Evaluating Multimodal Interaction of Robots Assisting Older Adults

Afagh Mehri Shervedani[1], Ki-Hwan Oh[1], Bahareh Abbasi[2], Natawut Monaikul[3], Zhanibek Rysbek[1], Barbara Di Eugenio[3], and Miloš Žefran[1]

***Abstract*— We outline our work on evaluating robots that assist older adults by engaging with them through multiple modalities that include physical interaction. Our thesis is that to increase the effectiveness of assistive robots: 1) robots need to understand and effect multimodal actions, 2) robots should not only react to the human, they need to take the initiative and lead the task when it is necessary. We start by briefly introducing our proposed framework for multimodal interaction and then describe two different experiments with the actual robots. In the first experiment, a Baxter robot helps a human find and locate an object using the Multimodal Interaction Manager (MIM) framework. In the second experiment, a NAO robot is used in the same task, however, the roles of the robot and the human are reversed. We discuss the evaluation methods that were used in these experiments, including different metrics employed to characterize the performance of the robot in each case. We conclude by providing our perspective on the challenges and opportunities for the evaluation of assistive robots for older adults in realistic settings.**

I. Introduction

With the world population rapidly aging, assistive robots promise to ease the societal burden of care for older adults. The primary focus of care for older adults is on the Activities of Daily Living (ADLs) so that they can continue to live independently, but companionship and socio-emotional support are also important. Increasingly it has been also recognized that helping caregivers may be as important as helping older adults directly.

Evaluation is a critical step in the deployment of assistive robots. Several types of evaluations are typically needed (see also [1], [2]):

- Functionality: does the technology work as intended? For example, does a human action recognition module reach a certain F score?
- Usability: is the user experience while interacting with the technology satisfactory? For example, can the user interact with the robot using unrestricted instructions, or are they limited to a set of keywords?
- Effectiveness: does the technology achieve the stated goal? For example, do older people using an assistive robot manage to stay healthier than those that don't?

[1]A.M. Shervedani, Z. Rysbek, K.H. Oh, and Miloš Žefran are with the Robotics Lab, Electrical and Computer Engineering Department, University of Illinois Chicago, Chicago, IL 60607 USA.

[2]B. Abbasi is with the Computer Science Department, California State University Channel Islands, Camarillo, CA 93012 USA.

[3]N. Monaikul and B. Di Eugenio are with the Natural Language Processing Lab, Computer Science Department, University of Illinois Chicago, Chicago, IL 60607 USA.

This work has been supported by the National Science Foundation grants IIS-1705058 and CMMI-1762924.

These different types of evaluation increase in complexity, with usability assessment requiring more complex studies than functionality assessment, and effectiveness assessment being significantly more demanding than usability assessment. This is especially true for applications of assistive robots in elderly care, and healthcare in general, where there are many challenges with recruiting subjects, the ability of technology to work in real-life settings, and the length of time needed to assess the health outcomes.

We focus this paper on our experiences with the evaluation of a multimodal interaction manager developed for assistive robots for older adults. The interactions during various activities of daily living (ADLs) between the human and the robot are expected to be inherently multimodal, such as force exchanges, pointing gestures, haptic-ostensive (H-O) actions, and speech. This is also confirmed by our previously collected corpus of human-human interactions between elderly individuals (elder role: ELD) and nursing students (helper role: HEL) assisting in ADLs [3]. Motivated by this, we proposed a Multimodal Interaction Manager (MIM) [4] that allows an assistive robot to process the actions of the human, generate appropriate responses, and make progress toward completing the task. The MIM is described in detail in Sec. III. The overview of the implementation and results of our experiments are provided in Sec. IV.

During the interaction, the robot may not be able to correctly translate speech to language, and determine other human actions from the readings of the sensors. For instance, the utterance "cup" might be misunderstood as "cop" which is not an object in the robot's data and results in the robot not being able to complete the task. Furthermore, the gestures of humans, for example pointing to the location, might not be recognized or the pointing direction may not be determined correctly.

Even though the robot may correctly interpret human actions, the robot may fail to correctly respond. For example, the robot may not have a complete representation of the task and could fail to determine what an appropriate response is to a particular human action. In such cases, the robot has to ask the participant to repeat the action until it finds a match in the model it uses for planning. The longer this interaction becomes, the less the user will expect from an assistive robot. We called these as *non-understandings* based on the definition in [5], and thus measuring the rate of unsuccessful attempts is important for evaluating the robot. More details about the evaluations and the discussion are in Sec. V and Sec. VI, respectively.

