Technological evaluation of gesture and speech interfaces for enabling dismounted soldier-robot dialogue

Kattoju, Ravi Kiran, Barber, Daniel, Abich, Julian, Harris, Jonathan

Ravi Kiran Kattoju, Daniel J. Barber, Julian Abich IV, Jonathan Harris, "Technological evaluation of gesture and speech interfaces for enabling dismounted soldier-robot dialogue," Proc. SPIE 9837, Unmanned Systems Technology XVIII, 98370N (13 May 2016); doi: 10.1117/12.2223894

SPIE.

Event: SPIE Defense + Security, 2016, Baltimore, Maryland, United States

Technological Evaluation of Gesture and Speech Interfaces for Enabling Dismounted Soldier-Robot Dialogue

Ravi Kiran Kattoju*, Daniel J. Barber, Julian Abich IV, Jonathan Harris
University of Central Florida (UCF), Institute for Simulation and Training (IST), 3100 Technology Parkway, Orlando, Florida 32826, USA.

ABSTRACT

With increasing necessity for intuitive Soldier-robot communication in military operations and advancements in interactive technologies, autonomous robots have transitioned from assistance tools to functional and operational teammates able to service an array of military operations. Despite improvements in gesture and speech recognition technologies, their effectiveness in supporting Soldier-robot communication is still uncertain. The purpose of the present study was to evaluate the performance of gesture and speech interface technologies to facilitate Soldier-robot communication during a spatial-navigation task with an autonomous robot. Gesture and speech semantically based spatial-navigation commands leveraged existing lexicons for visual and verbal communication from the U.S Army field manual for visual signaling and a previously established Squad Level Vocabulary (SLV). Speech commands were recorded by a Lapel microphone and Microsoft Kinect, and classified by commercial off-the-shelf automatic speech recognition (ASR) software. Visual signals were captured and classified using a custom wireless gesture glove and software. Participants in the experiment commanded a robot to complete a simulated ISR mission in a scaled down urban scenario by delivering a sequence of gesture and speech commands, both individually and simultaneously, to the robot. Performance and reliability of gesture and speech hardware interfaces and recognition tools were analyzed and reported. Analysis of experimental results demonstrated the employed gesture technology has significant potential for enabling bidirectional Soldier-robot team dialogue based on the high classification accuracy and minimal training required to perform gesture commands.

Keywords: Human-robot interaction, gesture recognition, speech recognition, human-robot team, gesture glove, multimodal communication

1. INTRODUCTION

Advancements in interactive technology have progressed the evolution of human-robot interaction (HRI) to reflect more intuitive forms of communication portrayed in human-human interaction (HHI). The ability of these technologies to engage natural modes of interaction through which information is conveyed, such as gesture and speech, have enabled efficient human-robot teaming[1, 2]. The notion of intuitive and natural communication with robots has been a motivating factor in a variety of domains, specifically military[3]. Soldier-robot teaming is expected to significantly increase across many future combat operations[4]. Teleoperated robots successfully performed 125,000 Intelligence, Surveillance and Reconnaissance (ISR) missions in Iraq and Afghanistan[5], but teleoperated robots require a dedicated operator and increases the cognitive burden on the operator. Presently, most Unmanned Ground Vehicles (UGV) deployed for ISR missions employ a particular technique of teleoperation enabled by a joystick, track ball, or touch screen[3]. However, as robots are developed with the capabilities of increased autonomy and ability to communicate through various modalities, significant advances in human-robot communication technology are required.

1.1 Background

Multimodal communication through interfaces that support natural language modalities such as gesture and speech have been demonstrated to be robust interaction methods for enabling bidirectional human-robot communication [1,6,7]. The integration of these interfaces for HRI plays a crucial role in facilitating the development of robust and autonomous robotic systems by supporting alternative interaction methods[1]. Autonomous robots equipped with gesture and speech interfaces establish communication structures permitting Soldiers to communicate to the robot without being encumbered by a teleoperating control device. In addition, these natural language interfaces developed by leveraging

*rkattoju@ist.ucf.edu; phone 1 407 882-1428; fax 1 407 882-1335; http://prodigy.ist.ucf.edu/

Unmanned Systems Technology XVIII, edited by Robert E. Karlsen, Douglas W. Gage, Charles M. Shoemaker, Grant R. Gerhart, Proc. of SPIE Vol. 9837, 98370N
 · doi: 10.1117/12.2223894

existing HHI principles[8] present a more intuitive approach with reduced training requirements, lowered task load, and increased situational awareness of human teammates[9,10] which benefit dismounted Soldier- robot teaming.

Advances in speech detection and recognition through the development of high fidelity microphones[11] and speech recognition software, such as the Microsoft Speech Platform SDK, Google API, and CMU pocket sphinx, have enabled the detection, recording and classification of speech commands to a high level of accuracy and reliability[12, 13]. Automatic Speech Recognition (ASR) software have improved significantly through the use of Hidden Markov Models (HMM) and other machine learning algorithms that utilize their ability to match input to speech models and perform classification using acoustic, lexicon, and language models to find the most probable sequence of words in speech[14,15]. Despite the increase in classification accuracy of ASR software for instructional commands, they are faced with challenges such as reduced classification accuracy when employed for translating high-level commands (natural and intuitive commands used in HHI). The classification accuracy of ASR is also affected when integrated with an autonomous robot and exposed to real world conditions due to extraneous noise, distortions, and multiple or simultaneous users. The Natural Language Processing (NLP) in the ASR algorithm functions to convert speech to text by mapping the speech components in the command to corresponding words in a dictionary. However, when used with complex or high level commands, it has also shown to impact the ASR classification accuracies[16]. Improving the ASR system accuracy and reducing the classification errors are the main challenges for developing a robust speech recognition software capable of accounting for the issues mentioned above. Besides speech, gestures are also an important aspect of natural language communication that have been employed for human-robot communication.

Gestures are distinctive actions containing brief segments of motion of the hand, arm or posture mapped to a signal, instruction, or a command. The effectiveness of gesture recognition depends on the capability of the hardware and software technologies employed to extract characteristic features and patterns from an input stream of gesture information[2]. The choice of gesture recognition hardware and software for a specific application are dependent on: 1) segmentation of the input stream of gesture data (beginning and end of a gesture), 2) choice of gestures for a particular application (in this case non-line-of-sight Soldier-robot interaction), 3) selection of best machine learning algorithms to accurately classify gestures, and 4) the ability of the recognizer to distinguish between closely related gestures by eliminating false positive gesture classification[2,17,18]. Over the last decade, gesture recognition for HRI has received extensive attention, where motion of the hand, arm, head and full body are tracked in different ways for determining their position and orientation[19] and translated to control commands for robots. Most commonly, gesture information is collected using mechanical, electro-magnetic[17], optical[20], and inertial sensors[1,17] and gesture classification is commonly conducted through the use of different machine learning algorithms such as HMM[21], Conditional random fields (CRF)[22-24], Support vector machine (SVM)[25], and Decision trees[26,27]. Research has also been conducted on finger tracking by utilizing electromyography[28, 29] and acoustic signal[30] sensors for identifying finger gestures through muscle data and acoustic signals transmitted through the skin[31]. However, a common approach for tracking hand and finger movement is through the use of instrumented inertial measurement unit (IMU) based gloves[2]. Inclination toward reducing burden caused by carrying a gesture capturing device in Soldier-robot communication have led to subsequent advancements in gesture recognition hardware which resulted in the creation of a custom gesture glove integrated with an Attitude, Heading and Reference System (AHRS)[1].The gesture glove (illustrated in Figure 1) eliminates the need to carry a physical control device, allowing human teammates to perform secondary functions or tasks with their hands while communicating with the robot. Research by Barber et al. compared performance of different IMU devices while classifying 21 unique gesture commands adapted from the U.S. Army field Manual[1, 32]. Although this gesture recognition system delivered high classification accuracies (98%) for command components under laboratory conditions which included training samples contributed from the users, the real-time performance and classification accuracies of high level intuitive commands without training samples from the users is warranted.

