Analysis of effective speech recognition in A Virtual Operating Room

Mark W. Scerbo
Levi Warvel
Samantha Zybak
Rebecca Kennedy
Amanda Ashdown
Kimberly Perry
Elizabeth Newlin-Canzone

Department of Psychology
Old Dominion University
Norfolk, VA
mscerbo@odu.edu, lwarvel@odu.edu, szybak@odu.edu

Yiannis E. Papelis
Menion Croll
Hector Garcia

Virginia Modeling Analysis & Simulation Center
Old Dominion University
Norfolk, VA
ypapelis@odu.edu, mcroll@odu.edu

ABSTRACT

Recently, we have developed a simulated environment for training surgical team members in judgement, decision-making, and technical ability. The Virtual Operating Room (VOR) is a fully immersive virtual environment that augments procedural task simulator training with a simulated OR context. The VOR is modeled on a standard OR and outfitted with both real and virtual equipment and a commercial medical simulator for laparoscopic cholecystectomy. Trainees communicate with virtual teammates using speech recognition software. One challenge with the VOR is developing a system for speech recognition that includes all reasonable surgical utterances while retaining flexibility between differences in user terminology. In the initial configuration, speech recognition rules were rigid and required operators to manually intervene for statements that went unrecognized. In the current configuration, we developed an extended state machine formalism that reduced the number of words needed to be detected in each utterance. The present paper compares previous results with the current version. Surgical residents were recruited to test system efficacy using free speech. Surgeons' perceptions about speech recognition with the current version were more positive compared to older version: levels of dissatisfaction with the vocal control declined from 20% to zero. Additionally, a third of the surgeons reported feeling as though the simulation was realistic due to the ability to communicate with the system via speech. However, some additional recognition issues were revealed, including lag in system response times, confusion regarding system inquiries and appropriate responses, the need to provide verbal feedback for certain procedural steps, and unexpected user responses. Preliminary results suggest that communication efficacy in the current VOR has improved, but issues such as system response time and speech recognition accuracy still pose a challenge.

ABOUT THE AUTHORS

Mark W. Scerbo, PhD, is Professor of human factors psychology at Old Dominion University. He leads a team of researchers and developers who study user interaction with medical simulation technology, develop new medical simulation models and technology, and investigate methods to integrate simulation into medical school curricula. Dr. Scerbo has over 30 years of experience researching and designing systems and displays that improve user performance in academic, military, and industrial work environments. He is a Fellow of the Human Factors and Ergonomics Society and the Society for Simulation in Healthcare and currently serves as Editor-in-Chief of the journal, *Simulation in Healthcare.*

Levi Warvel is a PhD student at Old Dominion University under the advisement of Dr. Mark Scerbo. His research interests include medical technology, simulation, mental workload, usability, augmented reality, and virtual environments.

Samantha Zybak is a MS student at ODU and a Human Factors Engineer for Addwin Consulting Corporation. She works on evaluating health IT software within the VHA for usability and patient safety concerns while providing recommendations for further evaluation and implementation.

Rebecca Kennedy is a User Experience researcher and consultant, and a Human Factors PhD candidate at ODU.

Amanda Ashdown, MS, is currently attending Old Dominion University in the Human Factors Psychology doctoral program. She works under the advisement of Dr. Mark Scerbo in the SURF lab studying simulation and user performance primarily in the area of healthcare. She is in her 4th year of graduate school and currently is involved in several projects at EVMS and Sentara. Her research interests focus on medical simulation, training, teamwork in healthcare, and usability of medical devices.

Kimberly Perry is a PhD student at Old Dominion University under the advisement of Dr. Mark Scerbo.

Elizabeth Newlin-Canzone, PhD is a Lead Human Factors Psychologist in the Center for Transforming Health at The MITRE Corporation. She received her doctorate in human factors psychology from Old Dominion University. She has significant research experience concerning safety in the aviation and healthcare domains and serves as co-lead on several projects with the National Patient Safety Partnership. This public-private partnership joined together leading pediatric hospitals with the goal to use predictive analytics on health data to better determine where patient safety events may occur. Liz currently uses her expertise to enhance the analysis of Veterans Administration (VA) health data to support the automated surveillance of the safety of Health Information Technology (HIT).

