2020

Multi-modal Personalisation in Long-Term Human-Robot Interaction

Irfan, Bahar

http://hdl.handle.net/10026.1/16756

http://dx.doi.org/10.24382/663
University of Plymouth

MULTI-MODAL PERSONALISATION IN LONG-TERM HUMAN-ROBOT INTERACTION

by

BAHAR IRFAN

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering, Computing and Mathematics

December 2020

Acknowledgements

First and foremost, I would like to thank my director of studies, Tony Belpaeme. This work would have not been possible without his continuous support and motivation. Thank you Tony, for giving me this chance and being an amazing supervisor. I could not have imagined a better PhD.

Secondly, I would like to thank Natalia Lyubova for her positiveness and strong belief in my work, supporting me even after leaving SoftBank Robotics Europe. I really wish that we had more chance to work together. I would also like to thank Michaël Garcia Ortiz for accepting to supervise me after Natalia left, and his support in my work. Furthermore, I would like to thank all the remaining supervisors of the APRIL team, Angelo Cangelosi, Torbjorn Dahl, Alban Laflaquière, Nikolas Hemion and Angelica Lim, for their various types of support, to the University of Plymouth and SoftBank Robotics Europe for the great hospitality and research environment, and to the European Union and the Marie Skłodowska-Curie Actions for funding this incredible project.

I would like to thank the people that I had the chance to work with at SoftBank. First of all, to my previous intern, Mehdi Hellou, for doing a great work in building a barista robot and conducting real-world studies on his own. Also, I would like to thank Rodolphe Gelin for the opportunity to conduct this internship project at SoftBank, and Alexandre Mazel for his supervision and feedback to Mehdi in my absence. I would also like to thank Maxime Busy for his help with the automatic speech recognition, and to Victor Paléologue and Jérôme Bruzaud for adapting my user recognition method for the newer versions of NAOqi. I would also like to thank the remaining members of AI Lab and Innovation department at SoftBank for all the great discussions and gatherings.

I would like to thank the team in Colombian School of Engineering Julio Garavito for a great collaboration. To Jonathan Casas for making everything work and his happy spirits, to Nathalia Céspedes Gomez for her immense help in gathering and analysing the data, and to Carlos A. Cifuentes and Marcela Múnera for this great project and their warm welcome in Bogotá. In addition, I would like to thank the medical specialists at Fundación Cardioinfantil-Instituto de Cardiología, Mónica Rincon Roncancio and Luisa F. Gutiérrez, for their invaluable suggestions and feedback, and to all the medical team and the patients at the clinic for accepting to integrate a robot as part of their rehabilitation for 3.5 years. I would like to thank the Royal Academy of Engineering for funding this impactful project. I would also like to thank Emmanuel Senft for a great collaboration for this work, being a great friend and for giving me the first tour of Plymouth.

I would like to thank James Kennedy for a fantastic internship opportunity at Disney Research. I was delighted to be part of the visionary Walt Disney Imagineering R&D team with an exciting, collaborative and successful project. I would also like to thank Anika Narayanan for her creative mind and all the members of Disney Research and R&D for their warm welcome.

I would like to thank all my friends and fellow researchers at the marvellous international HRI community, Ross Mead, Serge Thill, Brad Hayes, Séverin Lemaignan, Fotios Papadopoulos, Chris Wallbridge, Paul Baxter, among many others. Thank you all for your support, friendship and feedback for my work. I would like to thank my co-organisers at the Personalisation in Long-Term Human-Robot Interaction workshop, Aditi Ramachandran, Samuel Spaulding, Dylan F. Glas, Iolanda Leite and Kheng Lee Koay, for all their efforts in creating a great workshop. I would like to thank Ross Mead and Todd Terrazas for introducing me to AI LA, which was a great networking opportunity for me. Moreover, I would like to thank the pyAgrum library creator, Pierre-Henri Wuillemin for his substantial help in describing the library, on which my user recognition algorithm is built.

I would like to thank my friends at the University of Plymouth, Marta Romeo, Mina Marmpena, Samuele Vinanzi, Barış Serhan, Massimiliano Patacchiola, Riccardo Polvara, Leszek Pecyna, Emmanuel Senft, Chris Wallbridge, Yogangh Singh, Daniel Hernández García, Gabriella Pizzuto, Martina Fiori, Mohammad Thabet, Salvatore Lentini, and Tatsuro Yamada for taking part in my study for a month, and for all the lively gatherings and the great food.

A big thank you goes to each and every one of my friends in SoftBank, Paris, Plymouth, Turkey and all over the world.

I would like to thank the APRIL fellows, Alexandre Antunes, Oksana Hagen, Mina Marmpena, Pontus Loviken and Leszek Pecyna, for being a great team. I could not have think of a better one!

I would like to thank the ones who kept their prayers and thoughts with me, my sweet Vesile and İrem Hamulu, my dear Sabriye Alkaç, and ma chère Rosy Yvette Perroni, and to all others who are dear to me! I would also like to thank my grandmother's cat, Zikrettin, and our newly adopted dog, Sky, for their love and joy.

I would once again like to thank Alexandre Antunes, my best friend who is like a brother to me, for being there for bitter and sweet, and always giving me a hug when I need it.

I would particularly like to thank Valerio Biscione, for all the remarkable memories in more than two amazing years together, for his beautiful music, moral support and invaluable help with Bayesian networks.

A special thank you goes to Luca Grillotti, for his exceptional support, affection and patience through the most challenging part of my PhD, and for the sweet, wonderful and unforgettable two years.

I am sincerely grateful to my family, especially to my parents, Hanife İrfan and Nuri Coşkun İrfan, my sister, Elif Pınar İrfan, my grandmother, Yüksel Carban, and my uncle (a brother to me), Haluk Carban, for everything... For all your support, belief in my work and your love. I could not have done any of this without you. Thank you for always being there for me, even when so far apart. You all are the lights that guide me in the dark.

I would also like to thank my grandparents that passed away, Şekibe Nazmiye İrfan, İsmail İrfan, Salih Carban and to my great aunt Leman Büyükçubukçu. You are still in my heart and will always be. Thank you for watching over me from the skies.

Lastly, I would like to thank the founder of the Republic of Turkey, Mustafa Kemal Atatürk, for giving us, especially women, the chance to pursue dreams, take a lead in every area and learn to fly.

This thesis is dedicated to my parents, Hanife and Nuri Coşkun İrfan.

Author’s declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

No work submitted for a research degree at University of Plymouth may form part of any other degree for the candidate either at the University or at another establishment.

This work has been carried out by Bahar Irfan under the supervision of Prof. Dr. Tony Belpaeme, Dr. Michaël Garcia Ortiz, Dr. Natalia Lyubova, and Dr. Torbjorn Dahl. The work received funding from the EU H2020 Marie-Skłodowska Curie Actions Innovative Training Networks project APRIL (Applications of Personal Robotics for Interaction and Learning, grant no: 674868). Equal amount of time (18 months) has been spent at University of Plymouth and SoftBank Robotics Europe for this research, but no additional award will be awarded.

