Human Factors in Robots, Drones and Unmanned Systems, Vol. 179, 2025, 93–102
https://doi.org/10.54941/ahfe1006378

Designing Multimodal Human-Robot Interaction for Social Robots in Office Environments

Sebastian Pimminger, Werner Kurschl, Johannes Schönböck, and Gerald Adam Zwettler

University of Applied Sciences Upper Austria, School of Informatics, Communications and Media, Campus Hagenberg, Hagenberg, Austria

ABSTRACT

Social robots are becoming increasingly relevant and are expected to enter our professional lives. A key challenge lies in designing these robots to ensure their behaviors are easily and intuitively understood by the users and match their expectations. This paper focuses on the design and implementation of a multimodal interaction system for social robots in office environments. Following a user-centered design approach, we iteratively developed and evaluated a prototype through multiple user studies, addressing various office-related use cases such as welcoming arriving guests and guiding them to a room. The results provide key insights into the multimodal interaction user experience of our robot. This helped us to identify key requirements and features for natural and engaging interactions.

Keywords: Social robots, User-centered design, User study, Wizard-of-Oz, Prototyping

INTRODUCTION

Social robots are becoming increasingly important in various areas of everyday life, as major advances in artificial intelligence and robotics (e.g., cobots) have significantly improved the possibilities (Youssef et al., 2022). However, designing robots for office environments requires addressing challenges in functionality, appearance, interaction, and integration to meet the unique needs of busy, professional spaces. This also naturally raises the question of how the appearance and behaviour of a social robot should be designed, as well as which design features are essential.

In our previous work, we explored the *participatory design* of a social robot for office environments, where the initial focus was placed on the appearance (e.g., shape and structure, locomotion, face and facial features, material and texture, and apparel) and physical design. The participatory design was integrated into a user-centered design (UCD) process to gain a deep understanding of the users' work practices, perceptions, beliefs, and experiences with social robots (Pimminger et al., 2024). As a result of this context analysis, we identified key scenes in a user narrative that captured the relevant situations and interaction points that define the human-robot interaction (HRI) experience. These scenes represent key moments that significantly shape the users' overall experience with the envisioned robot. In our narrative, the key scenes include 1) establishing and maintaining contact with people (e.g., at the front desk), 2) navigating through the building, and 3) performing pick-up and delivery services. Based on these scenes, we have identified an initial set of relevant features and design elements categorized into 1) physical features (describing the appearance embodiment) and 2) non-physical features (describing behavioural characteristics). These features and the insights gained from the contextual analysis served as the basic requirements for a multimodal interaction system.

Building on this foundation, this paper explores the concept and requirements of a *multimodal interaction system* for social robots in office environments. We present our findings on HRI and demonstrate how it can be achieved through such a system. We have designed and implemented several prototypes and conducted a comprehensive analysis of the interaction between the users and the robot using multiple studies with Wizard-of-Oz experiments. This provided us with extensive feedback to refine our system and interaction requirements. The findings of our studies are discussed along the interaction modalities: 1) speech-based interaction, 2) visual signals, and 3) auditive signals. Our analysis revealed social presence and personalization as key elements to foster natural and engaging interactions.

BACKGROUND

Service and Social Robots: The broad field of service robots has been attracting increasing interest and is growing rapidly. These robots are becoming more and more integrated into our daily lives, assisting and supporting us in various tasks, such as vacuuming floors (Fink et al., 2013), delivering packages (Lee et al., 2021), or even performing complex tasks like folding laundry (Longhini et al., 2024). In professional environments, robots and humans increasingly share workspaces and collaborate safely and efficiently using industry 4.0 paradigms (Baratta et al., 2023). Such collaborative robot systems (cobots) are deployed in industrial and manufacturing scenarios, where they assist assembly workers (Banh et al., 2015) or support the transport of heavy loads (DelPreto & Rus, 2019). Beyond industrial applications, robots are increasingly integrated into frontline services where they interact with humans in a natural way and engage in social interactions. These applications range from the hotel and hospitality sector (Osawa et al., 2017) to the medical field (Ahn et al., 2019), where robots serve as receptionists (Sutherland et al., 2019).

