Investigating Opportunities for Active Smart Assistants to Initiate Interactions With Users

Jan Leusmann
LMU Munich
Munich, Germany
jan.leusmann@ifi.lmu.de

Jannik Wiese
LMU Munich
Munich, Germany
jannik.wiese@campus.lmu.de

Moritz Ziarko
LMU Munich
Munich, Germany
moritz.ziarko@campus.lmu.de

Sven Mayer
LMU Munich
Munich, Germany
info@sven-mayer.com

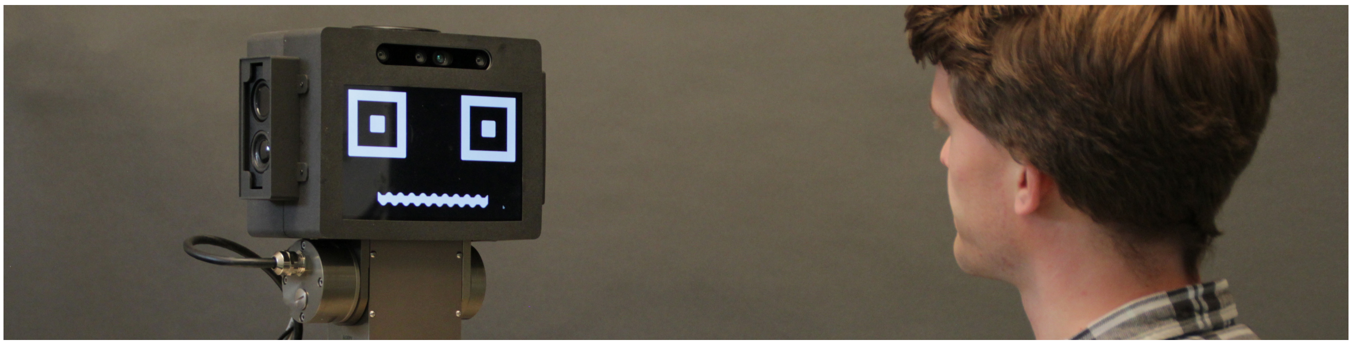

Figure 1: Picture of our system acting as an active smart assistant interacting with a user.

ABSTRACT

Passive voice assistants such as Alexa are widespread, responding to user requests. However, due to the rise of domestic robots, we envision active smart assistants initiating interactions seamlessly, weaving themselves into the user's context, and enabling more suitable interaction. While robots already deliver the hardware, only recently have the advancements in artificial intelligence enabled assistants to grasp the human and the environments to support such visions. We combined hardware with artificial intelligence to build an attentive robot. Here, we present a robotic head prototype discovering and following the users in a room supported by video and sound. We contribute (1) the design and implementation of a prototype system for an active smart assistant and (2) a discussion on design principles for systems engaging in human conversations. This work aims to provide foundations for future research for active smart assistants.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Interaction design**.

KEYWORDS

human computer interaction, human robot interaction, voice assistants, conversational agents

MUM '23, December 03–06, 2023, Vienna, Austria

ACM ISBN 979-8-4007-0921-0/23/12.
https://doi.org/10.1145/3626705.3631787

ACM Reference Format:
Jan Leusmann, Jannik Wiese, Moritz Ziarko, and Sven Mayer. 2023. Investigating Opportunities for Active Smart Assistants to Initiate Interactions With Users. In *International Conference on Mobile and Ubiquitous Multimedia (MUM '23), December 03–06, 2023, Vienna, Austria.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626705.3631787

1 INTRODUCTION

Currently, the interaction with smart assistants starts through user interactions. However, as domestic robots like vacuum cleaners become more prevalent, proactive engagement can enhance user experience. For instance, a vacuum cleaner could autonomously inquire about an ideal cleaning time, relieving users from explicit routine setups. Prior work has already investigated robots initiating interactions with users through eye contact [9, 27], human-like approaching behavior [8, 12, 26], or verbally [19]. Shi et al. [28] propose a model for constraints and expected behavior for humanoid robots initiating conversations. One challenge is to reliably detect the current user state and position to understand when and how these systems can approach users. While recent advancements allow reliable detection of objects [25], humans [14, 18, 22], and actions [3], there is a need to derive various human states for optimal system-user interaction moments. Humans excel at intuitively assessing suitable times for questions, unlike current device notifications that often disrupt users.

We propose a more human-like, implicit communication approach for requests to address this. In contrast to machines, humans can also include not observable factors to determine user states, e.g., interpersonal connections. On the other hand, machines need to rely on only observable factors. Thus, we explore possible observable human states and how systems can approach users in these different states.