A pilot study was conducted and reported on the ability of gesture and speech interfaces for classification of high level commands in a simulated ISR mission[6]. It focused mainly on classification of the high level commands using different speech recognition hardware and software, and illustrated the potential of gesture recognition for classifying those commands. Building on the pilot study, this paper focuses on the technological evaluation of gesture and speech classification technology across single and dual modalities for enabling bi-directional Soldier-robot dialogue during a simulated ISR mission. The classification results for gesture and speech detection hardware and recognition software across the single and dual modality combinations were analyzed and compared. The gesture and speech commands for

the spatial-navigation experiment task were extracted from U.S Army field manuals[32], Department of Defense (DoD) dictionary of Military terms[33], and Squad Level Vocabulary (SLV) surveys[35].

Figure 1. Instrumented gesture glove. IMU sensor is located on the back of the hand, with flex resistors sewn into the fingers of the glove. A wired module the user straps to the forearm contains XBee wireless hardware and battery.

1.2 Experiment Approach

The data acquired to conduct the present analyses were gathered from previous work[36, 37]. A repeated measures experimental design was developed to utilize gesture and speech modalities as independent variables and accuracy of gesture or speech recognition technology as dependent variables. Participants commanded an autonomous robot using gesture and speech modalities individually (single modality- gesture or speech only) and simultaneously (dual modality- gesture and speech together) for completion of three spatial-navigation scenarios that resembled ISR missions. Mission routes and levels of communication modalities were counterbalanced across participants to minimize order effects and impact of route sequence on modality.

2. METHODS

2.1 Participants

Forty-two participants (males = 21, females = 21) with a mean age of 21.4 (*SD* = 4.1) participated in the study. All participants were fluent in English and performed gestures with their right hand and arm. Qualified participants were awarded class credits as compensation.

2.2 Experiment Environment

A simplified model of an urban city block with two buildings surrounded by obstacles was designed for the simulated ISR mission experiment. The buildings represented by boxes were surrounded by small orange cones that acted as obstacles. In the simulated ISR mission, participants commanded the robot to navigate to predetermined locations and the robot autonomously circumnavigated through obstacles in order to reach designated waypoints. A route sequence map containing information regarding the location points and command sequence were provided to the participants. For each location point, the participant was required to deliver three specific high level commands to the robot: Report, Move, and Screen. The 'Report' command instructed the robot to report any obstacles in its line of sight, the 'Move' command instructed the robot to move to a particular location and the 'Screen' command instructed inspection of the region for targets. The high level command types and variations are described in the experiment procedure below (Table 1).

Table 1. High level command types and variations

Command Type	**Command Variations**
Report	Report obstacles to the {north, south, east, west} side of the {east, west} building
Move	Move to the {north, south, east, west} side of the {east, west} building
Screen	Screen
* For all commands, options with braces, {and}, are options	

2.3 Equipment

2.3.1 Gesture Recognition Hardware

A gesture glove was used to capture the participants' arm and hand movements. The gesture glove incorporated an Attitude, Heading and Reference System (AHRS) with the Razor 9 Degrees of Freedom (DOF) IMU containing four built in sensors: single axis gyro, dual axis gyro, triple axis accelerometer and triple axis magnetometer. The Razor IMU was selected based on the gauging capabilities of its accelerometers for forces up to ±16g to support gestures at higher forces and thereby preventing saturation of values. Additionally, the magnetometers in the Razor IMU enabled compass heading with their ability to reference global directions. Flexible resistors were embedded in the back of the glove fingers for measuring change in resistances caused by finger bend positions to indicate closed or open fist. Finally, an XBee module established wireless communication between the glove and computer used for gesture recognition. The AHRS and the XBee modules were assembled in an enclosure and mounted on the glove which permitted unimpeded motion of hand and arm, for performing gestures. Figure 1 above illustrated the gesture glove integrated with the IMU and XBee modules.

Additionally, a Microsoft Kinect was placed two meters in front of the participants on a 70 cm tall stand to collect skeletal and 3d depth information, while the participant was performing gestures. This information was utilized post-experiment to aid in transcription and verification of the gestures performed.

2.3.2 Speech Recognition Hardware

A dedicated clip-on wireless lapel microphone supporting frequencies ranging from 50 Hz to 15 kHz was used to capture speech commands with a sensitivity of ±3db. Furthermore, the Microsoft Xbox Kinect, utilized to record skeletal gesture data, also contained a multi-array microphone that detected and recorded speech commands with high fidelity.

2.3.3 Gesture and Speech Classification Software

A custom gesture classification software called the "Gesture Builder" was designed in-house to capture and process raw data from the gesture glove for classification of hand and finger gestures participants performed. The gesture recognition software utilized objects that constantly acquired and processed raw sensor data from the glove such as acceleration from the accelerometer, angular velocity from the gyrometer, orientation from the magnetometer, and finger position from flex resistors. The Gesture Builder was trained to recognize and classify the closing and opening of the hand and a set of nine distinct gestures (described in Table 2 below) gathered from the U.S. Army field Manual for Visual Signals[32] and modified American Sign Language[34]. After processing the raw data, gesture recognition classifiers in the Gesture Builder detected and classified gesture commands that included first, closing the hand into a fist to indicate the start of the command, then moving the arm and hand to perform the command, and finally opening the fist to indicate the end of the command.

All audio streams containing participants' recordings of the speech commands were converted to mp3 format at a bit rate of 192 kbps and were analyzed at 48 KHz with 16 bits per sample by a Commercial Off-the-Shelf (COTS) Microsoft speech detection platform. For the purpose of gesture and speech recognition, speech commands from the previously established SLV[35] were broken down to constituent parts and mapped with nine distinct gestures listed in Table 2. The speech recognition software was customized from the Microsoft Speech Platform SDK (version 11) to identify and classify constituent parts of the speech commands. Raw data from the microphones were detected, recorded, and processed to classify speech commands delivered by the participants.

Both gesture and speech recognition classifiers were run post-experiment to verify and validate the gesture and the speech command components respectively. Accuracy of command components and mean across command components were calculated for both gesture and speech modalities.

Table 2. The list of nine hand and arm gesture command components and descriptions of each.

Number	Gesture command	Description
1	Report Obstacles to	Adopted night visual signal for 'Start Engine' or 'Prepare to Move' derived from FM21-60[32].
2	Move to	Adopted visual signal for 'Assemble' or 'Rally' from FM21-60[32].
3	The North	Modified version of American Sign Language (ASL) [34] "North" gesture, but without the hand signal (eg. with closed fist).
4	The South	Modified version of American Sign Language (ASL) [34] "South" gesture, but without the hand signal (eg. with closed fist).
5	The East	Modified version of American Sign Language (ASL) [34] "East" gesture, but without the hand signal (eg. with closed fist).
6	The West	Modified version of American Sign Language (ASL) [34] "West" gesture, but without the hand signal (eg. with closed fist).
7	Side of	Modified version of American Sign Language (ASL) [34] "North" gesture, with hand signal (eg. with closed fist).
8	Building	Custom gesture developed for the experiment. Starting at head level with closed fist on your left side pointing away from the body, palm facing down, draw an imaginary box moving to right, down, left and back up to the start position while keeping fist parallel to the floor.
9	Screen	Custom gesture developed for the experiment. Starting at head level with closed fist pointing away from the body, palm facing down, draw capital letter "S" while keeping fist parallel to the floor.