II. Related Work

As assistive robots get more popular and society tends to utilize them more,the ground metrics become more critical for evaluation. Considering various aspects and applications of different SARs, different evaluation methods should be employed too.

In [6]–[8], it is shown that the user's experience could be refined by adding non-linguistic modalities to robots. In [6], the implementation is evaluated on the basis of human participants' answers to the questionnaire covering different metrics. In [7], particular metrics are ruled out based on the human participant's words and reactions; and the video recordings of the interaction are analyzed. They also provide the participants with a questionnaire and evaluate the interactions they had with the robot based on their responses to the questionnaire. In [8], the authors evaluate their approach by implementing it on a robot and measuring the length of the interaction.

The framework proposed in [6] is evaluated by theoretically analyzing the underlying model before implementing it on a robot. In [9], [10], Hierarchical Task Networks (HTNs) are introduced and evaluated theoretically.

In [11], a middleware system, DiscoRT, is developed and implemented to improve the performance of virtual and robotic conversational agents. The system is evaluated by running experiments on each virtual and robot agent and investigating the conceptual aspects of the experiments. To analyze the performance of their Interactive Hierarchical Task Learning Algorithms, the authors in [12] run simulations where human subjects interact with the simulated robot environment through their graphical user interface. They extract and measure some objectives associated with the task for evaluation. Similarly, in [13] the authors also run simulations with specific metrics for evaluating their Hierarchical Distributed Dialogue Architecture.

In [14], the proposed Hierarchical Deep Reinforcement Learning framework is evaluated by reporting the success rate, the average number of turns between the user and the agent, and the reward from the simulation experiments. In [15], the Distributed Play-based Role Assignment Algorithm is developed and tested by implementing it on a distributed team of robots for the RoboCup four-legged league. The task-completion time is used as an evaluation metric.

In the literature, however, we observe a lack of separate methodologies for evaluating the theoretical framework and implementation. In this work, we propose adopting different evaluation methods for different aspects of ASRs.

III. Multimodal Interaction Manager Framework

As shown in Fig 1, the Multimodal Interaction Manager (MIM) consists of three components: (a) the interpretation module, which interprets multimodal actions of the human observed by the robot; (b) the mediation module, which determines the action of the robot in response to the human; and (c) the execution module, which executes the action

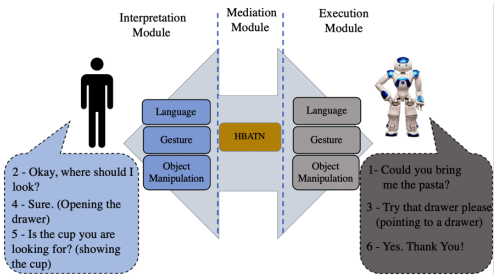

Fig. 1. The architecture for multimodal human-robot interaction. The figure is taken from [16].

of the robot. The task that was studied in detail in our work was the *Find* task, an interaction scenario in which a human and a robot work together to find an object in the environment. The core of our framework is Hierarchical Bipartite Action-Transition Networks (HBATNs) that model both agents simultaneously to maintain the state of a task-driven multimodal interaction and plan subsequent robot moves.

The ELDERLY-AT-HOME corpus [17], a publicly available corpus of human-human multimodal interactions, involves performing assisted ADLs, such as putting on shoes and preparing dinner. We developed the framework of HBATN on a subcorpus consisting of the interactions related to the *Find* task. That is, the elderly participant (ELD) would ask for an object, and the helper (HEL) would try to find it by asking follow-up questions.

The *Find* task can be decomposed into a set of subtasks to identify two main unknowns: the target object (O) and its location (L). The four main subtasks are determining the desired object type (*Det*(O_T)), determining a potential location to check (*Det*(L)), opening the location (*Open*(L)), and determining the actual object (*Det*(O)). These are modeled as Action-Transition Networks (AcTNets).

The AcTNet is a bipartite graph representing the states of both participants and their possible multimodal actions, which are defined as vectors consisting of linguistic features (the *dialogue act* (DA) [17] of the utterance and object or location words) and physical features (pointing gestures or *haptic-ostensive* (H-O) actions). The HBATN encompass these AcTNets allowing a robot to not only infer the state of its partner but also to plan its next action accordingly.