Parts of this thesis have been published and presented by the author at the following workshops:

Irfan, B., Lyubova, N., Garcia Ortiz, M., & Belpaeme, T. (2018b). Multi-modal open-set person identification in HRI. In *2018 ACM/IEEE International Conference on Human-Robot Interaction Social Robots in the Wild workshop*
URL: `http://socialrobotsinthewild.org/wp-content/uploads/2018/02/HRI-SRW_2018_paper_6.pdf`
PEARL (OA): `http://hdl.handle.net/10026.1/16699`

Irfan, B. (2019a). Multi-modal personalisation in long-term human-robot interaction. In *9th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL- EpiRob 2019), Workshop on Personal Robotics and Secure Human-Robot Collaboration*
URL: `https://www.fose1.plymouth.ac.uk/socem/crns/april/Oslo/APRIL_SECURE_ICDL_with_Abstracts.pdf`
PEARL (OA): `http://hdl.handle.net/10026.1/16702`

Additionally, parts of this thesis have been presented at the 2019 Behaviors.AI workshop as an invited talk:

Irfan, B. (2019). Multi-modal personalisation in long-term human-robot interaction. In *Behaviors.AI workshop*
URL: `https://behaviors.ai/workshop-2019/`
PEARL (OA): `http://hdl.handle.net/10026.1/16710`

The work on the generic and personalised barista robots was conducted at SoftBank Robotics Europe in collaboration with Mehdi Hellou, who was funded by SoftBank Robotics Europe as a research intern, under the main supervision of the author along with joint supervision of Alexandre Mazel. The barista robot experiment was conducted at Cité Internationale Universitaire de Paris. In addition to supervising Mehdi Hellou, the author designed the barista robot and the rule-based dialogue management system architectures, created the Barista Datasets, developed the evaluation methods for the user study, conducted the analysis of the study, and presented the work as the first author at the 2020 ACM/IEEE International Conference on Human-Robot Interaction:

Irfan, B., Hellou, M., Mazel, A., & Belpaeme, T. (2020b). Challenges of a real-world HRI study with non-native english speakers: Can personalisation save the day? In *Companion of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM
DOI: `https://doi.org/10.1145/3371382.3378278`
PEARL (OA): `http://hdl.handle.net/10026.1/15923`

The experiments on socially assistive robotics were conducted at Fundación Cardioinfantil Instituto de Cardiología (Bogotá, Colombia) in collaboration with Emmanuel Senft and the research group in Colombian School of Engineering Julio Garavito, under the Royal Academy of Engineering IAPP project Human-Robot Interaction Strategies for Rehabilitation based on Socially Assistive Robotics (grant IAPP/1516/137). Both of these institutions were visited for a month in April 2017 for research purposes. The author of this work designed and built the user recognition and personalisation systems used in the personalised robot condition of the study, in addition to contributing to the design of the overall study, the writing of published papers, and the presentation of the work as the first author at the 2020 IEEE International Symposium on Robot and Human Interactive Communication. Published collaborative work for the project:

Lara, J. S., Casas, J., Aguirre, A., Munera, M., Rincon-Roncancio, M., Irfan, B., Senft, E., Belpaeme, T., & Cifuentes, C. A. (2017a). Human-robot sensor interface for cardiac rehabilitation. In *2017 International Conference on Rehabilitation Robotics (ICORR)*, (pp. 1013–1018)
DOI: `https://doi.org/10.1109/ICORR.2017.8009382`
PEARL (OA): `http://hdl.handle.net/10026.1/10516`

Lara, J. S., Casas, J., Munera, M., Senft, E., Irfan, B., Gutiérrez, L. F., Pinzón, L., Belpaeme, T., Rincon, M., & Cifuentes, C. A. (2017b). Sensor interface for cardiac rehabilitation monitoring: Pilot clinical study. In *Proceedings of the IX Congreso Iberoamericano de Tecnologías de Apoyo a la Discapacidad (IBERDISCAP 2017)*
PEARL (OA): `http://hdl.handle.net/10026.1/16703`

Casas, J., Gomez, N. C., Senft, E., Irfan, B., Gutiérrez, L. F., Rincón, M., Múnera, M., Belpaeme, T., & Cifuentes, C. A. (2018a). Architecture for a social assistive robot in cardiac rehabilitation. In *2018 IEEE 2nd Colombian Conference on Robotics and Automation (CCRA)*, (pp. 1–6)
DOI: `https://doi.org/10.1109/CCRA.2018.8588133`
PEARL (OA): `http://hdl.handle.net/10026.1/13637`

Casas, J., Irfan, B., Senft, E., Gutiérrez, L., Rincon-Roncancio, M., Munera, M., Belpaeme, T., & Cifuentes, C. A. (2018b). Social assistive robot for cardiac rehabilitation: A pilot study with patients with angioplasty. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 79–80). ACM
DOI: `https://doi.org/10.1145/3173386.3177052`
PEARL (OA): `http://dx.doi.org/10.1145/3173386.3177052`

Casas, J., Irfan, B., Senft, E., Gutiérrez, L., Rincon-Roncancio, M., Munera, M., Belpaeme, T., & Cifuentes, C. A. (2018c). Towards a SAR system for personalized cardiac

rehabilitation: A patient with PCI. In *2018 ACM/IEEE International Conference on Human-Robot Interaction Personal Robots for Exercising and Coaching workshop*. ACM
URL: https://aiweb.techfak.uni-bielefeld.de/hri2018_workshop_robot_coach/paper/PREC2018_paper_5.pdf
PEARL (OA): http://hdl.handle.net/10026.1/16701

Irfan, B., Céspedes Gomez, N., Casas, J., Senft, E., Gutiérrez, L. F., Rincon-Roncancio, M., Munera, M., Belpaeme, T., & Cifuentes, C. A. (2020a). Using a personalised socially assistive robot for cardiac rehabilitation: A long-term case study. In *29th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, (pp. 124–130). IEEE
DOI: https://doi.org/10.1109/RO-MAN47096.2020.9223491
PEARL (OA): http://hdl.handle.net/10026.1/16697

Additional work was conducted on the effects of social facilitation, but it was not included in the thesis. The author equally contributed in the experimental design, execution, data analyses and writing, in addition to leading this research work, conducting the literature review, and presenting the work as the first author of the published paper at the 2018 ACM/IEEE International Conference on Human-Robot Interaction:

Irfan, B., Kennedy, J., Lemaignan, S., Papadopoulos, F., Senft, E., & Belpaeme, T. (2018a). Social psychology and human-robot interaction: An uneasy marriage. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 13–20). ACM
DOI: https://doi.org/10.1145/3173386.3173389
PEARL (OA): http://hdl.handle.net/10026.1/12299

The author coordinated and co-organised "Personalization in Long-Term Human-Robot Interaction" workshop[1] at the 2019 ACM/IEEE International Conference on Human-Robot Interaction:

Irfan, B., Ramachandran, A., Spaulding, S., Glas, D. F., Leite, I., & Koay, K. L. (2019). Personalization in long-term human-robot interaction. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (pp. 685–686). IEEE
DOI: https://doi.org/10.1109/HRI.2019.8673076
PEARL (OA): http://hdl.handle.net/10026.1/13844

Additionally, the author collaborated in a study for evaluating speech recognition for child-robot interaction, but this work was not included in this thesis. The author contributed to the analysis and writing of the results for the published study:

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., & Belpaeme, T. (2017). Child speech recognition in human-robot interaction: Evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 82–90). ACM
DOI: https://doi.org/10.1145/2909824.3020229
PEARL (OA): http://hdl.handle.net/10026.1/9152

Ghent University (Belgium) was visited for research purposes under the supervision of Prof. Tony Belpaeme. In addition, the data-driven dialogue models were trained and

[1] Proceedings available at: https://longtermpersonalizationhri.github.io/

evaluated at the Ghent University IDLab cloud servers for 6 months (February to August 2020).