2.4 Procedure

2.4.1 Experiment

Participants were required to complete training sessions on gesture and speech based interaction with the robot using gesture and speech commands. During the training sessions, participants were first educated on reading the route sequence map and command structures. Participants were shown examples and demonstrations of individual command types (Report, Move, and Screen) and practice ISR missions using both gesture and speech modalities.

Subsequent to the training session, participants were required to successfully demonstrate their ability to properly construct the gesture and speech commands. Qualifying in the first part of the quiz required participants to achieve at least 84% accuracy while performing a speech command based ISR mission comprising of eight reconnaissance locations (three command types for each reconnaissance location). The second part of the quiz evaluated participants' ability to implement and achieve at least 95% accuracy while performing hand and arm gestures. A total of nine gestures were developed and mapped to the speech command components in the command types as shown in Table 2.

Upon successful completion of both the gesture and speech quiz, participants were given instructions on interaction with the robot. After completing a command (gesture or speech), the robot provided feedback to the participant on a 30 inch monitor placed in front of the participant. The feedback message was displayed for three seconds and after the message on the display was cleared (and robot finished navigating) participants could initiate the next gesture or speech command. As stated previously, classification of the gestures and speech using software was done post-experiment. During the experiment, researchers determined if the participant issued the correct command, and if so, initiated the robots execution. This was done "Wizard of Oz" (WoZ) style, where the participants were told the robot actually classified gestures and speech in real time[36].

2.4.2 Measuring Gesture and Speech Classification Performance

The Gesture Builder acquired raw sensor data from AHRS and flexible resistors in the gesture glove, which were used to extract feature vectors for classification of gestures. Flexible resistors in the glove fingers measured the bend in the participants' fingers, which either was a closed fist or open hand indicating the beginning and end of a gesture command, respectively. The Gesture Builder application records raw data, processes and extracts features, and classifies gestures. Gesture extraction and classification from raw data is supported both real-time and post-hoc. Additionally, the Gesture Builder application had the capability to export gesture classifications to different real-time applications, enabling direct use of gestures for command and control of a robot. The effectiveness of the gesture glove and software was measured by determining the system's ability to accurately classify the nine distinct gestures (described above in Table 2) as command components and complete commands. During the experiment, the gestures performed by the participant were visually observed and validated by the researchers for robot execution. However, since the WoZ approached was used, the Gesture Builder's classification accuracy was validated post-experiment using transcriptions of performed gestures from skeletal model tracking data collected with the Microsoft Kinect.

Similarly, the Microsoft Speech Platform SDK was customized for detecting and classifying constituent parts of the speech commands. Effectiveness of the ASR was determined from accuracy of the software to detect and classify command components. The microphone levels and quality were validated before the start of each experiment and automatic gain control was disabled to prevent interference with the speech detection software. Participants then performed all experiment conditions in random order. The gesture and speech data were used to address following three research questions. These questions correspond to three analyses reported in the results section:

(i) Does gesture classification across single and dual modalities differ significantly from each other?
(ii) Does speech detection and classification across single and dual modalities differ significantly with respect to different hardware?
(iii) Does gesture and speech classification across single and dual modalities differ significantly from each other?

3. RESULTS

A one-way repeated measures ANOVA with Bonferroni correction was performed post-hoc for each of three different research questions. Degrees of freedom, F-scores, partial eta-squared and Cohen's *d* effect sizes using the conventional scale of .2, .5, & .8 (small, medium, & large, respectively) for command components (Report, North, South, East, West, Move, Side, Building and Screen) and overall are reported for each of the analyses. In the following analyses, 'overall' refers to the mean accuracy across command components. The sample size varied for each of these tests through list wise deletion of participants due to hardware issues or participant performance errors.

3.1 Analysis 1: Comparison of Gesture Classification Across Single and Dual Modalities

A one-way repeated measure ANOVA was performed on a sample size of 34 participants for gesture detection across single modality (gesture only) and dual modality (gesture and speech). The sample size decreased from 42 to 34 participants through list wise deletion due to hardware issues in the gesture glove flex resistors and participant performance error. No statistically significant differences were found between the modalities ($p > .05$; Table 3). Overall, the trend in the data shows that mean classification accuracies across gestures were higher for single modality ($M = .82$, $SD = .12$) compared to dual modality ($M = .81$, $SD = .14$), and the range of Cohen's *d* for the command components was < .001 for North to 0.3 for Building. Table 3 presents the main effects and effect sizes for the nine gesture command components and across single and dual modalities. Figure 2 illustrates the mean classification accuracies for gesture command components across both modalities.

Table 3. ANOVA results showing main effects for gesture classification across single and dual modalities. Command component, degrees of freedom, F-score, partial-eta-squared, and Cohen's *d* are all reported.

Command Component	D.F.	*F*	η^2_p	*d*
Report	1, 33	.06	.002	.03
North	1, 33	< .001	< .001	< .001
South	1, 33	.45	.013	.10
East	1, 33	.10	.003	.05
West	1, 33	1.43	.042	.25
Move	1, 33	.04	.843	.01
Side	1, 33	1.87	.054	.15
Building	1, 33	12.62	.277	.30
Screen	1, 33	.01	< .001	.02
Overall	1, 33	2.39	.07	.12

Note: $p > .05$

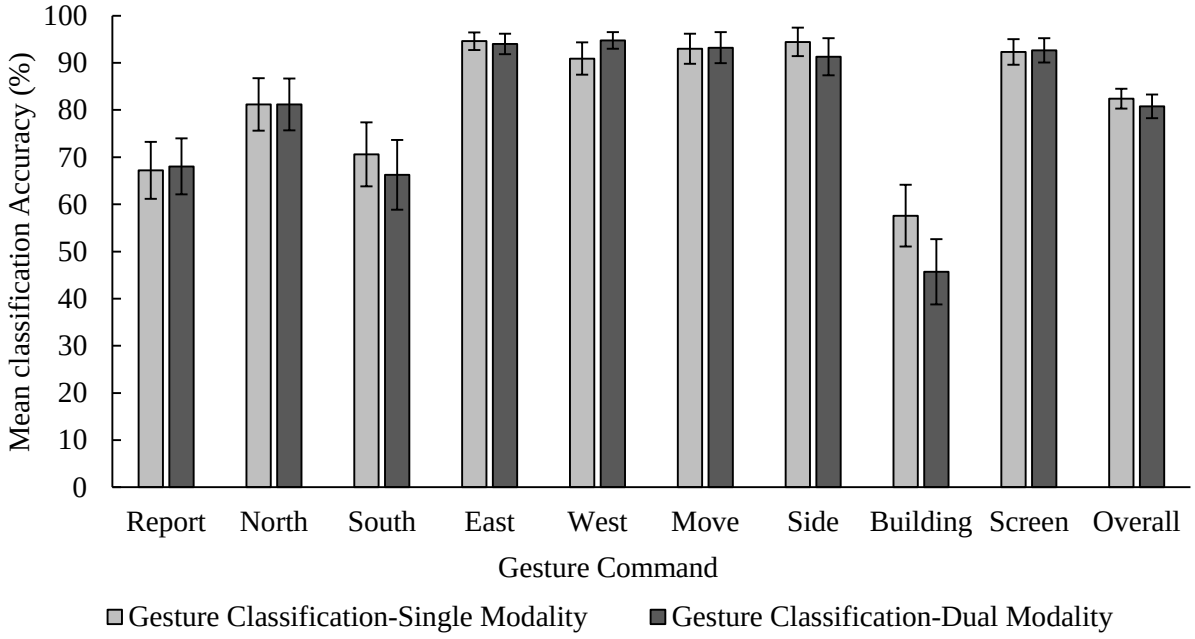

Figure 2. Mean classification accuracy for gesture classification across single and dual modalities. Error bars are included in the graph.