Yiannis Papelis is a Research Professor at the Virginia Modeling Analysis & Simulation Center, at Old Dominion University. Dr. Papelis obtained a Ph.D. from the University of Iowa, a Master's degree from Purdue and a BS degree from Southern Illinois University, all in Electrical & Computer Engineering. Dr. Papelis' research focuses on unmanned systems, autonomous robotics and use of virtual environments in a wide range of areas such as training, STEM education, design optimization and health-care. His research has been funded by numerous federal agencies as well as industry. Dr. Papelis is a member of the Society for Computer Modeling & Simulation and is currently serving as the Vice President of publications for the society.

Menion Croll has a BS in Computer Science from Virginia Tech and an MS in Computer Science from Old Dominion University. He started out developing combat systems and tactical display systems for NAVSEA. Since then he has integrated hardware control systems and virtual environments for NASA, developed a variety of serious games for training and education, and created applications for the web and mobile devices. His research interests include autonomous vehicles, virtual environments, and serious gaming.

Hector M. Garcia M.Arch. is a Senior Project Scientist at Old Dominion University's Virginia Modeling Analysis and Simulation Center, in the area of Visualization, Virtual Environments and Virtual Reality, integrating state of the art visualization systems with modeling and simulation applications, and the Scientist most closely involved with the CAVE (Cave Automatic Virtual Environment) on ODU's Norfolk Campus. Mr. Garcia received his Masters in Architecture from University of Houston in 1997. Mr. Garcia's expertise include the use of large scale visual simulation display systems, the use of tracking devices, robotics, and haptic devices used in training and education.
Mr. Garcia has more than 20 years' experience developing highly interactive Virtual Environments for training. He has been involved in a variety of research projects funded by NASA, NSF, ONR, AHRQ and private industry. Before joining Old Dominion University, Mr. Garcia spent 5 years as a Researcher at the University of Houston affiliated with the Virtual Environments Technology Laboratory working on several NASA projects for Astronaut training as well as NSF funded research for using Virtual Reality as a teaching tool.

Analysis of effective speech recognition in A Virtual Operating Room

Mark W. Scerbo
Levi Warvel
Samantha Zybak
Rebecca Kennedy
Amanda Ashdown
Kimberly Perry
Elizabeth Newlin-Canzone

Department of Psychology
Old Dominion University
Norfolk, VA
mscerbo@odu.edu, lwarvel@odu.edu, szybak@odu.edu

Yiannis E. Papelis
Menion Croll
Hector Garcia

Virginia Modeling Analysis & Simulation Center
Old Dominion University
Norfolk, VA
ypapelis@odu.edu, mcroll@odu.edu

INTRODUCTION

The need for simulator-based medical training has been well established over the last decade (Scerbo, 2005). Most of these simulators are designed to provide medical trainees with a safe and accessible alternative to treating real patients, allowing trainees to hone specific procedural skills without putting patients at risk. Simulators also allow residents a more flexible learning schedule that complies with the American Medical Association's 2003 restrictions on resident working hours (AMA, 2014). Although many simulators are advantageous to medical training programs, the focus on procedural skill development neglects other important areas of training that future doctors require, such as communication and critical thinking skills.

To address these types of skills, we developed the Virtual Operating Room (VOR). The VOR is a virtual immersive training environment that allows trainees to perform surgical skills within an operating room context and can be used to train and assess individual and team communication and decision-making (Baydogan, Belfore, Scerbo, & Saurav, 2009; Scerbo et al., 2006, 2007; Papelis et al., 2014). The VOR simulates an operating room through a combination of real and virtual equipment. Other surgical simulators are integrated into the simulation environment to represent the patient and target tissue areas and customized to provide real-time information about internal structures. Other members of the surgical team are represented with virtual agents and trainees communicate with them through a head-mounted microphone. The trainee has a number of actions available to them through vocal input, including responding to prompts, updating the team on progress, and inquiring about the patient's status.