In this work, we showcase an initial implementation of a future active smart assistant, see Figure 1. Related work shows that future smart systems can come in different designs and interaction possibilities [5, 7]. While such design designs are vast and can impact the perception and expectation of that system, the basic concept of an actively engaging system stays the same. Thus, we build a prototype system, considering both the needed functionality and the ability to communicate the robot's intent. In this initial work, we developed and showcased a more limited static robot using a pan-tilt unit (PTU), focusing on the essential aspects. The robot is equipped with a depth camera, a 4-channel microphone, stereo speakers, and a display showing an animated face. To derive the human position we chose to combine audio [4, 10, 15] and visual [18] channels for the best performance [6, 16]. With this, we developed a voice user interface circle and conversational logic for verbal interaction between the assistant and the human.

2 DESIGN CONSIDERATIONS

Initiating conversations between humans includes multiple steps and ways how this can occur [23], e.g., becoming physically copresent, greeting both audible or visible, touching, or introducing oneself. While humans are generally good at understanding these initiations [32], we currently do not know which concepts are similar or different to human-human interaction when adapting these ideas to human-robot interaction. In the following, we describe our considerations about the user state, how the system can initiate conversations, and how we envision interaction.

Observable User State. Humans can easily infer other humans' current state from many different information sources [21, 30]. This information can either be observable (facial expression, body language, or eye contact) or non-observable (personal experiences, stress level, or beliefs and values). From this observable and non-observable information, we adapt the way we approach others. In contrast to humans, machines need to rely on observable factors only to derive the current human state [13], which has already been done for smart home devices to improve user experience [31]. As a first step, we propose that observable factors are the number of people in the room, location of people, ongoing conversations or silence, and people entering and leaving the room [11]. Furthermore, the system should observe gaze direction [9] and the activity each person is doing [31]. The system can build knowledge about people by asking questions and recognizing people over time. Ultimately, the system can decide when and how to initiate conversations by taking these factors into account.

Robot Initiation. The system can initiate conversations based on observed human states using either verbal, non-verbal, or combined communication approaches. Verbal communication provides direct and fast information transfer. Non-verbal cues, delivered through moving expressions and facial cues, offer non-intrusive initiation, allowing users to choose whether they want to respond. Supporting verbal through non-verbal elements can enhance user comfort [29]. Deciding when and how often the system retries initiation after being ignored is a crucial consideration. Context, particularly the urgency of information, influences the approach, distinguishing between time-sensitive notifications and inquiries aimed at enhancing future interactions. Furthermore, the system chooses one target person, which should be approached [26].

Envisioned Interaction. We envision that an active smart assistant should precisely detect the position of people in the room and know for each person if it has seen them before and what information it already has about that person. The system should then be able to initiate conversations based on information it wants to give to users or needs from users. During conversations, the system should understand the user. Later, such systems could also be connected to other smart home devices to be able to gather information about and control the environment.

3 IMPLEMENTATION

We designed, constructed, and implemented an active smart assistant, see Figure 1. The system utilizes a Schunk PTU to enable horizontal and vertical rotation. The system features an 8" display for show a face via an HTML webpage, a Seed ReSpeaker 4-channel microphone, stereo speakers, and a Realsense D455 camera controlled by a Jetson Nano running Ubuntu 20.04 and ROS for communication. Additionally, we 3D printed a body for the system. We provide the STL files, part list, and build instructions via OSF[1].

We estimated the Direction of Arrival (DOA) of the audio signal using the onboard predefined DOA estimation of the ReSpeaker, which is based on GCC-Phat [4, 15]. We integrated mediapipe's pose detection and face mesh [18] along with voice activity detection [10] to determine user locations in the room. We fused the position of detected faces with detected bodies through Euclidean distance matching and took the midpoint between the shoulder landmarks as the position for each person.

The system utilizes facial expressions, including pupil movement, panning towards users, animated talking mouth, and naturalistic blinking. It can freely turn around to search for conversation partners or lock onto a specific person during conversation mode, following their movements. The system detects unknown people using face recognition [14]. We employ verbal interaction capabilities through speech-to-text using OpenAI Whisper [24] and text-to-speech using Nvidia Riva[2]. The dialogue management system employs a state machine, and we extract the users' intent through joint intent classification and slot tagging [2].

4 LIMITATIONS AND FUTURE WORK

Recent advancements in small-scale computing have progressed to levels where an Nvidia Jetson Nano can handle basic machine learning models, but they still face limitations with larger models necessitating additional computing power. One main challenge of our implementation is detecting speaker direction from audio-only when not all users are in the field of view of the camera. Due to noise and other audio sources at the same time, our system occasionally misdirects toward the non-dialogue audio source. While we can fix this through different interaction concepts, and the behavior of the system turning towards sounds that are not voices can also appear natural, it can still lead to interaction challenges. With our

[1] https://osf.io/9qv54/?view_only=cc677d8072d14899b5af0153fe865bc4
[2] https://docs.nvidia.com/deeplearning/riva/user-guide/docs/

work, future work can now look into how people actually like to interact with an active smart assistant. We propose to study different initiation methods from the robot in different user states. Finally, we propose that interaction can be improved even further by including even more world information through additional sensors.