3.2 Analysis 2: Comparison of Speech Detection and Classification Across Single and Dual Modalities

A 2 (Modality: Speech, Gesture and Speech) x 2 (Hardware: Kinect, Microphone) repeated measures ANOVA was performed on a sample size of 26 participants for speech detection using the Kinect and Microphone across single and dual modalities. The sample size decreased from 42 to 26 participants through list wise deletion due to poor recording quality of the lapel microphone and inaudible participants. Significant main effects were found between the modalities (Table 4). Overall, the trend in the data shows that mean classification accuracies across speech command components were higher for single modality ($M = .76$, $SD = .16$) compared to dual modality ($M = .23$, $SD = .20$). All speech command components except the Screen command showed significant main effects with *d*s in the range of 0.02 for Screen to 3.07 for West. Table 4 presents the main effects, and effect sizes for the nine command components across single and dual modalities. Figure 3 below illustrates the mean classification accuracies of speech classification for speech command components across both modalities.

Table 4. ANOVA results showing main effects for speech recognition across single and dual modalities. Command component, degrees of freedom, F-score, partial-eta-squared, and Cohen's *d* are all reported.

Command Component	**D.F.**	***F***	η^2_p	***d***
Report	1, 25	165.86**	.87	2.88
North	1, 25	121.87**	.83	2.44
South	1, 25	167.05**	.87	3.04
East	1, 25	135.58**	.84	2.83
West	1, 25	194.2**	.89	3.07
Move	1, 25	133.8**	.84	2.78
Side	1, 25	150.37**	.86	2.99
Building	1, 25	153.69**	.86	2.99
Screen	1, 25	0.01*	< .001	0.02
Overall	1, 25	157.858**	.86	2.9

Note: *$p < .05$, **$p < .001$

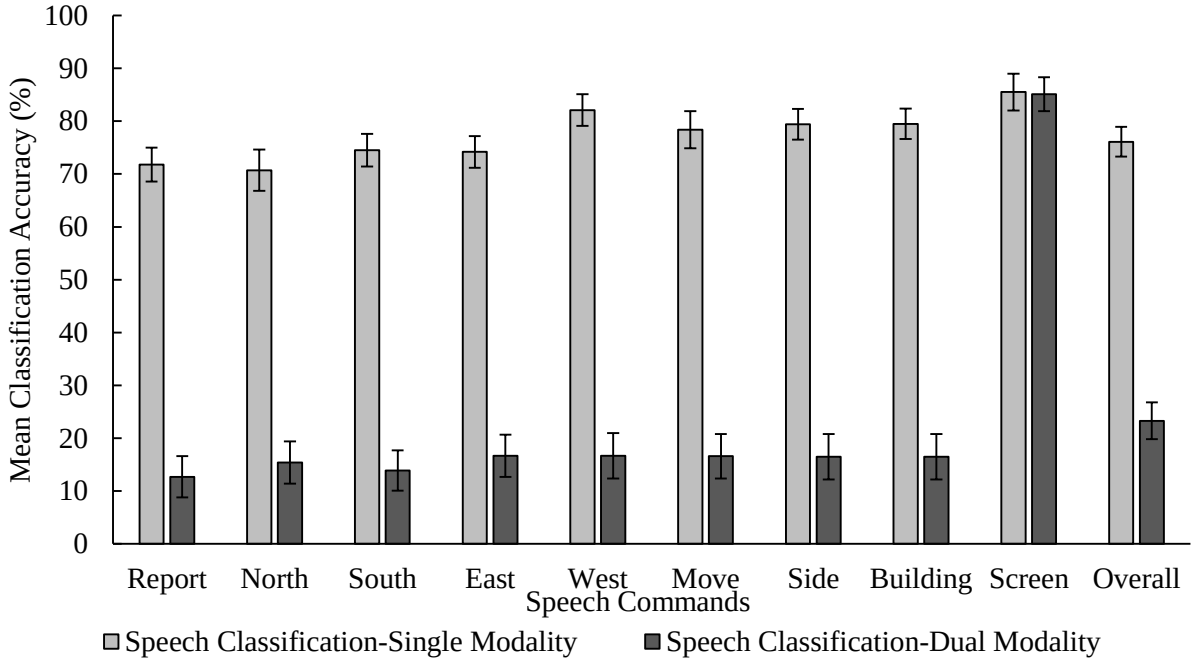

Figure 3. Mean classification accuracy for speech command components and overall across single and dual modalities. Error bars are included in the graph.

Significant main effects were found between the speech hardware for almost every command component (Table 5). The trend in the data shows that mean classification accuracies across speech command components were higher for speech detection using Kinect (M = .64, SD = .15) compared to Microphone (M = .35, SD = .2). All speech command components except the Screen command showed significant main effects with *d*s in the range of 0.55 for Screen to 1.79 for East. Table 5 presents the main effects and effect sizes for the nine speech command components using both hardware across single and dual modalities. Figure 4 below illustrates the mean classification accuracies for speech detection using Kinect and Microphone for speech command components across both modalities.

Table 5. ANOVA results showing main effects for speech recognition across Kinect and Microphone. Command component, degrees of freedom, F-score, partial-eta-squared, and Cohen's *d* are all reported.

Command Component	**D.F.**	***F***	η^2_p	***d***
Report	1, 25	33.26**	.57	1.24
North	1, 25	44.93**	.64	1.46
South	1, 25	55.47**	.69	1.72
East	1, 25	64.5**	.72	1.79
West	1, 25	53.64**	.68	1.31
Move	1, 25	43.88**	.64	1.40
Side	1, 25	52.17**	.68	1.54
Building	1, 25	51.95**	.68	1.53
Screen	1, 25	4.53*	.15	.55
Overall	1, 25	58.28**	.7	1.64