The author conducted a four-months research internship as a Research and Development Lab Associate between July and November 2019 under the supervision of Dr. James Kennedy at Disney Research, Walt Disney Imagineering Research and Development, Inc., Los Angeles, California (United States of America). This collaborative work was published and presented by the author at the 2020 ACM International Conference on Intelligent Virtual Agents, but it was not included in this thesis:

Irfan, B., Narayanan, A., & Kennedy, J. (2020c). Dynamic emotional language adaptation in multiparty interactions with agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20. New York, NY, USA: Association for Computing Machinery
DOI: `https://doi.org/10.1145/3383652.3423881`
PEARL (OA): `http://hdl.handle.net/10026.1/16700`

Word count of main body of thesis: **75535**

Signed: Bahar Irfan

Date: 4 December 2020

Abstract

MULTI-MODAL PERSONALISATION IN LONG-TERM HUMAN-ROBOT INTERACTION
Bahar Irfan

While short-term interactions with robots benefit from the novelty effect, long-term interactions can suffer from a decrease in user interest and engagement. Based on the previous research within Human-Robot Interaction (HRI), the thesis presented here is that user experience in long-term human-robot interactions can be improved by personalising the interaction through recognising users and recalling previously learned information.

User recognition is the first step towards personalising the interaction, however, there does not exist a reliable user recognition method for fully autonomous user recognition in long-term HRI for real-world applications. Correspondingly, this thesis proposes a Multi-modal Incremental Bayesian Network (MMIBN) model, which combines face recognition with soft biometrics and allows continuous, incremental and online learning of users, without the need for any preliminary training. We validated the robustness and reliability of this approach with a long-term (4-weeks) real-world study with 14 users and an artificially generated multi-modal long-term user recognition dataset with 200 users.

Following on from this work, we explored personalisation of the interaction in service robotics and socially assistive robotics, based on earlier evidence for the impact of personalisation on long-term interactions. We created the text-based Barista Datasets that contain simulated generic and personalised dialogues for interactions with a barista that recall and suggest user preferences in subsequent interactions in a coffee shop. Based on these datasets, we designed fully autonomous barista robots with MMIBN, automatic speech recognition (ASR) and a rule-based dialogue manager, and evaluated these robots with a real-world long-term (5-day) study with 18 non-native English speakers. The study demonstrated that personalisation mitigates negative user experiences that arise from unreliable speech recognition and the inflexible structure of the rule-based dialogue manager. Consequently, we explored the potential of the state-of-the-art data-driven dialogue models based on the Barista Datasets. The results showed that while data-driven models perform remarkably well in generic task-oriented dialogue, no model could perform sufficiently well for personalisation in long-term interactions.

Lastly, to demonstrate the real-world benefits of long-term HRI, we design a personalised robot to improve user motivation and adherence to the cardiac rehabilitation programme, and evaluate with a study that ran for 2.5 years at a hospital in Colombia. The robot individually tracked the patients' health progress and attendance throughout the programme, and provided personalised and immediate feedback based on continuous monitoring for 18 weeks. While the study could not be completed due to the outbreak of COVID-19, our initial findings (with 6 patients) showed that user engagement and motivation for the therapy and adherence were improved and maintained in the long-term interactions.

Overall, the work undertaken provides supporting evidence for our thesis and contributes fundamental stepping stones for future research in personalised long-term HRI to develop robots that can meet and maintain user expectations.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Long-term interaction is fundamental in many fields, such as education, rehabilitation, work environments and domestic applications. These fields also offer the opportunity to embed robots in assistive roles (Goodrich & Schultz, 2007; Leite et al., 2013; Bartneck et al., 2019). However, such application areas in the real world require reliable, durable and autonomous solutions for achieving feasible Human-Robot Interaction (HRI) over time. Unlike short-term interactions which benefit from the "novelty effect", long-term interactions can suffer from a decrease in user interest and engagement as the product ceases to be new to the user (Huttenrauch & Eklundh, 2002; Kanda et al., 2004; Salter et al., 2004; Gockley et al., 2005; Kanda et al., 2007; Sung et al., 2009; Leite et al., 2013). This arises from the 'one size fits all' approach based on a fixed set of behaviours, which is not suitable for repeated interactions. People have individual needs, likes and dislikes, preferences and personalities, hence adaptable systems are needed to learn from users and personalise the interaction. Personalisation, in addition, can help improve user engagement, and create a sense of familiarity over time to facilitate rapport and trust between the user and the robot (Dautenhahn, 2004; Bickmore & Picard, 2005; Kanda et al., 2010; Leite et al., 2013; Irfan et al., 2019).

Communication through verbal and non-verbal means is vital in interaction with humans (Mavridis, 2015). Conversational Artificial Intelligence (AI) refers to the speech and text-based systems, also known as *conversational agents* (e.g., virtual assistants and chatbots), that automate communication and create personalised experiences. Personal assistants such as Siri[1], Google Assistant[2], and Alexa[3] are increasingly being used for tasks

[1] `https://www.apple.com/siri/`
[2] `https://assistant.google.com`
[3] `https://developer.amazon.com/en-US/alexa`

such as calling or texting, querying news, weather or recipes, or playing music. However, these systems rely on uni-modal input in an interaction, such as text or speech. Hence, they lack the additional information that can improve the personalisation of the experience, such as visual input which helps to identify users, estimate their emotional states, and integrate the interaction context into the conversation. Robots are ideal platforms with multiple sensors to enable multi-modal communication and personalisation.

This thesis focuses on multi-modal personalisation in long-term HRI for real-world applications. As the first step towards personalisation, a multi-modal user recognition method is developed for enabling reliable autonomous and incremental recognition in long-term interactions. Two real-world studies are designed for long-term interactions in customer-oriented service and health-care domains. Subsequent interactions are personalised by recalling user attributes, preferences or behaviour patterns through multi-modal communication.