Despite this wide range of adoption, working in close collaboration scenarios and working side by side with a robot inevitably brings up barriers and challenges such as the risks of malfunction (Wang et al., 2023). It requires not only to consider the safety, but also the design of their interaction and the integration of social capabilities to drive adoption forward. On a high level, these robots must be able to perceive input and generate output that is understandable by their human counterparts. As described by Fong et al., (2003), these robots exhibit human social characteristics such as express and/or perceive emotions, communicate with dialogue, use natural cues (such as gaze), and exhibit distinctive personality and character.

User-Centered Design: To design such robots that interact socially with people, it is curial to gain a deep understand of their needs, behaviors and expectations and the context of use (Jung & Hinds, 2018). Applying a user-centered design approach offers a valuable approach for this by involving user throughout the process, providing feedback and evaluating design solutions in an iterative way. Recent examples of an applied UCD process in the context of robotics can be found in (Zhong & Schmiedel, 2021) and (Pimminger et al., 2024).

Multimodal Human-Robot Interaction: Pollmann & Ziegler (2020) argue that a special focus needs to be put on (personalized) HRI to achieve acceptance and long-term use of social robots. As a newly emerging approach to HRI, *multimodal HRI* allows users to communicate with a robot using various modalities, including voice, text, touch, or even eye movement and bio-signals like EEG and ECG (Su et al., 2023), aiming to improve the way humans and robots communicate with each other by enabling robots to understand and respond to these modalities in a natural and intuitive way. In (Wang, Zheng, Li & Wang, 2024) a survey on techniques applied in multimodal HRI is given, focussing on the four principal modalities vision, auditory and language, haptics, and physiological sensing. Especially Natural language processing and the recent advancements in Large Language Models (LLMs) play a key role in multimodal HRI by enabling speech recognition and comprehension. For example, Wang, Hasler, Tanneberg, Ocker, Joublin, Ceravola, Deigmoeller & Gienger (2024) present LaMI, an LLM-based robotic system to improve multimodal HRI, which coordinates user speech input with robot lid, neck, ear movements and speech output to produce dynamic, multimodal expressions.

CONTEXT ANALYSIS

Methods

In a first step, we conducted an initial context and domain analysis to gain a comprehensive understanding of our users to guide our design for the multimodal interaction of the envisioned robot. This first analysis step was aimed at 1) understanding work practices of employees, 2) identifying their needs and challenges and 3) analyzing the usage context and environment. The context analysis was conducted with users involved in various roles (e.g., assistants, HR staff, and a team lead in the payroll department) through ethnographic methods, including semi-structured interviews and workplace observations. For the analysis of the acquired data, we used common UCD methods such as user narratives, scenarios, thematic analysis, and personas. A detailed description of the process and results can be found in (Pimminger et al., 2024).

Insights and Results

With the context analysis described above, we were able to gain a deeper understanding of how our users consider the use of service robots and how they would like to engage with them in their daily routine. Our analysis revealed that our users desire support primarily for repetitive and non-creative tasks, administrative processes, and to prevent interruptions. Based on these insights, a user narrative was developed that positioned the robot in the front-desk domain. In this role, the service robot greets visitors, answers their questions, assists with navigation, guides them through the building, and redirects them to other personnel if necessary.