Note: *$p < .05$, **$p < .001$

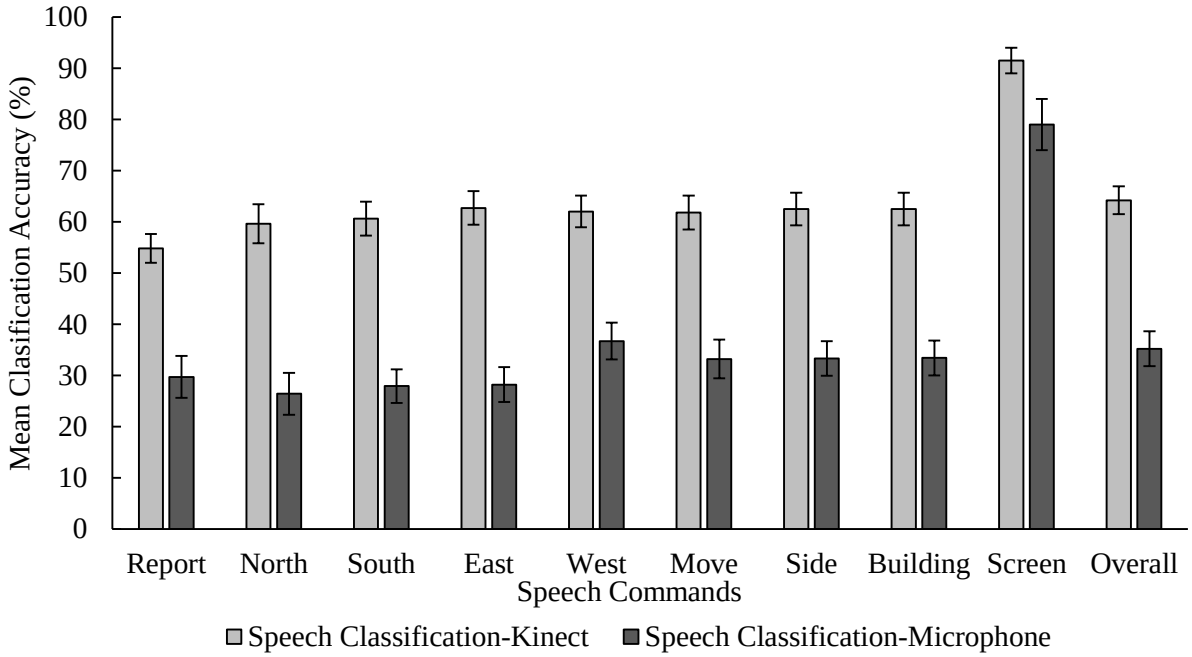

Figure 4. Mean classification accuracy for speech command components and overall across Kinect and microphone for both modalities. Error bars are included in the graph.

Among the speech command components, significant interaction effects were found between the type of modality and the speech hardware for Report, North, South and East commands while no significant interaction effects were seen in the other command components. The Kinect showed mean classification accuracies of 96.6 % (single modality) and 31.8 % (dual modality) while the lapel microphone showed 55.5 % (single modality) and 14.9 % (dual modality). Table 6 illustrates the interaction effects and effect sizes for the nine command components as follows.

Table 6. ANOVA results showing interaction effects for speech recognition between modality and hardware. Command component, degrees of freedom, F-score, and partial-eta-squared are all reported

Command Component	D.F.	F	η^2_p
Report	1, 25	14.68**	.37
North	1, 25	15.36**	.38
South	1, 25	13.65**	.35
East	1, 25	14.8**	.37
West	1, 25	4.04*	.14
Move	1, 25	6.07**	.20
Side	1, 25	8.81**	.26
Building	1, 25	8.56**	.26
Screen	1, 25	.43*	.02
Overall	1, 25	15.77**	.39

Note: *$p < .1$, **$p < .05$

The Microsoft Kinect exhibited higher classification accuracies compared to the lapel microphone. In order to show the trend in classification accuracies for the two different speech hardware (Microsoft Kinect and Lapel microphone) across the two different modalities (single and dual modality), a box plot was used to depict the minimum, maximum, median, first and third quartiles for speech recognition presented in Figure 5.

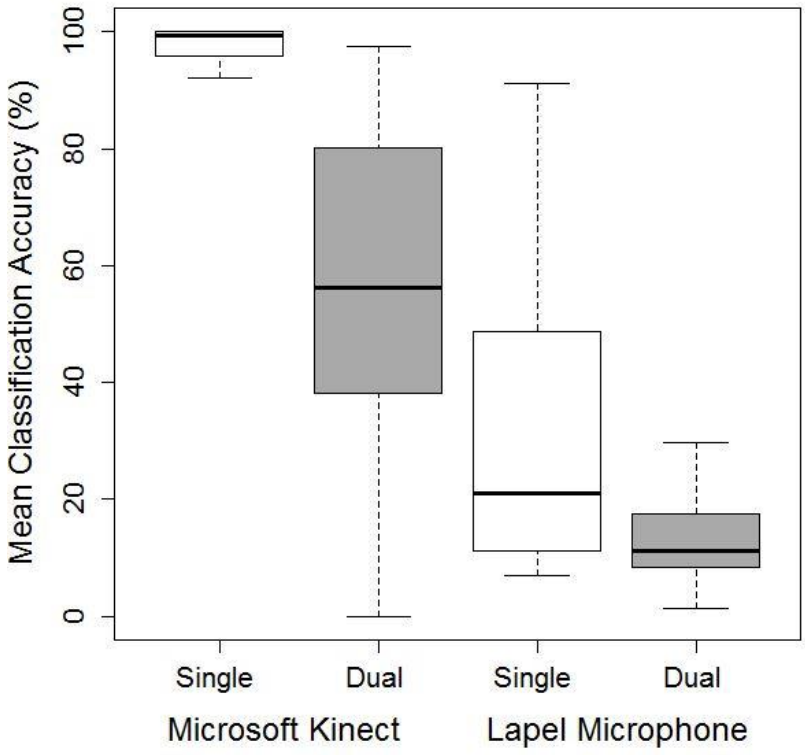

Figure 5. Mean classification accuracy for speech recognition across Kinect and Microphone across single and dual modalities showing minimum, maximum, median, first and third quartile values.

A sample size of 35 participants who successfully delivered the speech commands across both modalities were considered for comparing speech command durations (time taken from start to finish of a speech command). The mean, median and standard deviations of command duration (in seconds) for the commands: Report, Move, Screen and Overall (mean of Report, Move, Screen commands) are reported in Table 7.

Table 7. Mean command duration for delivery of speech commands across single (speech only) and dual (gesture and speech) modalities

Command	Single Modality			Dual Modality		
	Mean Duration (Secs)	SD	Median	Mean Duration (Secs)	SD	Median
Report	5.56	1.30	5.62	12.26	2.51	11.79
Move	4.76	.79	4.58	11.72	2.35	12.00
Screen	2.35	.21	2.30	2.57	1.06	2.35
Overall	4.26	.54	4.22	8.80	1.67	8.65

One way repeated measures ANOVAs were performed separately on the means and medians of the command durations for the speech commands: Report, Move, Screen, and Overall (Table 8 & 9). The overall mean command duration was calculated as the mean of the Report, Move and Screen command durations. Participants delivered speech commands significantly faster in the single modality than in the dual modality with mean command duration for single modality (M = 4.26, SD = .54) in comparison to dual modality (M = 8.80, SD = 1.67). The range of ds was .31 for Screen to 3.90 for Move. The command duration for delivering the Report and Move commands varied significantly and showed an increase across single to dual modalities, whereas the Screen command duration showed no significant differences across both modalities ($p > .05$).

Table 8. ANOVA results showing main effects for mean command durations for speech recognition across single and dual modalities. Command, degrees of freedom, F-score, partial-eta-squared, and Cohen's d are all reported.

Command	D.F.	F	η^2_p	d
Report	1, 34	272.69**	.89	3.39
Move	1, 34	362.21**	.92	3.90
Screen	1, 34	2.39*	.06	.31
Overall	1, 34	330.99**	.91	3.68

Note: *$p < .01$, **$p < .001$

Table 9. ANOVA results showing main effects for median of command durations for speech recognition across single and dual modalities. Command, degrees of freedom, F-score, partial-eta-squared, and Cohen's d are all reported.