Subsequently, we generalized our model to enable the robot to be either the ELD or HEL by decomposing the subtasks into what we call *primitive subtasks*. In this new formulation, *Det*(O_T) and *Det*(L) establishes the object type and its location (*Estab*), potentially followed up by verification (*Verify*) or questions specifying for more information (*Spec*), and *Det*(O) confirms the presence or absence of the desired object (*Finish*) in the current location or verify a physical object with the partner.

Subsequently, a classifier was developed to allow the robot to determine what is the primitive subtask that the interaction is currently in. The proposed classifier automatically annotates multimodal interaction data for our primitive subtasks. This can be used in turn to learn the topologies of

each subtask by extracting sequences of moves belonging to the same subtask and using well-established techniques for learning like Markov models. It is implemented on the MIM to infer the state of the human partner by comparing the observed human action with all possible actions in the HBATN with preference given to those in the predicted subtask. A comprehensive demonstration of the interaction can be found in [16].

IV. Implementation and Experiments

To test the full MIM depicted in Fig. 1, we implemented several components in both experiments to recognize and understand multimodal human actions (the Interpretation Module), and to generate robot actions (the Execution Module). In this section, we focus on the results and the evaluations.

A. Robot as the HEL

To evaluate the performance of our initial Multimodal Interaction Manager, the framework was implemented on Baxter Robot from Rethink Robotics. In this experiment, Baxter participated as HEL and human participants acted as ELD in the *Find* task. The human participant would give instructions to Baxter to guide it through the find task while Baxter robot would ask questions regarding the object and its location.

In this experiment, the ELD (human) rarely performs H-O actions. As a result, the Interpretation Module focuses only on interpreting human speech and gestures; and this is composed of three primary parts: a speech-to-text component, a pointing gesture recognition component, and the *Dialogue Processing & Modality Fusion* component for processing the utterance and gesture together.

The HEL (Baxter) should perform pointing actions as well as H-O actions. Baxter can also communicate to the ELD through generated speech. Therefore, the Execution module in this experiment performs H-O actions, pointing actions, and speech; and is composed of multiple subcomponents: pointing, H-O action, speech generation, and object recognition component that enables the robot to determine the object and its location. The output of the Mediation Module is an action vector defining the robot's next move; each execution component uses this vector to perform its respective action.

B. Robot as the ELD

To evaluate the feasibility of switching roles, we implemented the MIM with the refined HBATN. In this experiment, the NAO robot acted as the ELD and human participants acted as HEL in the interaction. Utilizing a different robot in this experiment confirms the fact that our framework is platform-independent.

In this experiment, the HEL (human) performs H-O and pointing actions as well as communications through speech. As a result, our Interpretation Module has to interpret not only the human's speech and pointing gestures but also their H-O actions. The Interpretation Module is composed of the followings: pointing gesture recognition, H-O action recognition, a speech-to-text component, an object recognition component, and most importantly the *Dialogue Processing & Modality Fusion* component. The last component performs DA and subtask classification, combines the results creating an input action vector, and transfers it to the Mediation Module.

The ELD (NAO) only needs to perform pointing gestures and speech. Thus, the Execution Module only takes care of the pointing gesture and speech, and thus it contains two components: pointing gesture execution and speech generation.

V. User Study and Evaluations

A. Experiment on Baxter

In a preliminary user study, seven participants were recruited to interact with Baxter. Each subject performed 4 trials to find one of the four objects, giving a total of 28 trials of *Find* task. Baxter would help the participants locate the object by talking and pointing. No script was provided to the participants.

The evaluation has been done by reporting: 1) Average length of the interactions as the mean duration and the mean number of moves; 2) The percentage of successful trials, (trials in which Baxter continued the interaction when the object was already found are counted as failed trials); 3) *Non-understanding* percentage of turns (throughout all trials) where the Baxter's interpretation of the human action can not be found in HBATN and Baxter needs to ask the participant to repeat their action; 4) The word error rate (WER) [18] and serious speech recognition errors (SSREs) [4] for checking the speech recognition component of the Interpretation Module; 5) The percentage of wrong pointing gestures that were either not recognized or not tagged with the correct intended location for gesture recognition component; 6) The percentage of wrong classified DAs compared with the manually-labeled DA tags; 7) Overall quality of the interaction by asking the participants to rate their experience on a 5-point Likert scale [4].