Although the VOR represents a step toward a more immersive and ecologically representative surgical environment, the inclusion of additional situational context increases complexity and makes the design process more challenging. One of the more difficult challenges results from the system's reliance on speech recognition to drive much of scenario progress (Papelis et al., 2014). Scenarios are based on surgical procedures and, in general, follow the appropriate sequence of steps. However, detecting progress can be difficult because the only objective measurable changes in the simulated procedure are incisions and ligation of structures on the physical organ model. Otherwise, detecting procedural steps requires the trainee to verbally indicate when each has been completed.

These difficulties have been identified and addressed in the VOR design (Papelis, 2014). Using the Dragon Naturally Speaking client, a speech recognition application was developed to detect the presence of key words in user utterances. A semantic interpretation grammar model based on the Speech Recognition Grammar Specification (W3C, 2004) was converted to a Finite State Automata (FSA) to increase the flexibility of key phrase recognition. Additionally, FSA formalism was extended using NULL and ANY transitions, allowing the system to accurately interpret statements across a wider range of potential utterances. Although the extended FSA formalism did enhance the ability to interpret trainee utterances, the task analyses on which predicted key words were based underrated the variability of potential phrasing. As a result, trainees could utter contextually relevant statements and questions that the system could not recognize, impeding scenario progression and necessitating manual intervention.

To address this limitation, efforts were made to enhance speech recognition module functionality by further extending the FSA. Additional testing and interviews with subject matter experts revealed that most procedural steps originally identified in the task analyses could be further reduced to single words or phrases that would likely be used in a particular step in the surgical process. Further testing revealed that some words were consistently misinterpreted by the speech recognition system, despite being clearly and correctly spoken by the trainee. For example, the word “trocar” was consistently recognized as “trucker”, “true car”, or “choker”. Because these words were not relevant in the surgical context, it was possible to include them in the grammar rules to improve recognition. In addition to including consistently misinterpreted words, adding common phrases to the grammar rules also enhanced the speech recognition functionality. For example, a surgeon could ask for a trocar or they could confirm that the trocars have been placed. These were two separate steps in the scenario and triggered different responses. The grammar rules addressed the difference in these two phrases by adding words such as “can”, “may”, “please”, and “pass” to the rule asking for the trocar, and words such as “placed”, “completed”, and “finished” to the rule confirming that the trocars have been placed. Another enhancement to the speech recognition concerned multiple sentence structures for the same step. For example, when asking for a trocar, the surgeon could ask as a formal question (“Can you pass me the trocar?”) or they could use a command sentence (“Hand me the trocar.”). The final enhancement to the speech recognition included adding synonyms into the grammar rules. Surgeons may use different words for the same item, for example they may refer to a trocar as a port.

The objective of this study was to evaluate of the updated grammar rule structure. Toward this end, a cohort of surgeons was recruited and performed a simulated laparoscopic cholecystectomy. Their communications within the VOR were captured and compared with those obtained in the original version of the VOR (Baydogan, Belfore, Scerbo, & Saurav, 2009; Scerbo et al., 2006, 2007). It was expected that the FSA and additions to the speech recognition grammar rules would improve recognition accuracy and user satisfaction over the original version of the VOR.

METHOD

Participants

Sixteen surgical residents (6 women and 10 men) were recruited from Eastern Virginia Medical School (EVMS) to perform a laparoscopic cholecystectomy (gall bladder removal) in the VOR. Each participant was at least in their second year of residency school and had assisted in, performed, and/or observed a laparoscopic cholecystectomy. Each participant received $50 as compensation for their participant in the study.

Materials

The VOR was rendered in a C.A.V.E. on the EVMS campus. Each of the three C.A.V.E. walls featured one of the three virtual team members with whom participants could communicate: circulating nurse on the left, anesthetist on the right, and the attending surgeon directly in front of participant. The physical patient model was raised on an operative platform in near the front wall. Trocars and simulated laparoscope were placed prior to each participant’s session and remained relatively fixed. The body cavity contained a customized version of the Simulab, Inc. LapTrainer system (Seattle, WA). The LapTrainer system was composed of a hard plastic model of the stomach and liver bed upon which a soft rubber gallbladder was adhered. Each gallbladder was modified by the researchers to detect cuts of the main structures. Cutting the correct or incorrect structures would inform the state machine and advance the scenario. Except for the modified gallbladder, all system interaction was done verbally via a headset microphone.