Command	D.F.	F	η^2_p	d
Report	1, 34	312.57**	.90	3.55
Move	1, 34	342.65**	.91	3.91
Screen	1, 34	2.08*	.06	.26
Overall	1, 34	353.26**	.91	3.90

Note: *$p < .01$, **$p < .001$

3.3 Analysis 3: Comparison of Gesture and Speech Classification Across Single and Dual Modalities

A 2 (Mode: Single, Dual) x 2 (Modality: Speech, Gesture) repeated measures ANOVA was performed on a sample size of 33 participants for gesture and speech detection across single and dual modalities. The sample size decreased from 42 to 33 participants through list wise deletion or due to participant performance error. As a result of the higher performance and mean classification accuracies of the Microsoft Kinect in speech recognition across both modalities, analysis 3 considered only speech detection using the Microsoft Kinect. Significant main effects were found between modes, such that the mean classification accuracies were higher for gesture and speech detection across the single modality (M = .90, SD = .07) in comparison to the dual modality (M = .55, SD = .14). All command components except the Screen command showed significant main effects with ds in the range of 0.18 for Screen to 3.34 for East. Table 10 presents the main effects and effect sizes for the nine individual commands for gesture and speech across single and dual modalities.

Table 10. Main effects for modes (single or dual) across gesture and speech modalities. Command component, degrees of freedom, F-score, partial-eta-squared, and Cohen's *d* are all reported.

Command Component	**D.F.**	***F***	η^2_p	***d***
Report	1, 32	166.92**	.84	1.80
North	1, 32	120.57**	.79	1.66
South	1, 32	84.20**	.73	1.71
East	1, 32	249.94**	.89	3.34
West	1, 32	140.60**	.82	2.53
Move	1, 32	170.89**	.84	2.39
Side	1, 32	159.63**	.83	2.86
Building	1, 32	229.27**	.88	2.01
Screen	1, 32	.868*	.03	.18
Overall	1, 32	245.16**	.89	2.96

Note: *$p < .1$, **$p < .001$

In general, significant main effects were found between the modalities, such that the mean classification accuracies were higher for speech detection ($M = .81$, $SD = .13$) in comparison to the gesture detection ($M = .63$, $SD = .13$). All individual commands except the Screen command showed significant main effects with *d*s in the range of 0.07 for Screen to 2.37 for East. Table 11 presents the main effects and effect sizes for the nine individual commands for gesture and speech across single and dual modalities. Figure 6 below illustrates the mean classification accuracies for gesture and speech classification for command components across both modalities.

Table 11. Main effects for modality (gesture or speech) across single and dual modalities. Command component, degrees of freedom, F-score, partial-eta-squared, and Cohen's *d* are all reported.

Command Component	**D.F.**	***F***	η^2_p	***d***
Report	1, 32	5.36*	.14	.52
North	1, 32	9.65*	.23	.76
South	1, 32	1.07*	.03	.26
East	1, 32	83.34**	.72	2.37
West	1, 32	95.56**	.75	2.17
Move	1, 32	58.86**	.65	1.80
Side	1, 32	37.48**	.54	1.74
Building	1, 32	2.61*	.08	.41
Screen	1, 32	.09*	.003	.07
Overall	1, 32	29.17**	.48	1.32

Note: *$p < .5$, **$p < .001$

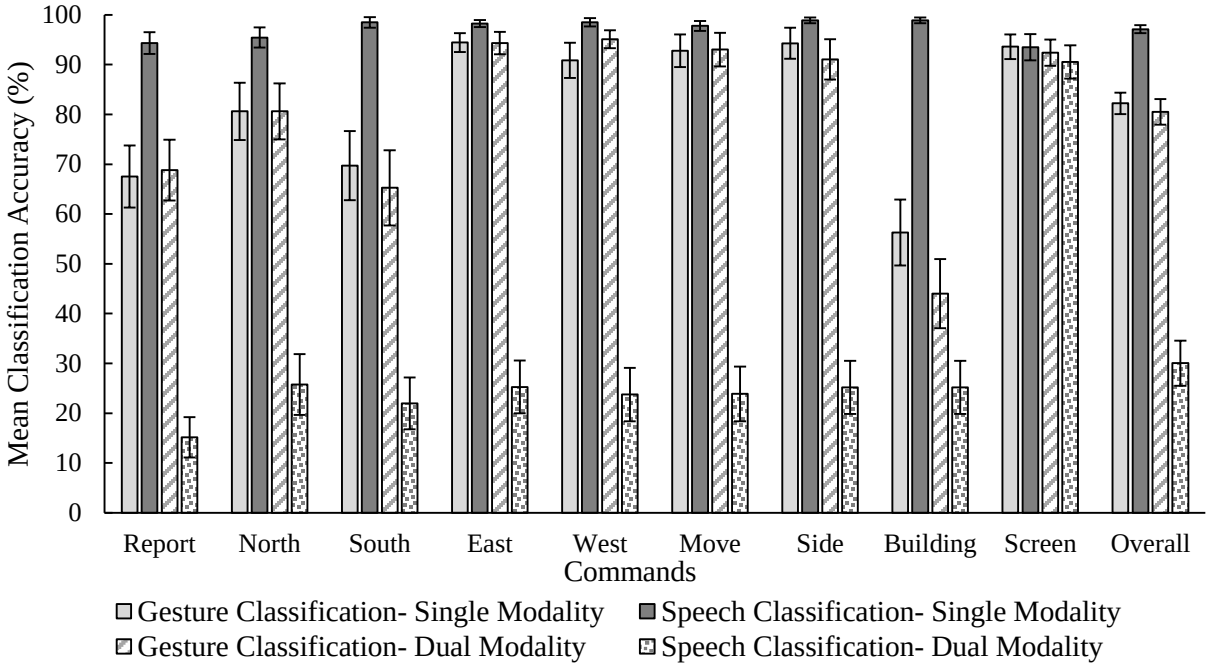

Figure 6. Mean classification accuracy for gesture and speech recognition across single and dual modalities. Error bars are included in the graph

The command and command components excluding the Screen command, reported significant interaction effects between modes (single and dual) and modality (gesture and speech) (Table 12). Gesture recognition showed mean classification accuracies of 82.2 % (single modality) and 80.5 % (dual modality) while the speech recognition showed 97.1 % (single modality) and 30 % (dual modality). Table 12 illustrates the interaction effects and significance of the nine command components for gesture and speech classification across both modalities.

Table 12. Interaction effects between modes (single or dual) and modality (gesture or speech). Command component, degrees of freedom, F-score, and partial-eta-squared are all reported.

Command Component	**D.F.**	***F***	η^2_p
Report	1, 32	246.47**	.89
North	1, 32	110.21**	.78
South	1, 32	83.21**	.72
East	1, 32	135.97**	.81
West	1, 32	141.56**	.82
Move	1, 32	183.48**	.85
Side	1, 32	166.04**	.84
Building	1, 32	79.93**	.71
Screen	1, 32	0.16*	.01
Overall	1, 32	198.6**	.86

Note: *$p < .01$, **$p < .001$

Speech classification across single and dual modalities showed significant differences while no significant differences were observed for the gesture classification across single and dual modalities. In order to show the trend in classification accuracies for the two different modes (single and dual) across the two communication modalities (gesture and speech), a box plot was used to illustrate the minimum, maximum, median, first and third quartiles (Figure 7).