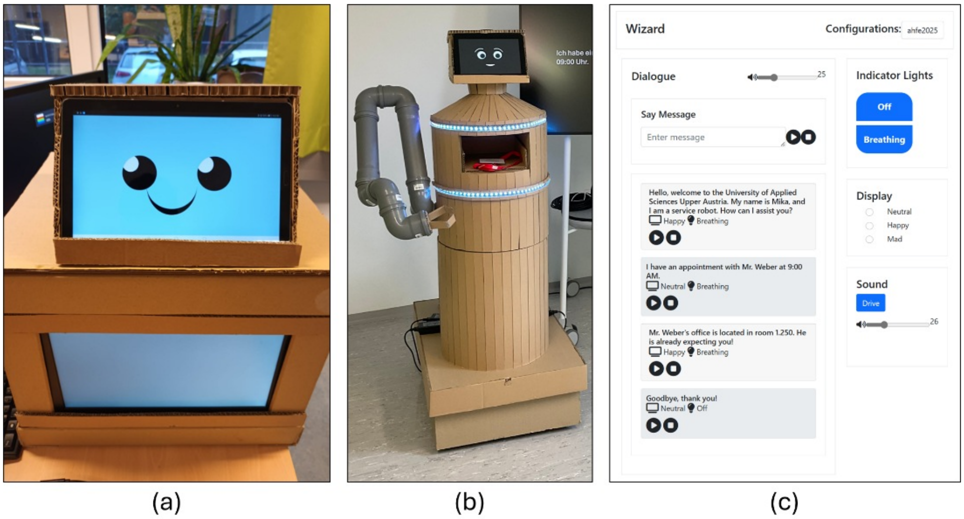
For interacting with the service robot, users expressed their preference for voice-based input and output. Our participants emphasized the importance of a robust speech processing that is capable of recognizing the instructions and translating them into concrete actions that the robot can then execute reliably and without errors. Additionally, they highlighted the need for visual feedback via a display or status indicators to convey task progress and other information. Complementary to voice interaction, tactile input methods such as touchscreens and gesture recognition could be useful, particularly in noisy or crowded environments. Beyond voice interaction, users also expressed their interest in alternative communication channels, such as text-based messaging through a chat window. For instances when the robot is out of sight or earshot, users suggested that it could be called via an instant messaging service.

PROTOTYPING

Following the UCD process, we first created design solutions based on the findings of the context analysis. First, sketches of the envisioned robot were drafted using pen and paper, which were then second transferred into an embodied version using cardboard and other low-cost materials.

Cardboard Prototyping: For early prototyping we created a mock-up robot of the real hardware (a MiR250 as mobile platform, a Universal Robots UR5e robotic arm, and a Robotiq 2F-85 two-finger gripper) using a cardboard engineering approach as described by Frens (2016). This material allows us to build fast and cost-effective iterations, which makes them ideal for prototyping. Additionally, modifications and add-ons can be implemented, and the components can be disassembled and at least partially reused. The initial iteration of the mock-up robot consisted of 1) a head and 2) an upper body as shown in Fig. 1a. In later iterations, we replicated all hardware components like the MIR250 using cardboard, resulting in a flexible prototyping platform for our subsequent studies and evaluations. The dimensions of the mock-up correspond to those of the actual hardware components, allowing for a realistic representation and look-and-feel. The final version used a modular design, with parts such as the lower body, upper body, and the head constructed as separate modules that can be easily detached and replaced independently.

Multimodal Interaction: The prototype is designed to interact through light, voice and facial expressions displayed on a screen, which can also present information. Therefore, it is equipped with a touch display, loudspeakers and two LED strips. Additional hardware, such as a laptop that is used to control the modalities, can be hidden behind the torso to remain unnoticeable. The prototype is controlled by a human “wizard”, who operates the robot’s modalities remotely using the interface shown in Fig. 1c. The custom-built Wizard-of-Oz interface can be used to control the basic modalities such as speech, sound, light indicator patterns and facial expressions or pictures on the display. To give the illusion of autonomous operation, the interface is equipped with a macro feature that combines parameters of the individual modalities which can be organized and played in predefined sequence aligning with the dialogue and interaction flow.



Hello, welcome to the University of Applied Sciences Upper Austria. My name is Mika, and I am a service robot. How can I assist you?
Happy Breathing

I have an appointment with Mr. Weber at 9:00 AM.
Neutral Breathing

Mr. Weber's office is located in room 1.250. He is already expecting you!
Happy Breathing

Goodbye, thank you!
Neutral Off

Figure 1: Examples of early prototypes include (a) the robot’s upper body and (b) a full robot made from cardboard, incorporating various input and output modalities. The prototypes can be controlled via (c) a Wizard-of-Oz interface.

EVALUATIONS

We conducted a multiple-case study with our prototype, gaining valuable insights and feedback on the multimodal interaction with the service robot. In our evaluation, we examined how users interact with the robot, how the different modalities are perceived, and how the robot’s appearance and behavior are evaluated. As methods in our studies, we chose 1) in-depth observations in order to analyze the interaction behavior with the robot, and 2) subsequent interviews to clarify assumptions and gain insights into the user experience. Additionally, 3) questionnaires were used to gather the participants’ impressions of the prototype as well as general demographic data.