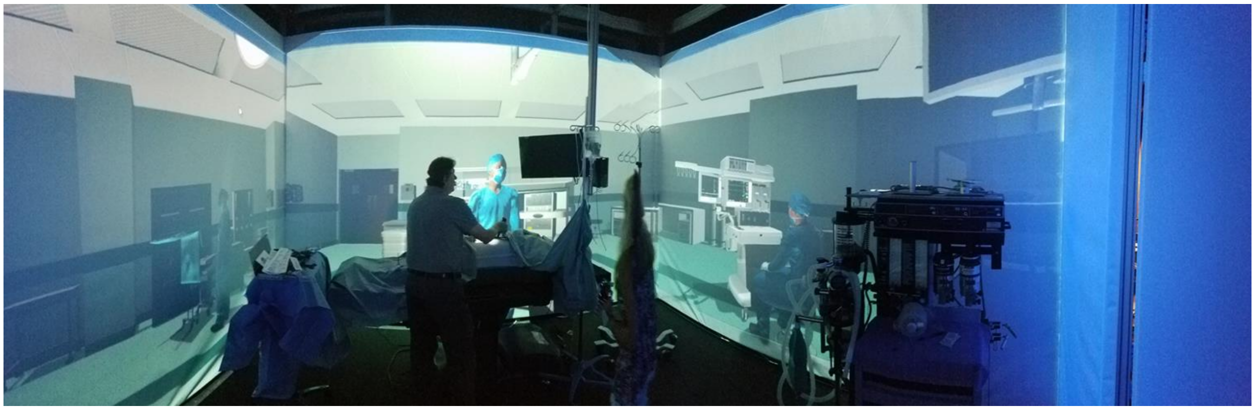

Figure 1. The Virtual Operating Room in EVMS C.A.V.E.

Procedure

First, participants were asked to fill out a demographic form. Next, each participant was asked to complete a simulated laparoscopic cholecystectomy scenario. Participants began the scenario with a brief timeout with the attending surgeon and would verbally indicate their current step after, either of their own volition as they naturally progressed or by answering yes or no prompts if a significant period of time had elapsed following the last step. No participant session exceeded an hour in duration. Following completion of the simulation, participants were given a post-experimental survey in which they rated aspects of the simulation's performance, identified the best and worst aspects of the system, and share their opinions and observations regarding their experience. Each was then thanked for their time and dismissed.

RESULTS

Accuracy of key word recognition was established by keeping track of the number of times the experimenters had to manually intervene to move the scenario along. The total number of manual transitions was compared to those observed with the original VOR in a 2008 pilot study. The level of user satisfaction was obtained from participant survey data.

Results suggested that the new modifications to the speech recognition grammar rules did indeed improve VOR performance over the earlier system. In the present study, the mean number of manual transitions required was 2.75 compared to a mean of 3.9 found in earlier testing. Participants largely indicated that the virtual team members in the VOR sounded and responded like real team members (M = 5.4 out of 7). Thirty-three percent of participants also indicated that open communication with the virtual team members and their ability to respond naturally was the most valuable aspect of the simulation. No participants identified the speech recognition system as the worst part of the simulation in the current build, representing a significant subjective improvement in performance compared to the 20% of participants found previously.

Although subjective outcomes were favorable, the speech recognition system still demonstrated some flaws. Twenty-five percent of participants found the worst aspect of system performance to be the lag between speech recognition and the response of virtual agents. Many participants indicated that they were uncertain if the system had interpreted their speech at all or if it was just processing a response. The lag between verbal input and response often lasted a few seconds which often led participants to repeat themselves. Additionally, participants felt that some of the responses were inappropriate for the question asked or lacked enough information relative to the question they asked.