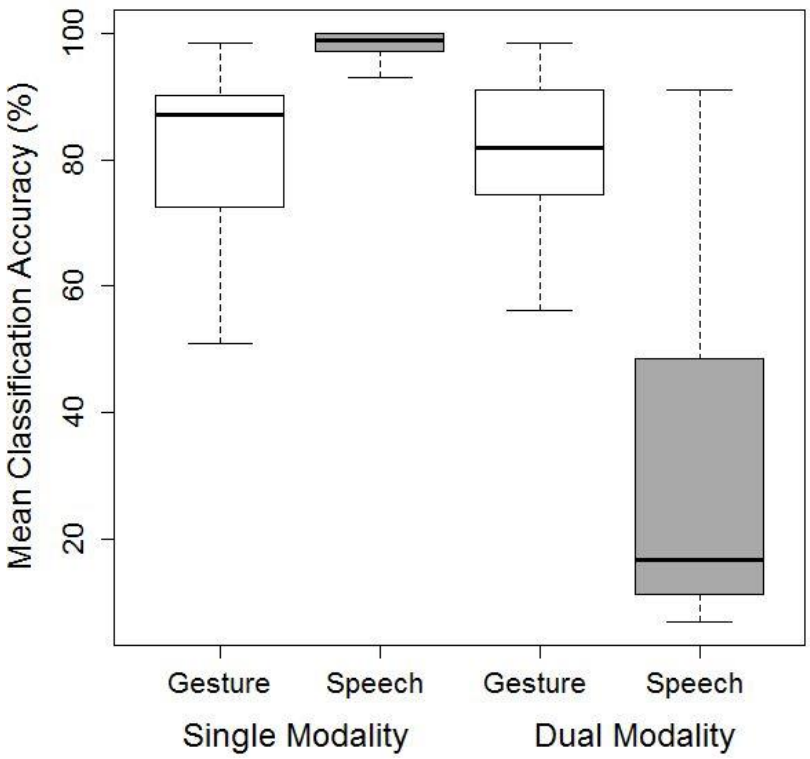

Figure 7. Mean classification accuracy for gesture and speech recognition across single and dual modalities showing minimum, maximum, median, first and third quartile values.

4. DISCUSSION

The goal for the current analyses was to evaluate the gesture and speech recognition classification technologies for commanding a robot teammate to perform spatial-navigation ISR missions. Three main research questions were addressed that evaluated if significant differences were found for gesture classification across single and dual modalities, speech detection and classification for different hardware across single and dual modalities, and gesture and speech classification across single and dual modalities. The results of these analyses clearly indicate the capability of gesture and speech recognition technologies in classifying spatial-navigation commands with a high level of accuracy.

The comparative analysis of gesture classification across single and dual modalities in section 3.1 showed that Gesture Builder across single modality (gesture only) had better performance overall, with a mean classification accuracy of 82.4 % compared to 80.8 % for dual modality (gesture recognition during simultaneous gesture and speech command delivery). The mean classification accuracy for the dual modality might have been due to the increased level of complexity of dual modality communication conveyance and the inability of the gesture classifier to concatenate long pauses in the gesture information while delivering gesture and speech commands simultaneously. Other contributing factors affecting the classification accuracies of gesture recognition were hardware issues in the gesture glove and absence of training sets from the participant for training the classifier to recognize gestures tailored to the wearer more accurately. Sometimes, the flex resistors in the glove fingers failed to detect and record the beginning and end of a gesture command which generated faulty gesture data from the glove and consequently resulted in misclassification of gestures. Future research should focus on developing a robust gesture glove by utilizing durable flex resistors in partially fingerless gloves. This reduces length of flex resistors in the glove fingers to eliminate unnecessary bending of the flex resistors. Additionally, utilizing training data from participants would enhance feature extraction and improve gesture recognition and classification accuracies across both modalities.

Results from the comparison of speech recognition across single and dual modalities in section 3.2 provide conclusive evidence revealing the ASR performed better for single modality (i.e. speech only) compared to dual modality (i.e. speech classification during simultaneous gesture and speech command delivery) with the classification accuracies of 76.1 % and

23.3 %, respectively. Speech recognition for all command components except the Screen command had mean classification accuracies in the range 70.7 - 79.5 % for the single modality compared to 12.7 - 16.7 % for the dual modality. The Screen command illustrated mean classification accuracies of 85.5% and 85.1% for single and dual modalities. The significant difference in the classification accuracies between the single and dual modalities was likely due to increased command duration for delivering speech commands during the dual modality which resulted in misclassification by the speech recognition software. The speech commands delivered during the dual modality were nearly two to three times longer in duration than the speech commands delivered during the single modality. The increased command duration while delivering commands in the dual modality (Table 7) was most likely due to the participants' attempting to obtain cues or prompts for the speech command from the corresponding gesture command being delivered simultaneously[37]. Another contributing factor to the lower classification accuracies was due to the Microsoft Speech Platform SDK's limited recognition range for speech command duration. Further, results of the comparison of speech recognition across both modalities for different hardware clearly illustrate that the Microsoft Kinect hardware performed better with 64.2% Overall command mean classification accuracy in comparison to 35.2% for the lapel microphone. Fidelity rate of the speech detection hardware plays a crucial role in digitizing the speech command delivered by the participant. The Microsoft Kinect with its four element microphone array has a higher reliability enabling better performance, compared to the lapel microphone. Figure 5 illustrates the interaction between hardware and modality in speech detection and recognition showing that the Kinect performed better across both modalities and could be a recommended choice for speech interface developers for Soldier-robot interaction.

The comparison of gesture and speech recognition across single and dual modalities in section 3.3 indicate that for single modality, ASR exhibited better performance than the gesture recognition with mean classification accuracies of 97.1% and 82.2%, respectively, and for dual modality, the gesture recognition exhibited better performance than the speech recognition with mean classification accuracies of 80.5 % and 30 %, respectively. These results were based on the comparison of classification accuracy of command components. The Screen command was detected and classified consistently by gesture and speech classifiers across both single and dual modalities with high classification accuracy ranging between 91.5 - 93.5 %. This was most likely due to the relatively lower cognitive burden and task load compared to other command components and also due to the fact that the Screen command was also a command component in itself (i.e. it was a single word command). The gesture recognition across both single and dual modalities showed no significant differences whereas speech recognition showed significant differences across single and dual modalities. It is possible that speech recognition during the dual modality was affected by 1) increased command duration for the delivery of speech commands while performing equivalent gesture sequences, 2) the increased task load on the participant and 3) the Microsoft Speech Platform SDK's inability to recognize and concatenate slowly delivered speech commands. Further research must be directed towards building gesture and speech classifiers that would account for the lag and lead in command durations in gesture and speech command delivery during dual modality functions to enable integration of gesture and speech interfaces for HRI that reflect HHI and facilitate more intuitive and natural, real time Soldier- robot communication.

5. CONCLUSION

The investigation of the three research questions facilitate evaluation of gesture and speech technologies for dismounted Soldier-robot dialogue. The results show evidently that both gesture and ASR technologies demonstrated classification accuracies ranging between 82 - 98% for single modality and 30 - 80% for dual modality during human to robot command delivery. However, gesture recognition showed more promising results with no significant differences in classification accuracies across both modalities. In addition, the Microsoft Kinect exhibited better performance and reliability in comparison to the lapel microphone across both modalities which would enable developers to design more robust HRI interfaces using the gesture glove and the Kinect, taking in to consideration, the complexity of gestures and speech commands, ambient noise, distortions in the communication channel, and environmental conditions during military operations. Finally, the integration of dual modality natural language interfaces for HRI present great potential and scope for improvement in classifying spatial-navigation commands using gesture and speech interfaces for enabling non-line-of-sight communication between Soldier and robot through the development of more robust gesture and speech interfaces, and classification models.