B. Experiment on NAO

In another preliminary user study, six participants were recruited to interact with NAO. Each subject performed 5 to 6 trials with a total of 28 *Find* task trials. Participants would help NAO find the object it had in mind from a specific location.

Similar to the previous experiment, various metrics are reported for the evaluation: 1) The percentage of successful trials; 2) *Non-understanding* percentage of human turns (throughout all trials) in two categories where NAO asks the participant to repeat their action, and NAO fails to answer the participant's question or to follow their instruction; 3) The percentage of non-understandings in which various components make mistakes, particularly if an action is not accounted in the HBATN; 4) The speech-to-text (STT) accuracy [16]; 5) The accuracy of DA classifier; 6) The accuracy of H-O action recognition; 6) The Accuracy of pointing gesture recognition; 7) The accuracy of subtask classifier.

Avg. Duration	Avg. # Turns	Successful Trials	Non-Understandings	WER	SSREs	Wrong Pointing	SSRE & Wrong Pointing	Wrong DAs	Avg. User Rating
1m 45s	15.6	85.7%	11.7%	16.3%	23.4%	28.9%	1.2%	11.1%	4

TABLE I

PERFORMANCE RESULTS OF THE MIM ON THE *Find* TASK WITH BAXTER AS THE HEL. THE TABLE IS TAKEN FROM [4].

Avg. # Moves	Successful Trials	Non-Understandings	STT Accuracy	DA Accuracy	H-O Accuracy	Pointing Accuracy	Subtask Accuracy
19	84.8%	32.6%	84.8%	57%	83.1%	96%	49.3%

TABLE II

PERFORMANCE RESULTS OF THE MIM ON THE *Find* TASK WITH NAO AS THE ELD. THE TABLE IS TAKEN FROM [16].

DA Failure	Speech Failure	H-O Failure	Pointing Failure	Subtask Failure	DA & Subtask Failure	Model Failure
55.5%	43.4%	22.2%	2%	92.9%	51.5%	14.1%

TABLE III

PERCENTAGE OF NON-UNDERSTANDINGS IN WHICH ERRORS IN EACH COMPONENT OCCUR. THE TABLE IS TAKEN FROM [16].

VI. DISCUSSION

For a detailed discussion of the results of the experiments, the reader is referred to [4], [16]. The focus of this paper is on how our evaluation methods could practically be utilized in evaluating assistive robots.

One vital aspect of an interaction between a human and a robot is the duration of the interaction (a type of usability). As we pointed out before, one important application for assistive robots is to help older adults. The longer one interaction takes the more frustrated the human becomes. It also could highly affect the efficiency of task completion because if the older adult gets tired they are less interested in engaging in the task. Based on our experiment, we can declare that the average duration of the interactions is less than the frustration threshold of humans. Moreover, the average user Likert ratings of 4 out of 5 greatly supports our argument. However, most of the subjects were young people, and the ratings may decrease when more elderly subjects are involved. Since our studies remain at a more theoretical level and are not immediately relevant for applications in the real world, evaluation with older adults remains part of our future work.

An alternative aspect of usability is the overall success rate of the interactions between humans and robots. Humans tend to expect an assistive robot to act similarly to a human. Since humans are extremely adept at completing interactions, they expect similar performance from a robot. The success rates reported in Tables I and II show that our proposed framework achieves very good performance in this regard too.

Evaluating the components of a system can reveal many hidden issues. Each sub-system is closely connected to its adjacent components and it is likely that an error made by one component dramatically affects the overall performance of the robot and the success of the interaction. Evaluating components thus provides important insight for the robotic community by identifying possible failure points and establishing the relative importance of different components. For instance, the detailed results of sub-component accuracy/ failure we reported in Tables I, II, and III explain unsuccessful trials and non-understanding turns [4], [16]. In particular, DA classifier accuracy of 57% contributes to a non-understanding rate of 32.6% in the NAO experiment.

The discussion above raises the question of which is more important: the theoretical framework itself or the implementation? Will a superior implementation of a mediocre framework outperform a mediocre implementation of a superior framework? Clearly, the theoretical framework needs to be implemented on a real robot, or at least in a simulation, to be evaluated. However, the limits on time and resources frequently prevent researchers from spending sufficient effort on implementation.