1.1 Scope and Key Concepts

HRI covers an extensive research area, as outlined in Goodrich & Schultz (2007). This research focuses on exploring multi-modal personalisation in long-term HRI for real-world applications. More specifically, this work addresses user identification for real-time HRI over extended periods of time, and personalised conversational agents as service robots and socially assistive robots. The following subsections define the scope of our research focus and the key concepts in more detail, accompanied by the justifications of our restrictions where necessary.

1.1.1 Long-Term Human-Robot Interaction

In addition to helping design systems for long-term HRI in a variety of fields, long-term interaction studies allow observing changes in user behaviour and experiences over time, and the true impact of an offered solution (Kanda et al., 2004; Leite et al., 2013). The duration for a study to be considered as *long-term interaction* varies in the literature (ranging with studies lasting 25 minutes (Salter et al., 2004) to six months (Sung et al., 2009)), due to several factors, such as the number of interactions with a robot, the length of the interaction, the number of users interacting with the robot simultaneously, and the complexity of the robot's behaviour (Leite et al., 2013). Leite et al. (2013) define long-term interaction as follows:

> *"An interaction can be considered as 'long-term' when the user becomes familiarised with the robot to a point that her perception of such robot is not biased by the novelty effect anymore."*

Based on this definition, we designed long-term user studies with varying durations depending on the complexity and the goal of the task: five consecutive days (Chapter 7), four weeks (Chapter 4), and 18 weeks with two sessions per week (Chapter 9).

1.1.2 Personalisation and Long-Term Memory

Blom (2000) suggests that users or designers personalise their products primarily to "facilitate the work" (i.e., for "enabling access to information content", "accommodating work goals" or "accommodating individual differences") or "accommodate social requirements" (i.e., for "eliciting an emotional response" or "expressing identity"). Based on this taxonomy, our motivation for personalisation in this thesis falls under facilitating the work through accommodating individual differences and work goals, particularly by tracking the progress of patients to increase the user engagement in the therapy (Chapter 9) and improving the user experience and increasing the robot efficiency through decreasing the time it takes to take an order by recalling previous orders (Chapter 6).

In order to create personalised HRI, certain type of information should be acquired from the user and recalled in the subsequent interactions. Based on the previous research for adaptation in natural and synthetic systems, Wood et al. (2011) suggest using a biologically-inspired *long-term memory* that contains semantic (e.g., name, preference), as well as, episodic information ("who, what, where, when"), by using contextual, spatial, and temporal information about previous interactions. Acquiring a diverse set of facts for achieving a higher level of personalisation requires making use of various sensors, thus, creating the necessity for multi-modal personalisation. Consequently, in this thesis, we rely on visual, audio, sensory and text-based input to acquire and recall semantic and episodic information.

Humans are capable of learning incrementally, that is, they can expand their knowledge over time, which is a desirable feature for personalisation in long-term HRI because new users or new attributes might need to be learned. However, *incremental learning* is not sufficient to achieve fully adaptable systems, because preferences and appearances of known users can change over time. For example, if a user gets a hair cut or starts wearing glasses, user recognition might fail to recognise the user, and this would persist over time because the model is not updated for this user. In comparison, humans can continuously adapt to changing circumstances by updating their prior beliefs. Such adaptation can be made possible by *online learning (OL)* in robots, in which the model is updated sequentially with incoming information. We apply incremental and online learning throughout this thesis for learning new users as well as updating the models.

The information acquired during an interaction can be stored and recalled by two types of approaches: *knowledge-base (KB)* and *data-driven*. KB is a structured database with entities and their corresponding values and relations; whereas, data-driven approaches rely on extracting the structures and values from the data itself. KB approaches are suitable for long-term interactions, because the database can be expanded and updated. However, only data-driven approaches that rely on *end-to-end* learning allow incremental and online learning, because the system is trained from input-to-output as a whole. Thus, KB approaches are more robust because they rely on explicit storage of data; whereas data-driven methods are more flexible because *a priori* or structured knowledge in terms of rules or templates are not necessary to store or use (Gao et al., 2019; Yan, 2018). In this thesis, we use a combination of KB and end-to-end data-driven approaches in our architectures to account for this trade-off.

1.1.3 Multi-modal User Recognition

Deploying robots in the real world for long-term interactions requires a high level of autonomy. In other words, tele-operated robots or the Wizard-of-Oz (WoZ) method, which rely on a human operator to control the robot unbeknownst to the user, are not suitable for real-world applications. Consequently, for achieving personalisation in long-term interactions, the robot should autonomously recognise users without intrusive methods or external devices, such as QR tags or access cards. This requires recognising and updating the users continuously and incrementally and possibly starting from a state without any known user.

Face recognition (FR) has been the technique that is most prominently used for biometric user identification due to its non-intrusive character. Most state-of-the-art FR methods are based on deep learning approaches (e.g., Taigman et al. (2014); Sun et al. (2014); Parkhi et al. (2015); Schroff et al. (2015)), however, these techniques are made for recognising users that are already in the user database. Only a few deep learning techniques exist for recognising unknown users (Bendale & Boult, 2016; Ge et al., 2017). Moreover, these models are not suitable for incrementally learning unknown users due to the *catastrophic forgetting* problem, which refers to the drastic loss of performance on previously learned classes when a new class is introduced (McClelland et al., 1995; McCloskey & Cohen, 1989; Parisi et al., 2019). One solution to overcome this problem is by re-training the network after the introduction of a novel user. However, this requires storing the previous samples, which could create a prohibitive computational burden in long-term deployments. Furthermore, it would require a significant amount of time to retrain with a growing number of users and samples (Bendale & Boult, 2015), especially on low computational power systems such as robots, which makes them unsuitable for real-time HRI. Thus, we need systems that allow scaling and support incremental learning of new classes in addition to recognising previously learned classes, which is termed *open world recognition* (Bendale & Boult, 2015). There exist a few approaches that were designed for open world recognition (Bendale & Boult, 2015; De Rosa et al., 2016; Fei et al., 2016; Rudd et al., 2018), however, none of these approaches is applied to user recognition.

Relying solely on FR can result in recognition failures due to inaccuracies in the data, such as low lighting conditions or blurry images (Wójcik et al., 2016). This problem, along with the incorrect recognitions arising from the similarities between users, can be overcome by combining multi-modal sources of information, similar to how humans recognise each other (e.g., using voice to recognise a person in a dark room). For example, ancillary physical or behavioural characteristics, called *soft biometrics (SB)*, such as age and gender, can be used in combination with *primary biometrics* (i.e., biometrics that can help uniquely identify a person, such as FR or fingerprints), or other SB to improve the recognition performance (Jain et al., 2011; Dantcheva et al., 2016; Scheirer et al., 2011). Most robots, due to the rich sensor suite they carry, lend themselves well to multi-modal recognition.