The multiple-case study consisted of three studies conducted between August and November 2024, focusing on the interaction between arriving guests and the robot in a reception scenario, following the defined user narrative from above. For *Study A* we recruited eight participants (3f, 5m) and explored the interaction design for guiding a person to a room in a controlled lab environment. *Study B* involved 14 participants (6f, 8m) and focused on the robot’s personality in a receptionist scenario that reflects the initial contact, where the robot welcomed an arriving guest. *Study C* used an iteratively refined version of the navigation scenario from Study A, this time conducted in a real-life field setting with 19 participants (10f, 9m).

All studies used the same prototype robot and setup with the same modalities: speech input/output, light signal indicators, and touchscreen displays for face and other information. Each study was conducted as a Wizard-of-Oz experiment using the previously described interface. Certain parameters, such as speech rate, voice characteristics and light signal patterns and colors, were adapted and slightly modified for each study. In all study runs, the participants were given a brief introduction to the specified study scenario and provided with dialogue scripts. They could either read their lines from the provided sheet (or a screen in Study B) or speak freely along the lines. The robot's responses were unknown to the participants in advance. Then, the participants were asked to interact with the robot while a minimum of two observers were present to take notes during the observation. After completing the interaction, participants filled out questionnaires and we conducted a semi-structured interview to capture their initial impressions as well as issues they encountered, and other positive and negative aspects of the interaction. Finally, a demographic questionnaire concluded each study.

The analysis and synthesis of the individual studies into a multiple-case study were conducted through several systematic steps: First, each study was analyzed separately, highlighting its specific results and key findings. Subsequently, the results were compared with a focus on multimodal HRI, aiming to identify commonalities, differences and recurring patterns. The consolidated results are described and interpreted in the next section.

FINDINGS AND DISCUSSION

In this section, we present the key findings of our user studies from the prototype evaluation, focusing on the different modalities of interaction and their impact on the user experience. The aim is to identify areas for improvement, key requirements, and highlight features. Based on these findings, we identified key elements of social presence and personalization that contribute to a natural and engaging interaction.

Interaction With the Prototype

Speech-based Interaction: In general, speech-based interaction was perceived as pleasant, straightforward, fluent and comprehensible. However, some participants criticized the slow reaction times or delayed speech in certain situations, even though a human was controlling the robot via the Wizard-of-Oz interface and responded as quickly as possible. There were also overly lengthy and monotonous greetings and explanations (e.g., regarding room positions) and repetitive responses during extended interactions. In this case, the participants expressed a desire for the possibility to interrupt and control the speech output via a barge-in function (e.g., stop a lengthy introduction of the robot or repeat the last utterance). Based on the findings and suggestions from the participants indicate that improvements could be achieved by shortening and dynamically tailor parts of the dialogue (e.g., greetings), adjusting the speech tempo, and allowing users to steer the dialogue.

Visual Signals: The visual elements, such as the light ring indicators and the animated, expressive face of our robot, can be seen as helpful cues that communicate the robot's status and support the different phases of interaction. Animation effects (e.g., reading and loading animations) are generally perceived as realistic, friendly, and good cues for status indication. Some participants point out and appreciate the use of the light ring effects that make the robot appear more lifelike. However, a major issue in all studies was that the light signals (mainly from the light ring indicators) were often not distinctively differentiated or simply misinterpreted, and their underlying meaning frequently remained unclear and therefore ignored altogether (e.g., not understanding when the robot switches to a lighthouse pattern during prolonged information processing to indicate its current operational status). We also found that often these signals are not seen and recognized at all (e.g., turn indicators in the navigation scenario). Variations in eye movements, such as a wandering gaze, sometimes come across as too distracting. This indicates a need for clearer, more user-friendly visual cues to better convey their meaning and functionality. For example, colors could be used to differentiate various states (e.g., error, processing, confirmation, etc.).