Finally, while evaluation in real-world applications is clearly the ultimate test, it is becoming increasingly common to evaluate assistive robots in simulations. How to properly interpret the simulation results and reduce the cost of real-world evaluation remains an important topic for future research.

VII. CONCLUSION

In this work, we summarized our previous studies of assistive robots capable of multimodal interaction and described in detail the metrics used for their evaluation. These metrics should be of general interest and we hope that our insights can benefit other researchers in the area of assistive robots. We provided the motivation for using various metrics and showed that defining metrics tailored to evaluation of different components of the overall system can help explain its overall performance.

REFERENCES

[1] K. M. Tsui, D. J. Feil-Seifer, M. J. Matarić, and H. A. Yanco, "Performance Evaluation Methods for Assistive Robotic Technology," in *Performance Evaluation and Benchmarking of Intelligent Systems*, R. Madhavan, E. Tunstel, and E. Messina, Eds. Boston, MA: Springer US, 2009, pp. 41–66.

[2] M. Jung, M. J. S. Lazaro, and M. H. Yun, "Evaluation of Methodologies and Measures on the Usability of Social Robots: A Systematic Review," *Applied Sciences*, vol. 11, no. 4, p. 1388, Jan. 2021.

[3] L. Chen, M. Javaid, B. Di Eugenio, and M. Žefran, "The roles and recognition of haptic-ostensive actions in collaborative multimodal human–human dialogues," *Computer Speech & Language*, vol. 34, no. 1, pp. 201–231, 2015.

[4] B. Abbasi, N. Monaikul, Z. Rysbek, B. Di Eugenio, and M. Žefran, "A multimodal human-robot interaction manager for assistive robots," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6756–6762.

[5] G. Hirst, S. McRoy, P. Heeman, P. Edmonds, and D. Horton, "Repairing conversational misunderstandings and non-understandings," *Speech Communication*, vol. 15, no. 3-4, pp. 213–229, 1994.

[6] C.-M. Huang and B. Mutlu, "Learning-based modeling of multimodal behaviors for humanlike robots," in *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2014, pp. 57–64.

[7] J. K. Lee and C. Breazeal, "Human social response toward humanoid robot's head and facial features," in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 4237–4242. [Online]. Available: https://doi.org/10.1145/1753846.1754132

[8] J. Hemminahaus and S. Kopp, "Towards adaptive social behavior generation for assistive robots using reinforcement learning," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, 2017, pp. 332–340.

[9] K. Erol, J. Hendler, and D. S. Nau, "HTN planning: Complexity and expressivity," in *AAAI*, vol. 94, 1994, pp. 1123–1128.

[10] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[11] B. Nooraei, C. Rich, and C. L. Sidner, "A real-time architecture for embodied conversational agents: beyond turn-taking," *ACHI*, vol. 14, pp. 381–388, 2014.

[12] A. Mohseni-Kabir, C. Rich, S. Chernova, C. L. Sidner, and D. Miller, "Interactive hierarchical task learning from a single demonstration," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 205–212.

[13] M. Gašić, D. Kim, P. Tsiakoulis, and S. Young, "Distributed dialogue policies for multi-domain statistical dialogue management," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5371–5375.

[14] B. Peng, X. Li, L. Li, J. Gao, A. Celikyilmaz, S. Lee, and K.-F. Wong, "Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning," *arXiv preprint arXiv:1704.03084*, 2017.

[15] C. McMillen and M. Veloso, "Distributed, play-based role assignment for robot teams in dynamic environments," in *Distributed Autonomous Robotic Systems 7*. Springer, 2006, pp. 145–154.

[16] N. Monaikul, B. Abbasi, Z. Rysbek, B. Di Eugenio, and M. Žefran, "Role switching in task-oriented multimodal human-robot collaboration," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 1150–1156.

[17] L. Chen, M. Javaid, B. Di Eugenio, and M. Žefran, "The roles and recognition of haptic-ostensive actions in collaborative multimodal human–human dialogues," *Computer Speech & Language*, vol. 34, no. 1, pp. 201–231, 2015.

[18] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Prentice Hall, 2008, vol. 2.