One architecture that has been used in previous research that allows combining multi-modal information for recognition is a Bayesian network (BN) (Scheirer et al., 2011). BN is a probabilistic graphical model which represents conditional dependencies of a set of variables through a directed acyclic graph. However, typically BNs are created with a conditional probability table based on *a priori* knowledge of the dependencies, termed *likelihoods*, of various sources of information. In the case of missing *a priori* data, e.g., for new users or with changes in the user appearance, the assumptions for dependencies might not hold, which, in turn, would affect the overall recognition performance. Nevertheless, it is possible to achieve incremental and online learning with BNs by updating the likelihoods of the network (Bauer et al., 1997; Cohen et al., 2001b,a; Lim & Cho, 2006; Liu & Liao, 2008).

Following these aspects, this thesis aims to design a multi-modal user recognition system for long-term HRI using an incremental BN with online learning, as described in detail in Chapter 3 and evaluated in Chapter 4 and 5.

1.1.4 Conversational Artificial Intelligence

This thesis focuses on the social perspective of long-term HRI. Breazeal (2002) define robots that communicate with and understand the responses of the users in a personal way as *sociable robots*. Communication in HRI may refer to verbal or non-verbal interaction, such as nonlinguistic utterances, body movements, facial expressions, colour and eye gaze (see works by Bethel & Murphy (2008); Mavridis (2015); Saunderson & Nejat (2019) for extensive surveys in the literature). However, non-verbal interaction is not sufficient to convey various type of information. For example, while nonlinguistic utterances similar to the beeping sounds of R2D2 in the Star Wars movies may result in the categorical perception of the sounds, and the subtle differences between utterances might not be understood by users (Read & Belpaeme, 2012, 2016). Nevertheless, non-verbal interaction is an important aspect of human-human communication, and it can increase efficiency and robustness of the interaction when used in conjunction with verbal information. In order to achieve a natural HRI with mutual understanding, it is desired to have both verbal and non-verbal communication (Mavridis, 2015). Moreover, HRI poses a challenge over uni-modal interaction due to the robot's embodiment, with the user assuming multi-modal capabilities based on the various sensors of the robot (e.g., camera, microphones, speakers, tablet) (Goodrich & Schultz, 2007; Rickert et al., 2007). In this thesis, we focus on multi-modal natural language interactions based on speech, by making use of available technologies for natural language processing (NLP), such as automatic speech recognition (ASR), natural language understanding (NLU), natural language generation (NLG) and text-to-speech (TTS), in addition to acquiring sensory information and using touchscreen interfaces for text or image-based interaction to reduce the errors in ASR, wherever necessary. We use non-verbal communication, such as the body movements[4], gaze and touch-based interaction to increase the naturalness of the interaction. Natural language interactions in HRI can refer to the uni-directional exchange of information, e.g., the human instructing the robot or the robot instructing the human, or bi-directional interaction, such as a dialogue. In this thesis, we examine bi-directional interaction that involves multi-modal exchange of information between the robot and the user.

Conversational agents that are restricted to conversing about a narrow domain are called closed-domain dialogue systems. In contrast, open-domain dialogue systems can converse about a variety of topics. These dialogue systems can be further categorized for their application purpose: *general-purpose* or *task-oriented* (goal-oriented or goal-driven) dialogue systems. Task-oriented dialogue systems can only do a particular task (or a small set of tasks), such as booking a restaurant, whereas general-purpose dialogue systems are capable of performing any task, such as "chit-chat" (Yan, 2018). Designing conversational agents that are similar to humans requires using open-domain and general-purpose dialogue systems, however, building such systems is extremely challenging (Gao et al., 2019). Moreover, the ASR systems are not reliable enough to understand open-domain spoken dialogue in HRI. In addition, such systems are not necessary in most of the real-world application areas in HRI (e.g., education, rehabilitation, customer service, search and rescue). For example, a customer service robot deployed at a coffee shop only needs to know the menu and request information from the customers accordingly. Hence, for such applications, task-oriented closed-domain dialogue systems are sufficient and more appropriate.

Traditional conversational AI relies mostly on *slot-filling* methods through rule-based dialogue management systems (RBDMSs) based on predefined if-then-structures and templates with a set of slots to be filled during the dialogue, which are the basis of most commercial systems (Yan, 2018; Gao et al., 2019; Bordes et al., 2016). For example, the Mitsuku[5] chatbot, which is the five-time winner of the Loebner Prize Turing Test in general-purpose open-domain dialogue, is based on Artificial Intelligence Markup Language (AIML), which is a rule-based language written on XML. The Turing test (Turing, 1950)

[4]Using animated speech feature of NAOqi: `http://doc.aldebaran.com/2-4/naoqi/audio/alanimatedspeech-api.html`

[5]`https://www.pandorabots.com/mitsuku/`

evaluates the dialogue system's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from that of a human, based on general-purpose open-domain dialogue. The evaluator needs to understand if the conversation partner is a machine or a human, based on text-based interactions through a computer. In the recent years, the research (especially by Google, Facebook, IBM, Microsoft and Amazon) is shifting towards creating data-driven conversational agents (e.g., neural approaches) to compete with RBDMSs, in order to create more flexible systems that do not require any feature engineering or domain-specific handcrafted rules (e.g., Sutskever et al. (2014); Graves et al. (2014); Sukhbaatar et al. (2015); Rajendran et al. (2018); Shum et al. (2018); Ram et al. (2018)). Only recently Google's Meena chatbot, which is an end-to-end neural conversational model, outperformed Mitsuku (Adiwardana et al., 2020), and Facebook AI's Blender outperformed Meena (Roller et al., 2020). However, they both used an excessive amount of computational power to train the model[6], which has incited a discussion both in the fields of NLP[7] and HRI that defended the advantages of using RBDMSs for their low computational power requirement and the lack of necessity for training, which are especially crucial for robots. Even though, data-driven approaches require smaller datasets in training for task-oriented closed-domain dialogue, variations that can arise from user utterances, as well as, the differences between user needs can be challenging for achieving personalisation in long-term interactions, thus requiring sufficient data from each individual or methods that allow transferring common dialogue knowledge between users (Mo et al., 2016). Using an RBDMS, allows structuring the interaction through a set of rules, such that the relations between users need not be learned or transferred. However, deploying robots in the real world brings about the challenges of ASR errors, compared to text-based approaches that benefit from a more robust NLP, which decrease the robustness of RBDMSs. Moreover, rule-based approaches require the user to respond in a particular manner, which can be time consuming and frustrating (Williams et al., 2018; Bartneck et al., 2019).

At the time of writing, only a few studies (Kasap & Magnenat-Thalmann, 2012; Zheng et al., 2019; Churamani et al., 2017) explored fully autonomous personalisation in dialogue

[6]Meena is a neural conversational model with 2.6 billion parameters trained on a TPU-v3 Pod (2048 TPU cores) for 30 days on 341 GB of text, filtered from public domain social media conversations (Adiwardana et al., 2020). Blender is a neural conversational model with 9.4 billion parameters trained on filtered public domain social media conversations (1.50B comments), and fine-tuned with a blend of Wikipedia, crowd-sourced conversations based on personality or grounded in an emotional situation (a total of 76k utterances) (Roller et al., 2020).