We designed a dialog management system using a simple state machine. The active smart system determines the progression of the dialog to the next step by assessing the understanding of the human response. Future work can enhance this interaction by incorporating additional factors, e.g., knowledge about the humans' and robots' uncertainty, to create a more natural conversation [17].

Our prototype used a smart assistant as a foundation and integrated robotic features via the PTU. This allows the system to rotate and pan towards users with a displayed face, which has been shown to enhance users' perception of the system [1, 20]. Future enhancements may include additional robotic elements, such as a body, arms, or hands, to further embody the system [33].

5 CONCLUSION

In this work, we present an active smart assistant. We build a robotic head that can turn towards detected humans and sense speaker directions. This system can then automatically initiate conversations with users and understand the user's intent. We designed the first version of a conversational logic unit for the system to decide when to initiate conversations. Furthermore, we discuss our design considerations to understand and act according to the current user state. With this work, we propose a groundwork for future research to investigate what interaction with active smart assistants can look like.

REFERENCES

[1] Elizabeth Broadbent, Vinayak Kumar, Xingyan Li, John Sollers 3rd, Rebecca Q. Stafford, Bruce A. MacDonald, and Daniel M. Wegner. 2013. Robots with Display Screens: A Robot with a More Humanlike Face Display Is Perceived To Have More Mind and a Better Personality. *PLOS ONE* 8, 8 (2013). https://doi.org/10.1371/journal.pone.0072589

[2] Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for Joint Intent Classification and Slot Filling. (2019). https://doi.org/10.48550/arxiv.1902.10909

[3] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. 2016. Online Action Detection. arXiv:1604.06506 [cs]

[4] Joseph Hector Dibiase. 2000. *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays.* Ph. D. Dissertation.

[5] Rolf Dieter Schraft, Birgit Graf, Andreas Traub, and Dirk John. 2001. A Mobile Robot Platform for Assistance and Entertainment. *Industrial Robot: An International Journal* 28, 1 (2001), 29–35. https://doi.org/10.1108/01439910110380424

[6] Jannik Fritsch, Marcus Kleinehagenbrock, Sebastian Lang, Gernot A. Fink, and Gerhard Sagerer. 2004. Audiovisual Person Tracking with a Mobile Robot. *Proc. Int. Conf. on Intelligent Autonomous Systems* (2004).

[7] Randy Gomez, Deborah Szapiro, Kerl Galindo, and Keisuke Nakamura. 2018. Haru: Hardware Design of an Experimental Tabletop Robot Assistant. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/3171221.3171288

[8] Van Bay Hoang, Van Hung Nguyen, Trung Dung Ngo, and Xuan-Tung Truong. 2023. Socially Aware Robot Navigation Framework: Where and How to Approach People in Dynamic Social Environments. *IEEE Transactions on Automation Science and Engineering* 20, 2 (2023), 1322–1336. https://doi.org/10.1109/tase.2022.3174141

[9] Mohammed Moshiul Hoque, Yoshinori Kobayashi, and Yoshinori Kuno. 2014. A Proactive Approach of Robotic Framework for Making Eye Contact with Humans. *Advances in Human-Computer Interaction* 2014 (2014). https://doi.org/10.1155/2014/694046

[10] Fei Jia, Somshubra Majumdar, and Boris Ginsburg. 2020. MarbleNet: Deep 1D Time-Channel Separable Convolutional Neural Network for Voice Activity Detection. (2020). https://doi.org/10.48550/arxiv.2010.13886

[11] Daphne Karreman, Lex Utama, Michiel Joosse, Manja Lohse, Betsy van Dijk, and Vanessa Evers. 2014. Robot Etiquette: How to Approach a Pair of People?. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (Bielefeld, Germany) *(HRI '14)*. Association for Computing Machinery, New York, NY, USA, 196–197. https://doi.org/10.1145/2559636.2559839

[12] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. 2015. May I Help You? Design of Human-like Polite Approaching Behavior. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 35–42. https://doi.org/10.1145/2696454.2696463

[13] Wesley Kerr and Paul Cohen. 2010. Recognizing Behaviors and the Internal State of the Participants. In *2010 IEEE 9th International Conference on Development and Learning* (Ann Arbor, MI, USA, 2010-08). Ieee, 33–38. https://doi.org/10.1109/devlrn.2010.5578868

[14] Davis King. 2017. High Quality Face Recognition with Deep Metric Learning.