6. ACKNOWLEDGEMENTS

This research was sponsored by the U.S. Army Research Laboratory (ARL) and was accomplished under Cooperative Agreement Number W911NF-10-2-0016. The views and conclusions contained in this document are those of the author's and should not be interpreted as representing the official policies, either expressed or implied, of ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] Barber, D., Lackey, S., Reinerman-Jones, L. and Hudson, I., "Visual and tactile interfaces for bi-directional human robot communication," In SPIE Defense, Security, and Sensing. International Society for Optics and Photonics. pp. 87410U-87410U (2013).

[2] LaViola, J.J., "3d gestural interaction: The state of the field," International Scholarly Research Notices 2013, (2013).

[3] Hutchins, S., Cosenzo, K., Barnes, M., Feng, T., Pillalamarri, k., "Soldier-Robot Teaming: Effects of Multimodal Collaboration on Team Communication for Robot Reconnaissance," In ARL Technical report 5385, (2010).

[4] Barnes, M. and Jentsch, F., [Human-robot interactions in future military operations], Ashgate Publishing Company, (2010).

[5] Burke, D., Schurr, N., Ayers, J., Rousseau, J., Fertitta, J., Carlin, A. and Dumond, D., "Multimodal interaction for human-robot teams," In SPIE Defense, Security, and Sensing. International Society for Optics and Photonics. pp. 87410E-87410E (2013).

[6] Harris, J., & Barber, D., "Speech and gesture interfaces for squad-level human-robot teaming," In SPIE Defense and Security. International Society for Optics and Photonics. pp. 90840B-90840B (2014).

[7] Cockburn, J., Solomon, Y., Kapadia, M. and Badler, N., "Multi-modal human robot interaction in a simulation environment," Technical report, University of Pennsylvania (2013).

[8] Shah, J. and Breazeal, C., "An empirical analysis of team coordination behaviors and action planning with application to human–robot teaming," Human Factors: The Journal of the Human Factors and Ergonomics Society.Vol.52 (2), pp.234-245 (2010).

[9] Capstick, E., Pomranky, R., Dungrani, S. and Johnson, T., [Soldier Machine Interface for Vehicle Formations: Interface Design and an Approach Evaluation and Experimentation], Army Research Laboratory, (2010).

[10] Lackey, S., Barber, D., Reinerman, L., Badler, N.I. and Hudson, I., "Defining next-generation multi-modal communication in human robot interaction," In Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications. Vol. 55, No. 1, pp. 461-464 (2011).

[11] Microsoft, "Xbox One," 2016, http://www.xbox.com/en-US/xbox-one/innovation (9 March 2016).

[12] Apple Inc., "Siri," 2016, http://www.apple.com/ios/siri/ (9 March 2016).

[13] Google, "Research at Google," 2016, http://research.google.com/pubs/SpeechProcessing.html (9 March 2016)

[14] Zaykovskiy, D., "Survey of the speech recognition techniques for mobile devices," Proc. of DS Publications (2006).

[15] Jiang, H., "Confidence measures for speech recognition: A survey," Speech communication. Vol.45 (4), pp.455-470 (2005).

[16] Kaljurand, K. and Alumäe, T., "Controlled natural language in speech recognition based user interfaces," Springer Berlin Heidelberg. pp. 79-94 (2012).

[17] Welch, G., and Foxlin, E., "Motion tracking survey," IEEE Computer graphics and Applications. vol. 22. No.6. pp.24-38 (2002).

[18] LaViola, J., "A survey of hand posture and gesture recognition techniques and technology," Brown University, Providence, RI 29 (1999).

[19] Bowman, D., Kruijff, E., LaViola, J., Poupyrev, I., "An introduction to 3-D user interface design," Presence, vol. 10, no. 1, pp. 96–108 (2001).

[20] Kim, D., Hilliges, O., Izadi, S., Butler, A.D., Chen, J., Oikonomidis, I. and Olivier, P., "Digits: freehand 3D interactions anywhere using a wrist-worn gloveless sensor," In Proceedings of the 25th annual ACM symposium on User interface software and technology, ACM. pp. 167-176 (2012).

[21] Starner, T., Weaver, J. and Pentland, A., "Real-time American sign language recognition using desk and wearable computer based video," Pattern Analysis and Machine Intelligence, IEEE Transactions. Vol.20. pp. 1371-1375 (1998).

[22] Wang, R., Paris, S. and Popović, J.,"6D hands: markerless hand-tracking for computer aided design," In Proceedings of the 24th annual ACM symposium on User interface software and technology. ACM. pp. 549-558 (2011).

[23] Ellis, C., Masood, S.Z., Tappen, M.F., Laviola Jr, J.J. and Sukthankar, R., "Exploring the trade-off between accuracy and observational latency in action recognition," International Journal of Computer Vision. Vol.101.pp.420-436 (2013).
[24] Elmezain, M., Al-Hamadi, A., Sadek, S. and Michaelis, B., "Robust methods for hand gesture spotting and recognition using hidden markov models and conditional random fields," In Signal Processing and Information Technology (ISSPIT), 2010 IEEE International Symposium on. pp. 131-136 (2010).
[25] Huang, D.Y., Hu, W.C. and Chang, S.H., "Vision-based hand gesture recognition using PCA+ Gabor filters and SVM," In Intelligent Information Hiding and Multimedia Signal Processing. Fifth International Conference on. pp. 1-4 (2009).
[26] Zhang, X., Chen, X., Li, Y., Lantz, V., Wang, K. and Yang, J., "A framework for hand gesture recognition based on accelerometer and EMG sensors," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 41(6), pp.1064-1076 (2011).
[27] Fang, G., Gao, W. and Zhao, D., "Large vocabulary sign language recognition based on hierarchical decision trees," In Proceedings of the 5th international conference on Multimodal interfaces. ACM. pp. 125-131 (2003).
[28] Saponas, T.S., Tan, D.S., Morris, D., Balakrishnan, R., Turner, J. and Landay, J.A., "Enabling always-available input with muscle-computer interfaces," In Proceedings of the 22nd annual ACM symposium on User interface software and technology. ACM. pp. 167-176 (2009).
[29] Saponas, T.S., Tan, D.S., Morris, D., Turner, J. and Landay, J.A., "Making muscle-computer interfaces more practical," In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM. pp. 851-854 (2010).
[30] Harrison, C., Tan, D. and Morris, D., "Skinput: appropriating the body as an input surface," In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM. pp. 453-462 (2010).
[31] Harrison, C., Benko, H. and Wilson, A.D., "OmniTouch: wearable multitouch interaction everywhere," In Proceedings of the 24th annual ACM symposium on User interface software and technology. ACM. pp. 441-450 (2011).
[32] U.S. Army, "FM 21-60 Visual Signals," Washington D.C., (1987).
[33] DoD, U. S., "Department of Defense dictionary of military and associated terms," Joint Publication, pp.1-02 (2007).
[34] Costello, E., [Random House Webster's American Sign Language dictionary], Random House Reference, (1998).
[35] Barber, D., Wohleber, R. W., Parchment, A., Jentsch, F., and Elliott, L. "Development of a Squad Level Vocabulary for Human-Robot Interaction," In Virtual, Augmented and Mixed Reality. Springer International Publishing. pp. 139-148 (2014).
[36] Abich IV, J., Barber, D. J., & Reinerman-Jones, L. (2015). "Experimental Environments for Dismounted Human-Robot Multimodal Communications," Virtual, Augmented and Mixed Reality. Springer International Publishing. pp. 165-173 (2015).
[37] Abich, J., Barber, D.J., [Under review]. "The Impact of human-robot multimodal communication on mental workload, usability preference, and expectations of robot behavior," (2016).