[7]See, for example: `https://twitter.com/eturner303/status/1223976313544773634`

for long-term HRI. However, none of these studies was conducted in the real world. In this thesis, we explore both the applicability of a RBDMS (Chapter 7) and data-driven approaches (Chapter 8) to real-world long-term interactions with task-oriented closed-domain dialogue that focuses on individually personalising HRI over time.

1.1.5 Real-World Applications

It is important to find an application area for long-term HRI, where personalisation can be valuable in the real world. There exists several studies that show the advantages using social and assistive robots in influencing, educating or training people, especially in long-term interactions (Huttenrauch & Eklundh, 2002; Kanda et al., 2004; Severinson-Eklundh et al., 2003; Robins et al., 2004; Werry et al., 2001). There is growing evidence that many long-term interactions require (mutual) adaptation (de Ruyter et al., 2005; Severinson-Eklundh et al., 2003; Robins et al., 2004; Forlizzi & DiSalvo, 2006; Goodrich & Schultz, 2007). Moreover, it has been shown that people adapt to robots in the service industry over the long-term over a wide range of tasks (Green & Eklundh, 2003; Sung et al., 2009), and personalisation of conversational agents in healthcare and medicine is effective in increasing user engagement and user task performance (Matarić et al., 2007; Kocaballi et al., 2019). Based on these aspects, we explored two areas of application: service robotics and socially assistive robotics.

A *service robot* is a robot that performs useful tasks for humans or equipment excluding industrial automation applications, based on its current state and sensing without human intervention[8]. Service robots exist in a wide range of areas, ranging from domestic use to deployment for customer service. This thesis focuses on the use of customer-oriented service robots (Chapter 7), specifically as an order-taking robot in a coffee shop which recognises users and recalls the user's previous orders in subsequent interactions, similar to a barista in a local coffee shop.

Socially assistive robotics (SAR) refers to the assistive and supportive robotics applications in social interactions (Feil-Seifer & Matarić, 2005). This field faces several challenges, such as the proximity and vulnerability of the human in the interaction, the potential of unanticipated patterns and a noisy real-world environment (Goodrich & Schultz, 2007). Due to these challenges, it is important to structure social interaction, such that the therapy

[8]Definition from the International Organization for Standardization (ISO 8373): `https://www.ifr.org/service-robots/`

is not negatively affected. Moreover, because the robot will be deployed in the real world with non-expert users (e.g., doctors, nurses, patients), thus, it should not require an operator or extensive training (Feil-Seifer & Matarić, 2005). In other words, it should be autonomous and require minimum effort from users and medical staff. In addition, it must be designed in accordance with the requirements of the therapy in collaboration with a medical team and it must conform to the changing routines and demands of the patient and carers (Matarić & Scassellati, 2016). Moreover, the ultimate goal of SAR is to alter the long-term behaviour of the user in accordance with the behavioural, therapeutic, or educational goals that the robot was designed to support. SAR is used in a variety of therapy fields, such as a therapeutic tool for children, the elderly, stroke patients, rehabilitation and other special-needs populations requiring personalised care (Matarić & Scassellati, 2016).

Cardiac rehabilitation (CR) is a therapy used to prevent cardiovascular disease or to treat a patient after a post-cardiovascular event. One of the most critical issues of CR is the lack of adherence of the patients to the therapy process. However, there is evidence that an embodied agent can increase compliance and adherence to the therapy (Deng et al., 2019). Based on these findings in the literature, we designed a multi-modal personalised socially assistive robot to increase patients' engagement, motivation and adherence to the long-term CR programme in a hospital (Chapter 9). The robot processes the sensory information from the patient to track the patient during the session and through the programme, as well as to recall the previous sessions to provide personalised feedback on the patient's health and progress throughout programme.

1.2 The Thesis

The main thesis that this document seeks to put forward is as below.

> User experience in long-term human-robot interactions can be improved by personalising the interaction through recognising users and recalling previously learned information.

This thesis raises a series of additional research questions (RQ) that shaped our objectives and evaluation approaches. As defined in the main thesis and previously outlined in the scope of this work, we explore personalisation of the interaction through user recognition (concerning RQ1) and acquiring and recalling of information (RQ2-4) for application in the real world (RQ5-6).

- RQ1: *Which user recognition algorithms are applicable to long-term recognition in the real world?*

  As previously established in this chapter, long-term HRI requires a high level of autonomy in real-world applications. Hence, in order to personalise the interaction, the users should be continuously and incrementally recognised without intrusive methods or external devices. However, as further addressed in Section 2.2, a reliable user recognition algorithm does not exist for open world user recognition in long-term HRI, which brought upon the following research objective (RO) in this thesis:

- RO1: **Build a user recognition algorithm suitable for fully-autonomous long-term HRI in the real world that allows incremental and online learning of users.**

  Based on this objective, we created a multi-modal incremental Bayesian network with online learning for user recognition, as described in Chapter 3 and evaluated in Chapter 4 and 5.

- RQ2: *How should the robot communicate with users to acquire and convey information?*

  As presented in Section 1.1.4, this thesis focuses on natural language interactions through verbal and non-verbal communication. Based on this research question and the motivation to achieve naturalness in the interaction that relies on mutual understanding, we explore bi-directional interaction for our user studies in which the robot requests, receives and delivers information.

- RQ3: *Which dialogue architectures are appropriate for long-term interactions in the real world?*

  Conversational Artificial Intelligence is a broad research topic ranging from task-oriented closed-domain to general-purpose open-domain systems using rule-based dialogue management systems to data-driven approaches, as highlighted in Section 1.1.4. This research question explores the rule-based and data-driven state-of-the-art approaches described in Section 2.3, and evaluates their potential for real-world interactions in Chapter 7 and 8.

- RQ4: *Which type of information should be recalled for personalisation?*

Inspired by the suggestion of Wood et al. (2011), we obtain and recall both semantic and episodic information for personalising long-term interactions. However, the specific information that should be acquired depends on the context of the application, which brought upon the following research question:

- RQ5: *Which are the real-world application areas where personalisation can make an impact in long-term interactions?*

  As outlined in Section 1.1.5, we identified that customer-oriented service robotics and socially assistive robotics fields are suitable to evaluate our research questions for long-term interactions in the real world. Consequently, we developed the following objectives:

- RO2: **Design a personalised customer-oriented service robot to improve user experience and increase the efficiency of the task.**

  We designed a barista robot that recalls the previous orders of a user, in order to create the personalised experience of a local coffee shop, in addition to decreasing the number of turns necessary for an order. We initially designed the interaction by generating text-based Barista datasets, as described in Chapter 6. Correspondingly, we designed a robot with a rule-based dialogue management system in Chapter 7. We explored the applicability of data-driven approaches in Chapter 8.

- RO3: **Design a personalised socially assistive robot to increase user engagement, motivation and adherence to the therapy.**

  We collaborated with the Colombian School of Engineering Julio Garavito and doctors and therapists in Fundación Cardioinfantil Instituto de Cardiología (Bogotá, Colombia) in designing a personalised socially assistive robot to be applied in cardiac rehabilitation programme. The designed system, our contributions and the real-world study are described in detail in Chapter 9.

