speech uttered by the user. These improvements in SPEAR shall enable the users to control UGVs completely hands free so that they do not have to turn speech recognition off when they have to converse with others, significantly improving the usability of the system in Operational scenarios.

To demonstrate the capability of good speech command discrimination a very basic algorithm was incorporated into SPEAR. Before the inclusion of this algorithm 100% of conversational speech could potentially be "inserted" as commands. The inclusion of this algorithm has reduced that amount to 80% of conversational speech that could be inserted as commands with no drop in correct recognition of valid commands. Furthermore, a "tuning knob" has been included that changes the sensitivity of the speech recognizer so that it can reject more and more conversational speech at the cost of also rejecting a certain percentage of valid commands.

For example, if the knob is set to a value setting of zero (on a scale of 0 to 10), the recognizer could potentially insert 80% of conversational speech as commands with no drop in correct recognition of valid commands. But if the knob if set to 5 out of 10, the recognizer would insert only up to 20% of conversational speech with a drop of only 5% of correct commands. The goal in Phase II shall be to allow command insertions from less than 5% of conversational speech while rejecting less than 5% of valid commands.

Brain computer interface evaluation

Industry evaluated two commercial-off-the-shelf (COTS) brain computer interfaces from two companies – Neurosky and Emotiv.

Figure 5 Neurosky Headset

Initial set up and use of the Neurosky system was very straightforward. However there were significant issues with the system itself.

Within a short period of time, the evaluator felt that he could raise and lower the concentration and meditation bars of the application with relatively good accuracy. The evaluator did find great difficulty in maintaining the bars at any specific level, and if the evaluator's attention left the application for even a moment it would change the value of the bars rendering previous readings useless, resulting in very low command recognition accuracy. In addition, latency issues plagued the system.

Figure 6 Emotiv Headset

The Emotiv system, unlike the Neurosky, required great effort to use and calibrate. It requires 16 foam tips to be wet with saline solution and then twisted on to the headset. The foam tips would routinely fall out of the headset when trying to place it on the evaluator's head. The first use required a period of three hours to get the system calibrated and all 16 sensors registering a good contact. After much practice, it typically took 20 minutes just to get the headset adjusted properly before the program reported a usable signal.

The evaluator also had great difficulty getting the Emotiv headset to register his thoughts. The signal was completely indistinguishable from random noise. Seemingly at random it would register a command. The evaluator could not find any correlation between his thoughts and what was registered on the headset's software. The evaluator expressed great frustration with the system.

Two main problems plagued both headsets: low command recognition accuracy, and a long latency. These two issues combined made it very difficult to isolate where the accuracy problems originated. Typically there was about a second delay before the headsets would pick up on a thought, if they ever did. The evaluator could not distinguish between the headset not picking up his thoughts (a command deletion) or if the system was still processing.

Both training manuals stated that training involved both the headset learning to recognize an individual's thoughts, and the person learning to concentrate think in a way that was conductive to recognition. The training was made difficult because the evaluator never knew if he was thinking correctly or if the headset was not recognizing his thoughts. Poor accuracy and lack of feedback both contributed to the difficulty in training.

Industry did not find much, if any benefit in using BCI technology for robotic control in office, much less operational situations. The systems are inaccurate, have long latency, are difficult to use, are difficult to maintain, and require complete concentration for their use. Industry does not recommend further evaluation of these technologies for control of UGVs.

Integration of target selection using different overlay methods

Industry also evaluated target selection with different overlay methods for use in vocal selection. Industry experimented with several color segmentation algorithms and determined that color scene parsing is an effective method of extracting objects for a video image when the scene has several objects have distinct texture. For more cluttered scenes, segmentation would need to be performed using a segmentation algorithm that incorporates geometry and texture.

Industry also implemented a grid overlay that calculated a major ground plane in the video and then divided this plane into grid cells. This method was useful for certain scenarios such as navigation where there is not necessarily a desired goal with distinct visual features. The overlays for each of these two methods were given letter or number labels and tested to be effective for area selection using the SPEAR system.