[15] Charles Knapp and G. Clifford Carter. 1976. The Generalized Correlation Method for Estimation of Time Delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24, 4 (1976), 320–327. https://doi.org/10.1109/tassp.1976.1162830

[16] Sebastian Lang, Marcus Kleinehagenbrock, Sascha Hohenner, Jannik Fritsch, Gernot A. Fink, and Gerhard Sagerer. 2003. Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (Vancouver, British Columbia, Canada) *(ICMI '03)*. Association for Computing Machinery, New York, NY, USA, 28–35. https://doi.org/10.1145/958432.958441

[17] Jan Leusmann, Chao Wang, Michael Gienger, Albrecht Schmidt, and Sven Mayer. 2023. Understanding the Uncertainty Loop of Human-Robot Interaction. https://doi.org/10.48550/arXiv.2303.07889 arXiv:2303.07889 [cs.HC]

[18] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. (2019). https://doi.org/10.48550/arxiv.1906.08172

[19] Kazuki Mizumaru, Satoru Satake, Takayuki Kanda, and Tetsuo Ono. 2019. Stop Doing It! Approaching Strategy for a Robot to Admonish Pedestrians. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI '19)*. IEEE, Daegu, Korea (South), 449–457. https://doi.org/10.1109/hri.2019.8673017

[20] Ali Mollahosseini, Hojjat Abdollahi, Timothy D. Sweeny, Ron Cole, and Mohammad H. Mahoor. 2018. Role of Embodiment and Presence in Human Perception of Robots' Facial Cues. *International Journal of Human-Computer Studies* 116 (2018), 25–39. https://doi.org/10.1016/j.ijhcs.2018.04.005

[21] D. Patel, Steve Fleming, and James Kilner. 2012. Inferring Subjective States through the Observation of Actions. *Proceedings of the Royal Society B: Biological Sciences* 279, 1748 (2012), 4853–4860. https://doi.org/10.1098/rspb.2012.1847

[22] Manoranjan Paul, Shah M. E. Haque, and Subrata Chakraborty. 2013. Human Detection in Surveillance Videos and Its Applications - a Review. *EURASIP Journal on Advances in Signal Processing* 2013, 1 (2013), 176. https://doi.org/10.1186/1687-6180-2013-176

[23] Danielle Pillet-Shore. 2018. How to Begin. *Research on Language and Social Interaction* 51, 3 (2018), 213–231. https://doi.org/10.1080/08351813.2018.1485224

[24] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. (2022). https://doi.org/10.48550/arxiv.2212.04356

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 779–788. https://doi.org/10.1109/cvpr.2016.91

[26] Satoru Satake, Takayuki Kanda, Dylan F. Glas, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2009. How to Approach Humans? Strategies for Social Robots to Initiate Interaction. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (La Jolla, California, USA) *(HRI '09)*. Association for Computing Machinery, New York, NY, USA, 109–116. https://doi.org/10.1145/1514095.1514117

[27] Shayla Sharmin, Mohammed Moshiul Hoque, S. M. Riazul Islam, Md. Fazlul Kader, and Iqbal H. Sarker. 2021. Development of Duplex Eye Contact Framework for Human-Robot Inter Communication. *IEEE Access* 9 (2021), 54435–54456. https://doi.org/10.1109/access.2021.3071129

[28] Chao Shi, Michihiro Shimada, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2012. Spatial Formation Model for Initiating Conversation. In *Robotics*. The MIT Press, 305–312. https://doi.org/10.7551/mitpress/9481.003.0044

[29] Mei Si and Joseph Dean McDaniel. 2016. Using Facial Expression and Body Language to Express Attitude for Non-Humanoid Robot: (Extended Abstract). In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (Singapore, Singapore) *(AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1457–1458.

[30] Robert P. Spunt and Ralph Adolphs. 2019. The Neuroscience of Understanding the Emotions of Others. *Neuroscience Letters* (2019), 44–48. https://doi.org/10.1016/j.neulet.2017.06.018

[31] Shashi Suman, Francois Rivest, and Ali Etemad. 2022. Towards Personalization of User Preferences in Partially Observable Smart Home Environments. arXiv:2112.00971 [cs]

[32] Paul Thagard and Ziva Kunda. 1997. *Making Sense of People: Coherence Mechanisms.* Hillsdale, NJ: Erlbaum.

[33] Joshua Wainer, David Feil-seifer, Dylan Shell, and Maja Mataric. 2006. The Role of Physical Embodiment in Human-Robot Interaction. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication* (2006-09). IEEE, 117–122. https://doi.org/10.1109/roman.2006.314404