Two of the six primary steps within the simulation, the trocar placement and lights out steps, were found to be largely ignored by participants and required manual transitions. The timeout step responses demonstrated the greatest degree of accuracy at 75%, likely due to being a simple confirmation response. All responses that failed to trigger the next step automatically were due to inadequate verbalization (IV) on the part of the participant. The body cavity bleed check step was largely missed with only 16.7% of participants responding affirmatively to the prompt. The speech recognition system failed to recognize most responses due to unexpected language, with a general tendency towards participant confusion over what aspect of the cavity they were supposed to be checking. Despite task analyses indicating that checking the abdominal cavity for bleeding was an important part of the surgical process, participants responded to the prompt with statements such as “Check the cavity for what?” and “checking the cavity is not actually a surgical step”. This trend toward participant confusion may be a result of vague prompt phrasing (“Have you checked the cavity?”).

The speech recognition system seemed to have the most difficult processing utterances related to the duct and artery clip step of the procedure. A third of participant’s responses were interpreted correctly, with three being negative responses to the prompting question and one a formal statement that the clips were in place. The remaining 66.7% of responses failed to be automatically processed for a variety of reasons: two utterances adhered to existing grammar rules but the system failed to recognize the words (e.g., “nope”), three responses were inadequately verbalized to be recognized (e.g., word spoken too quickly in succession or too quietly), two responses were phrased in unexpected ways (e.g., “almost” instead of “yes” or “no”) , and one response being incorrectly interpreted as a statement of completion rather than a request for materials (i.e., “click” included as a grammar rule for “clips” with “clip” being sufficient to engage transition from duct and artery clip step).

DISCUSSION

The overall performance of the speech recognition system demonstrated a general improvement over the earlier version of the VOR. The number of manual transitions necessary in the 2008 version of the system was greater than those needed by the newer version. Additionally, subjective data indicated that participants found the interactive vocal features of the VOR to be the most valuable aspects of the simulation, while also reporting significantly less dissatisfaction comparable to the 2008 cohort. However, the system also demonstrated some downfalls. Unexpected utterances, inadequately annunciated statements, and a general lack of responses to low-stakes procedural steps accounted for most speech recognition failures. Lag time between participant input and system response was also significant enough to cause participants to doubt that the system had indeed heard them. Some of the attending surgeon’s scripted statements also appeared to be vague, confusing participants and introducing utterance variability.

Many of these issues may not be overly difficult to address. In the case of the body cavity bleed check step, most participants responded with confusion but in fairly consistent ways. Participants frequently asked for clarification (e.g., “I’m sorry?”, “what do you mean?”, or “check the cavity for what?”). As such, one solution would be to add a clarification rule to the cavity check step. A simple grammar rule checking for any utterance for clarification would allow the system to clarify the intended goal of the step and quickly lead users back to the intended task flow. Two steps in the current scenario/procedure (trocar placement and lights out) could be effectively eliminated all together by removing the need for the user to interact with the system to drive the events. Instead, these events could be programmed to occur automatically at the very start of the procedure, or after a timed interval. Although task analyses suggested that these two steps were important, neither seemed to have a significant effect on participant outcomes. Another simple adaptation may be to extend the speech recognition requirement for the steps that require clipping the cystic duct and artery. Currently, the grammar rule only requires a single word to be recognized (either “clipped”, “clipping”, “clip”) to exit the step and begin the next. Although the grammar rule was designed to address observations regarding system performance during this step, it may be too generic. Any utterance about clips, be it a statement informing the attending surgeon that placement is complete or a simple request for the clip applier, is sufficient to enact the rule. Changing the rule to require an additional common word, such as “placed” or “finished”, would likely solve this problem.

In addition to modifications suggested by speech recognition errors, participants also demonstrated a few previously unpredicted utterances not directly relating to VOR progress. Many of these utterances indicated that residents thought differently about task progress than the subject matter expects who informed our initial design. Whereas experts tended to define each step as complete or incomplete, residents recognized an intermediary state in task progress. One example came as a response to a binary question, “Have you finished clipping the cystic duct and artery?”. Rather than yes or no, the participant responded with “almost”. This may be a valid response to the question. Indeed, “almost” conveys both the current state of the surgeon’s progress within the step as well as a projection of the near future state. Currently, the attending surgeon agent is programmed to repeat any inquiry that is ignored or answered with a “no” after 30 seconds. The inclusion of grammar rules that reflect estimations of a future state might allow for more adaptable intervals to be programmed between repeat inquires. Answering “almost” could result in a shortened elapsed interval to the next repetition relative to a “no” response which could then be given a longer interval.