The following is a summary of tasks performed under this category during Phase I:

- Target Selection:
  - Integration with SPEAR speech recognition system.
  - Implemented scene parser using color segmentation.
  - Implemented scene parser using ground plane grid.
  - Tested different display methods for the overlays.
  - Tested selection using mouse input then using voice commands.
  - Created command interpreter interface.
  - Investigated laser pointer target referencing.
  - Investigated Microsoft Kinect gesture recognition software.
- Autonomy Development

During Phase I Industry designed and modified an existing navigation and manipulation systems to incorporate obstacle avoidance and visual servoing control techniques. This system were tested in simulation with voice selection and simple goto actions were found to be feasible and less operator intensive than teleoperation. More complex scenarios that include uneven terrain with cluttered terrain will be evaluated in Phase II. The Phase I tasks can be summarized as follows:

- Incorporated manipulation technology from other industry projects.
- Designed high level state machines for servoing with path planning.
- Tested grid navigation autonomy.
- Integrated navigation path planning system into Industry's Mobile Manipulation System.

- Integration
  - Tested manipulation behaviors on a small UGV, with navigation through simulation
  - Interfaced robot payload control with voice selected actions.
  - Range Sensor investigation.

Human Factors Evaluation of Proposed Systems

The purpose for this project is to evaluate the use of heads-up, hands-free systems in the operation of a UGV. To determine the usefulness of the systems, two factors have been identified that will impact the user and are pertinent to the context with which the system will be implemented: situational awareness (SA) and workload. SA is widely used as a measure to evaluate performance, decision making, and focus while operating a given system. Traditionally, SA is defined as the "perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" [4]. Workload is commonly defined as the degree to which information processing, mental effort, or cognitive resources are taxed, relative to their capacity [5].

Researchers have used a combination of both objective and subjective measures to measure both SA and Workload. Objective measures of SA include: real-time probes continuously presented based on time or location trigger in task [6]; task interruption (Situation Awareness Global Assessment Technique (SAGAT)[7]; and performance based inferences. Subjective measures of SA include: the Participant Situation Awareness Questionnaire (PSAQ) [8]; the Situation Awareness Rating Technique (SART) [9]; and expert ratings. Objective measures of workload include: EEG (Event-Related Potential, Resonance Imaging), ECG (Heart Rate Variability, Blood Volume), and eye tracking [10]. Subjective measures of workload include: NASA-Task Load Index (NASA-TLX) [10], Multiple Resource Questionnaire (MRQ) [11]; and the Subjective Workload Assessment Technique (SWAT) [12].

Considering the types of measures available for assessing SA and workload, two contributors to operational mission performance and success, an approach for evaluating the use of heads-up, hands-free systems is proposed. For SA, the planned experiment will utilize objective (performance based inference), subjective (PSAQ), and performance based measures of accuracy, reaction time, and reaction distance. In particular for the performance based assessment, performance will be evaluated independently for each task environment (from the perspective of the participant and the UGV) to more accurately determine the specific effects of the experimental manipulations on the operator's SA. For workload, the proposed study will utilize both objective (ECG) and subjective measures (NASA-TLX) to completely capture the operator's state.

Heads up display assessment

Industry completed a market survey of different soldier worn displays. Several types of displays, including helmet mounted and wrist mounted displays were considered. Overall the Parvus system, with a wrist mounted display from Trident Systems seems to be the most feasible, but helmet mounted displays from Rockwell Collins Inc. (RCI) will also be evaluated under the Phase I contract. RCI is the sole provider to the U.S. Army of soldier worn displays for the Mounted Soldier Program, and is also one of the teams competing the Nett Warrior program. By evaluating displays from RCI, Industry will therefore be considering displays that the U.S. Army is either already using or will be evaluating.