Auditive Signals: Non-verbal auditory signals, such as (artificial) motor sounds, can be helpful for indicating movement and various operational states of the robot (e.g., only a limited or no interaction is possible while the robot is moving). These sounds were generally described as helpful. However, the volume, particularly during navigation scenarios, was often perceived as too low and difficult to notice. Auditory signals were also the most conservatively applied and least utilized modality in our prototype. The participants suggested several improvements. Additional auditory signals could communicate states of the robot, such as indicating a dialog turn or providing danger warnings. For example, a sound could signal that the robot is "thinking" or processing information, thereby supporting visual cues. Furthermore, a dynamic adjustment of the volume to match louder environments would improve the perception of this modality.

Fostering Natural and Engaging Interactions

Social Presence: The robot is perceived as friendly, professional, kind, and helpful, contributing to a positive emotional connection with the user. However, that analysis of the studies showed that nonverbal communication (e.g., gaze behaviors) should be further refined to create a more natural interaction. A noticeable issue is the discrepancy between the expected and perceived voice profile (at the time of the studies, the robot had a feminine name but a voice that was perceived as rather masculine). This mismatch causes confusion and weakens the empathetic interaction. Additionally, a more balanced interaction, where the robot alternates between proactive and user-initiated interaction, could result in more natural and emotionally engaging dialogues. Also, (active) movements would make the robot's presence feel livelier and more social. Another important aspect for strengthening the presence would be the ability to recognize individuals. This would foster a long-term, personalized relationship between the robot and its users, making the interaction more personal and meaningful.

Personalization: Personalization is another key element for natural, engaging interaction, as it helps tailor the interaction to the individual needs of users. For example, participants emphasized that regular and consistent use of personal address (e.g., using the user's name) is valuable and makes the interaction feel more personal, which can, in turn, foster a stronger emotional connection. At the same time, there is a desire that the robot is more adaptable to the individual preferences of the users (e.g., speech tempo, voice, dialogue style). This would also require the robot to recognize users across multiple interactions and adapt to previous conversations. Such capabilities would strengthen a long-term personalized relationship between the robot and its users.

CONCLUSION

Designing multimodal HRI is a challenging task. In this paper, we identified key requirements and features that enable natural and engaging interaction between humans and robots.

Our findings show that speech-based interaction was generally well-perceived but could be enhanced by optimizing the dialogue structure, adjusting speech tempo and adding features to steer and control the flow of the conversation. Visual signals, while appreciated for making the robot appear alive, require a clearer differentiation to convey their meaning effectively. Auditory signals offer potential to complement the visual cues and enrich the interaction. Based on these findings, we identified social presence and personalization as a critical role in fostering personal and engaging interactions with social robots.

As future work, we are currently implementing these improvements and refine the multimodal interaction. We also explore how to apply our insights and findings to other tasks within an office environment, such as pick-and-delivery services and generally assisting employees in saving time on recurring tasks. By offering adaptability and personalization capabilities, we hope to further improve the usability and appeal of our social robot.

ACKNOWLEDGMENT

The research leading to these results has been accomplished within the project *MARIE - Mobile Robotic Assistance for Office Environments*, funded by the Upper Austrian government under grant programme "Programm des Landes OÖ zur Stimulierung der Erschließung/Erweiterung von zukunftsweisenden Forschungsfeldern bei den OÖ außeruniversitären Forschungseinrichtungen im Zeitraum 01.01.2022 – 31.12.2029". The content is solely the responsibility of the authors and does not necessarily represent the official views of the research sponsor.

REFERENCES

Ahn, H. S., Yep, W., Lim, J., Ahn, B. K., Johanson, D. L., Hwang, E. J., Lee, M. H., Broadbent, E. & MacDonald, B. A. (2019), Hospital receptionist robot v2: Design for enhancing verbal interaction with social skills, in “2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)”, IEEE, New York, NY, USA, pp. 1–6.

Banh, A., Rea, D. J., Young, J. E. & Sharlin, E. (2015), Inspector baxter: The social aspects of integrating a robot as a quality inspector in an assembly line, in “Proceedings of the 3rd International Conference on Human-Agent Interaction”, HAI ‘15, Association for Computing Machinery, New York, NY, USA, pp. 19–26.

Baratta, A., Cimino, A., Gnoni, M. G. & Longo, F. (2023), “Human robot collaboration in industry 4.0: A literature review”, Procedia Computer Science 217, 1887–1895.