Another tendency of participants was to ask questions about the equipment, procedure, or patient in general. Some examples included requests to focus the screen on the laparoscopic tower, questions regarding how much control one had over the laparoscope, updates on the patient's vitals, and clarifications regarding the precise way the procedure should be done within the simulated context or when the simulated procedure could be considered complete. While none of these communications drive the scenario forward, they may represent unique opportunities to improve immersive experience of the VOR. Adding simple responses to these questions may increase user confidence that the system understands their communications. Also, inquiries surrounding patient status (vital signs) represent an opportunity to predict or anticipate future states. Being able to get vital signs before a critical event may give users an opportunity to increase their awareness of the patient's status and increase their ability to recognize potential adverse states.

CONCLUSION

The development of effective speech recognition within the VOR is highly dependent on effective grammar rules. The extended grammar rules improved system performance when compared to earlier results, reducing the need to manually intervene and transition between states and increased participant satisfaction with speech recognition. However, lag times between user utterances and cued responses are problematic, introducing opportunities for confusion poor perception of system performance. Similarly, the variety of phrases generated by participants underscores the importance of establishing a more comprehensive set of possible statements and inquiries. Certain procedural steps within the VOR may also need to be revisited for relevance, such as verbalizing trocar placement or asking for the lights to be turned off. For the VOR to adequately represent the operating room theater, the speech recognition system must be able to respond in a more expeditious and intuitive manner to a wider range of probable surgeon utterances. Although the new speech recognition modifications have indeed enhanced recognition accuracy and user satisfaction relative to earlier results as predicted, it is clear that some challenges remain that must be addressed before the VOR can be considered a suitable alternative experience for a genuine operating room.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the guidance provided by Dr. Rebecca Britt from Eastern Virginia Medical School and the Virginia Modeling, Analysis, & Simulation Center for supporting the project.

REFERENCES

American Medical Association (2014). *AMA duty hours policy*. Retrieved January 18th, 2017, from https://www.ama-assn.org/sites/default/files/media-browser/public/about-ama/councils/Council%20Reports/council-on-medical-education/cme-rpt5-a-14.pdf

Baydogan, E., Belfore, L. A., Scerbo, M.W., & Saurav, M. (2009). Virtual operating room team training via computer-based agents. *International Journal of Intelligent Control and Systems*, *14*, 115-122.

Papelis, Y.E., Croll, M., Garcia, H, Scerbo, M.W., & Kennedy, R. (2014). *Behavior authoring and run-time management of computer agents for a virtual operating room training environment.* MODSIM World 2014, Paper No. 1454, (pp. 1-7). Hampton, VA: MODSIM World.

Scerbo, M.W., Bliss, J.P., Schmidt, E.A., Hanner-Bailey, H., & Weireter, L.J. (2005). Assessing surgical skill training under hazardous conditions in a virtual environment. In J.D. Westwood et al. (Eds.), *Medicine meets virtual reality*, 13, (pp. 436-442). Amsterdam: IOS Press.

Scerbo, M.W., Belfore, L. A., Garcia, H. M., & Weireter, L. J., Jackson, M., Nalu, A., & Baydogan, E. (2006). *The virtual operating room*. Proceedings of the Interservice/Industry Training, Simulation and Education Conference, Paper No. 2711, (pp. 1-9). Arlington, VA: National Training and Simulation Association.

Scerbo, M.W., Belfore, L.A., Garcia, H.M., Weireter, L.J., Jackson, M.W., Nalu, A., Baydogan, E., Bliss, J.P., & Seevinck, J. (2007). *A Virtual operating room for context-relevant training.* Proceedings of the Human Factors & Ergonomics Society 51st Annual Meeting (pp. 507-511). Santa Monica, CA: Human Factors & Ergonomics Society.

W3C (2004). *Speech Recognition Grammer specification Version 1.0*, Retrieved March 2014 from: www.w3.org/TR/speech-grammar.