  In both of these applications, we used the user recognition system that we developed for our first objective (RO1) for identifying users.

- RQ6: *What is the impact of personalisation in long-term human-robot interaction?*

  We evaluated the impact of personalisation through real-world studies, as described in Section 7 and Section 9.

1.3 Research Overview

The thesis and the research questions are explored in this document through building and designing a series of systems applicable to long-term interactions, followed by experimental evaluations in the real world. Initially, a literature review was conducted in long-term human-robot interaction, long-term memory systems, personalisation, and real-world application areas for personalised robots (Section 2.1), followed by user recognition (Section 2.2) and conversational agents (Section 2.3) to define the scope of this work and find answers to our research questions.

1.3.1 User Recognition

Based on our findings, multi-modal user recognition was identified to provide reliable identification that would overcome problems in real-world applications. Moreover, as described in Section 1.1.3, incremental and online learning were determined to be vital for autonomous long-term interactions. Due to the lack of an existing model that satisfied these constraints, we built a multi-modal incremental Bayesian network with online learning for identifying users, which is the first in combining soft biometrics with a primary biometric for open world user identification in real-time HRI, as detailed in Chapter 3.

We validated our system in a real-time HRI scenario with 14 participants (10 males, 4 females, with age range of 24-40) and collected a total of 66 images per user over four weeks period, as described in Chapter 4. We used a Pepper[9] robot (SoftBank Robotics Europe) with NAOqi[10] software modules that provided the face recognition similarity scores, along with gender, age, and height estimations that we used as input modalities, in addition to the time of interaction. In order to obtain reliable ground truth values, the participants enrolled by entering their name, gender, age, and height through the tablet interface of the robot, followed by a picture taken by the robot. This enrolment process occurred within the first week of the study for the participants, that is, the participants were not enrolled at the same time, which resembled a real-world situation. The robot was placed in the kitchen of an office of the Centre for Robotics and Neural Systems (CRNS) at the University of Plymouth, as shown in Figure 1.1. The participants were PhD

[9] https://www.softbankrobotics.com/emea/en/pepper
[10] http://doc.aldebaran.com/2-5

Figure 1.1: A user is interacting with a Pepper robot during the user recognition study[11].

students or researchers working in the office, or visiting it frequently. The participants interacted with the robot at the beginning and the end of the day and throughout the day whenever they wished (mostly during their coffee or lunch breaks). This enabled us to achieve a natural level of interaction similar to the real world. At each interaction, the robot would autonomously recognise the user, and request from the user a confirmation of the estimated identity. The robot communicated its requests and feedback verbally, but the participants interacted with the robot only through the tablet interface to ensure reliable data collection. This study showed that the proposed model is suitable for real-world human-robot interaction experiments for user recognition in real-time, in addition to enabling us to optimise the parameters of the Bayesian network.

However, we could not generalise the results obtained from the user study to larger populations due to the limited population size and the narrow age range of the users. Obtaining a dataset which captures a diverse set of characteristics for a large number of users over long-term interactions is a laborious task in HRI. Thus, we created a multi-modal long-term user recognition dataset (Chapter 5) based on the images of 200 celebrities obtained from the IMDB-WIKI dataset (Rothe et al., 2015, 2018), which contains images taken at events or still frames from movies. We used proprietary algorithms of the Pepper robot to obtain multi-modal biometric information from these images (namely, face recognition scores for similarities between users, and gender and age estimates), while the height and time of interaction were artificially generated to simulate a long-term HRI scenario similar to the one in our earlier work. We defined two datasets of varying sizes: (1) where each user is observed precisely ten times, e.g., ten return visits to a robot

[11] Video demonstration for human-robot interaction within the study: `https://youtu.be/Ix98k6_-2Zc`

therapist, and (2) where each user is encountered a different amount of times (10 to 41 times). Moreover, we defined two sets of timing: (1) patterned interaction times, where the user will be encountered certain times on specific days similar to HRI in rehabilitation and education areas, and (2) random interaction times, such as in domestic applications with companion robots, in which it is likely to encounter the user at any time of the day. We evaluated the proposed multi-modal incremental Bayesian network with and without online learning in comparison to face recognition, soft biometrics and a state-of-the-art open world recognition method Extreme Value Machine (EVM).

1.3.2 Conversational Artificial Intelligence

As previously stated in the scope of this thesis, we used task-oriented closed-domain dialogue within a customer-oriented service robot context as an order-taking barista in a coffee shop for evaluating the impact of personalisation in long-term recognitions.

Due to the lack of available corpora for human-human interaction or human-robot interaction and the challenges in collecting thousands of interactions, it was necessary to generate artificial datasets for coffee shop interactions with a barista for ordering drinks and accompaniments. Subsequently, three text-based datasets were generated, as described in Chapter 6: (1) *Barista Dataset*, (2) *Personalised Barista Dataset*, (3) *Personalised Barista Dataset with Preferences Information*. The *Barista Dataset* is based on the transaction between the barista and a customer similar to that in a coffee shop chain: the drink and accompaniment order is taken, confirmed and changed if necessary, the customer's name is taken and used to note the location of the order, followed by a goodbye phrase. In the *Personalised Barista Dataset*, the customer is recognised, and the previous most common (or most recent) order of the customer is recalled to ask if the customer would prefer to have that again, thereby, decreasing the number of conversational turns necessary to make an order. The identified user information is provided by the identification number and the name of the user, which simulates the information obtained from a user recognition system in HRI. This dataset allows training and evaluating data-driven approaches to extract orders from the data when the user is known. The final dataset, *Personalised Barista Dataset with Preferences Information*, provides the most common (or most recent) order information along with the user information to be able to compare the performance of data-driven approaches with the same information in rule-based dialogue management systems (RBDMS).

Based on these datasets, we designed non-personalised and personalised barista robots with RBDMS in Chapter 7. The RBDMS relies on template matching and dialogue state tracking to match the user responses to the phrases in the rules used to create the *Barista Dataset* and *Personalised Barista Dataset*. We combined the user recognition system that we developed in Chapter 3 with NAOqi voice activity detection[12] and Google Cloud Speech-to-Text[13] for online automatic speech recognition (ASR). We optimised the ASR to be adaptable to non-native speakers on an audio dataset that we collected through 12 non-native English speakers (with slightly accented English, but with high English proficiency levels) reading the same five monologues from *Personalised Barista Dataset* to the robot. We used the *Adapted Pepper*[14] robot (shown in Figure 1.2), which has an improved microphone system with higher signal-to-noise ratio compared to an off-the-shelf robot.

We conducted a 5-day real-world study in the coffee bar of an international student campus, Cité Internationale Universitaire de Paris, with 18 non-native English speakers

Figure 1.2: *Adapted Pepper* is taking the order of a user as a barista robot[15].

[12]http://doc.aldebaran.com/2-5/naoqi/audio/alspeechrecognition-api.html
[13]https://cloud.google.com/speech-to-text
[14]Created for MuMMER project: http://mummer-project.eu.