Conclusion and Proposed System Summary

Industry will continue to build on the development done in Phase I to provide a heads-up and hands-free UGV control solution for UGV control. The proposed architecture that

will be demonstrated at the end of Phase II for hands-free control of UGVs is described as follows: The UGV operator wears the SPEAR headset and a wrist worn or helmet-mounted display. The Operator calls up a virtual grid overlay on the video display or image segmentation through the use of speech commands, and uses it to direct the UGV to a specific location. He directs the UGV to using one of three modes: speech commands issued at a normal tone of voice, whisper commands, or teeth click sequences mapped to commands.

The three types of signal – whisper, teeth clicks and speech, are captured from the SPEAR earpiece with an in-ear microphone. The whisper commands and teeth clicks can be used where the operator intends to maintain silence to conceal his location from enemy personnel. Whisper commands can be used in quiet or low noise environments while teeth-clicks and speech control can be used in both low and high noise environments.

The user will be able to give speech commands to control the UGV without needing to depress a Push-to-Talk (PTT) button. This will enable the user to control the UGV hands free and allow the Operator to 'speak' to the UGV in the same way that he/she speaks with his/her team mates.

The operator gets feedback from the UGV in the form of audio played in the speaker of the SPEAR earpiece. When the UGV reaches its desired location, for example, he would receive an audio message notification, allowing him to maintain his SA, and eliminating the need to constantly look down at the video display.

In its proposed Phase II effort, Industry will also focus on solving three main problems: Segmenting an image to allow for a 'destination' for the robot to be indicated by voice command; Navigating the robot to the 'destination' indicated, without relying on GPS and by using local Obstacle Detection and Avoidance (OD/OA); Integrating these capabilities into current UGVs through an open architecture using the JAUS - Joint Architecture for Unmanned Systems interface, with sensors and processing power compatible with UGVs, will be a key goal for Phase II.

REFERENCES

[1] Rabiner, L. R. and Wilpon, J. G., "A simplified robust training procedure for speaker trained, isolated word recognition systems", J. Acoust. Soc. Am. 68(5), Nov 1980, pp. 1271-1276.

[2] Young, S., "The HTK Book", Manual for the Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/download.shtml

[3] Shin, Y., Nam, S., An, C., and Powers, E., "Design of a time-frequency domain matched filter for detection of non-stationary signals", International Conference on Acoustics, Speech and Signal Processing, ICASSP 2001.

[4] Endsley, M. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, Vol 37(1), 32-64.

[5] Moray, N. E. (1979). Mental workload: Its theory and measurement. New York, NY: Plenum Press.

[6] Jones, D., & Endsley, M. (2004). Use of Real-Time Probes for Measuring Situational Awareness. International Journal of Aviation Psychology, Vol 14(4), 343-367.

[7] Endsley, M., op. cit.

[8] Strater, L., Endsley, M., Pleban, R., & Matthews, M. (2001). Measures of Platoon Leader Situational Awareness in Virtual Decision-Making Exercises. U.S. Army Research Institute for the Behavioral and Social Sciences.

[8] Ryu, K., & Myung, R. (2005). Evaluation of Mental Workload With a Combined Measure Based on Physiological Indices During a Dual Task of Tracking and Mental Arithmetic. International Journal of Industrial Ergonomics, Vol 35, 991-1009.

[9] Taylor, R. M., 1990, Situational Awareness Rating Technique (SART): The Development of a Tool for Aircrew Systems Design, in AGARD-CP-478, Situational Awareness in Aerospace Operations. Neuilly Sur Seine, France, 3-1 – 3-17.

[10] Ryu, K., & Myung, R., op cit.

[11]Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX: Results of Empirical and Theoretical Research. In: Hancock, P.A., Meshkati, P. (Eds.), Human Mental Workload. Elsevier, Amsterdam, 139–183.

[12]Reid, G. B., Potter, S. S., and Bressler, J. R. (1989). Subjective Workload Assessment Technique (SWAT): A User's Guide. Wright Patterson Air Force Base, OH: Harry G. Armstrong Aerospace Medical Research Laboratory.