DelPreto, J. & Rus, D. (2019), Sharing the load: Human-robot team lifting using muscle activity, in “2019 International Conference on Robotics and Automation (ICRA)”, IEEE, Montreal, Canada, pp. 7906–7912.

Fink, J., Bauwens, V., Kaplan, F. & Dillenbourg, P. (2013), “Living with a vacuum cleaning robot”, International Journal of Social Robotics 5(3), 389–408.

Fong, T., Nourbakhsh, I. & Dautenhahn, K. (2003), “A survey of socially interactive robots”, Robotics and Autonomous Systems 42(3), 143–166. Socially Interactive Robots.

Frens, J. J. (2016), Cardboard Modeling: Exploring, Experiencing and Communicating, Springer International Publishing, Cham, pp. 149–177.

Jung, M. & Hinds, P. (2018), “Robots in the wild: A time for more robust theories of human-robot interaction”, J. Hum.-Robot Interact. 7(1).

Lee, D., Kang, G., Kim, B. & Shim, D. H. (2021), “Assistive delivery robot application for real-world postal services”, IEEE Access 9, 141981–141998.

Longhini, A., Wang, Y., Garcia-Camacho, I., Blanco-Mulero, D., Moletta, M., Welle, M., Alenyà, G., Yin, H., Erickson, Z., Held, D., Borràs, J. & Kragic, D. (2024), “Unfolding the literature: A review of robotic cloth manipulation”, Annual Review of Control, Robotics, and Autonomous Systems.

Osawa, H., Ema, A., Hattori, H., Akiya, N., Kanzaki, N., Kubo, A., Koyama, T. & Ichise, R. (2017), What is real risk and benefit on work with robots? from the analysis of a robot hotel, in “Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction”, HRI ‘17, Association for Computing Machinery, New York, NY, USA, pp. 241–242.

Pimminger, S., Kurschl, W., Schönböck, J., Slabihoud, R., Froschauer, R. & Zwettler, G. A. (2024), Towards a user-centered design approach for the development of social assistive robots in office environments, in “Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments”, PETRA ‘24, Association for Computing Machinery, New York, NY, USA, pp. 45–54.

Pollmann, K. & Ziegler, D. (2020), Social human-robot interaction is personalized interaction, in “Workshop on Behavioral Patterns and Interaction Modelling for Personalized Human-Robot Interaction 2020”, Fraunhofer IAO, Stuttgart, Germany.

Su, H., Qi, W., Chen, J., Yang, C., Sandoval, J. & Laribi, M. A. (2023), “Recent advancements in multimodal human–robot interaction”, Frontiers in Neurorobotics 17.

Sutherland, C. J., Ahn, B. K., Brown, B., Lim, J., Johanson, D. L., Broadbent, E., MacDonald, B. A. & Ahn, H. S. (2019), The doctor will see you now: Could a robot be a medical receptionist?, in “2019 International Conference on Robotics and Automation (ICRA)”, IEEE, New York, NY, USA, pp. 4310–4316.

Wang, C., Hasler, S., Tanneberg, D., Ocker, F., Joublin, F., Ceravola, A., Deigmoeller, J. & Gienger, M. (2024), Lami: Large language models for multimodal human-robot interaction, in “Extended Abstracts of the CHI Conference on Human Factors in Computing Systems”, CHI EA ‘24, Association for Computing Machinery, New York, NY, USA.

Wang, T., Zheng, P., Li, S. & Wang, L. (2024), “Multimodal human–robot interaction for human-centric smart manufacturing: A survey”, Advanced Intelligent Systems 6(3).

Wang, X., Zhang, Z., Huang, D. & Li, Z. (2023), “Consumer resistance to service robots at the hotel front desk: A mixed-methods research”, Tourism Management Perspectives 46, 101074.

Youssef, K., Said, S., Alkork, S. & Beyrouthy, T. (2022), “A survey on recent advances in social robotics”, Robotics 11(4).

Zhong, V. J. & Schmiedel, T. (2021), A user-centered agile approach to the development of a real-world social robot application for reception areas, in “Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction”, HRI ‘21 Companion, Association for Computing Machinery, New York, NY, USA, pp. 76–80.