(11 males, 7 females) within the age range of 22-47, as described in Chapter 7. We compared three conditions: (1) enrolment, (2) non-personalised robot and (3) personalised robot. Enrolment is the first interaction with the robot, which is non-personalised. The participants interacted with the same condition during the duration of the study.

Furthermore, we explored the applicability of the state-of-the-art data-driven approaches highlighted in Section 1.1.4 and described in detail in Section 2.3. We evaluated the approaches on the three generated artificial text-based datasets with different sizes of datasets, as described in Chapter 8 to observe the performance of the approaches based on the provided information, size of the datasets and varying tasks.

1.3.3 Socially Assistive Robotics

For evaluating long-term HRI in an assistive real-world application, we designed a personalised socially assistive robot for cardiac rehabilitation programme which has been used throughout the therapy of patients in Fundación Cardioinfantil Instituto de Cardiología (Bogotá, Colombia). We used a NAO[16] robot (SoftBank Robotics Europe) in our study, along with a touchscreen interface and a sensory interface to receive feedback from the patient and obtain medical measures, as shown in Figure 1.3.

In the study, we designed three conditions: (1) *control*, (2) *social robot* and (3) *personalised robot*. In each of these conditions, a sensory interface obtains the patient's heart rate, posture, cadence, step length and speed and the inclination of the treadmill. In addition,

Figure 1.3: Socially assistive robot setup for cardiac rehabilitation programme at Fundación Cardioinfantil Instituto de Cardiología (Bogotá, Colombia).

[15] Demonstration of the robot behaviours is available at: https://youtu.be/eAOnHlDuHqw
[16] https://www.softbankrobotics.com/emea/en/nao

a touchscreen interface is used to request the exertion level of the patient. In the *control* condition, there is only a touchscreen that provides online and continuous monitoring and visualisation of the obtained sensory information. The tablet does not provide any verbal feedback to the patient in order to closely resemble the conventional CR programme. A socially assistive robot is used in the *social robot* condition to provide immediate feedback if any of the sensory values exceed the given limits, motivate the patient throughout the session and alert the doctor in case of emergencies. In the *personalised robot* condition, the patient is recognised through the user recognition system defined in Chapter 3. The name of the patient is used periodically throughout the session to personalise the content of the feedback provided in the other robot condition. Moreover, the difficulty level of the session and the performance of the patient is compared to the previous sessions to motivate the patient for the therapy and the upcoming sessions, at the beginning and end of a session. In addition, the attendance of the patient is tracked to ensure that the patient is adhering to the CR programme. The study has finished for the control condition, but could not be completed for the social and personalised robot conditions due to the outbreak of COVID-19.

1.4 Contributions

This research work contributed both technically (by developing software in multiple projects) and scientifically (by evaluating the impact of the approaches) to the state-of-the-art in HRI, specifically on personalisation in long-term interactions. This section highlights the contributions of this thesis and indicates the relevant chapters and the published work, in addition to stating other contributions that are not included in the main body of the document.

1.4.1 Main Contributions

- The first contribution, and one of the cornerstones of this work, is the **design and implementation of Multi-modal Incremental Bayesian Network (MMIBN)**, which is a **multi-modal user identification system that supports online learning**. It is the **first** method for **sequential and incremental learning in open world user recognition** that allows **starting from a state without any known users**. In addition, this proposed approach is the **first in combining soft biometrics with a primary bio-**

metric for open world user identification in real-time in human-robot interaction (Chapter 3; Irfan et al. (2018b, under review)).

- We propose an online learning method for Bayesian networks based on Voting Expectation Maximization (EM) (Cohen et al., 2001a,b) and Maximum Likelihood estimation that accounts for modelling the noise in the modalities and uses an adaptive learning rate based on the frequency of user appearances. This method relies on supervised learning, through direct or indirect (e.g., through dialogue) confirmation of the identity, to adapt the likelihoods of the modalities within the Bayesian network (Section 3.2.5; Irfan et al. (under review)).

- We introduced the **quality of the estimation** and **long-term recognition performance loss** to **decrease the number of incorrect recognitions** and **create a balance between identifying known users and unknown users for long-term interactions**, respectively (Chapter 3; Irfan et al. (2018b, under review)).

- We **evaluated MMIBN in a user study for four weeks** that showed the proposed model is **applicable for real-world HRI experiments for user recognition in real-time**. In addition, the **proposed model outperformed base face recognition** in terms of **higher identification rate** (Chapter 4; Irfan et al. (2018b)).

- We **created a multi-modal long-term user recognition dataset** with 200 users of varying characteristics based on the IMDB-WIKI dataset (Rothe et al., 2015, 2018) for evaluating the model with a large number of users (Chapter 5; Irfan et al. (under review)).

- We **evaluated MMIBN with the multi-modal long-term user recognition dataset** and showed that the proposed model **significantly outperforms base face recognition, soft biometrics and a state-of-the-art approach in open world recognition**. **Online learning** was found to **decrease recognition performance compared to using a non-adaptive model** for our proposed user recognition model, which could be due to the accumulating noise of the identifiers. However, online learning was shown to **equalise the recognition performance between users**, thereby, **decreasing the biases in the system caused by face recognition** (Chapter 5; Irfan et al. (under review)).

- We **created text-based simulated Barista Datasets for generic and personalised**

task-oriented closed-domain dialogue based on interactions of an **order-taking barista in a coffee shop**, in order to train and evaluate rule-based and data-driven approaches (Chapter 6).

- We **designed a fully autonomous barista robot with user recognition, automatic speech recognition and a rule-based dialogue management system** (Chapter 7; Irfan et al. (2020b)).

- We **evaluated the barista robot** in a **real-world study with non-native English speakers for five days**, which is the **first study for fully autonomous personalisation in dialogue for long-term HRI conducted in the real world**. The study showed that **personalisation can mitigate interaction failures and the negative user experience** (Chapter 7; Irfan et al. (2020b))[17].

- We **explored** the potential of the state-of-the-art **data-driven dialogue models** in **generic and personalised long-term interactions** within **continual** and **few-shot learning** contexts based on the text-based Barista Datasets. The experiments demonstrated that a generative model, **Sequence-to-Sequence (Seq2Seq)**, achieves **near-perfect accuracy in generic long-term interactions**, however, **no model is suitable for personalised long-term interactions**. Nonetheless, a retrieval-based attention model, **Memory Network**, **shows potential**, in addition to **performing well in generic long-term interactions** (Chapter 8).

- We **designed a personalised socially assistive robot** for **cardiac rehabilitation to improve user motivation and adherence** in the **real-world long-term (18 weeks) clinical therapy of patients**. We designed a clinical study to **compare the conventional CR programme** with **a generic robot** with continuous monitoring and immediate feedback, and **to a robot with personalised feedback based on patients' progress and attendance** in the therapy. The study took place in a **hospital in Colombia** and ran for **2.5 years**. The **personalised robot** was **perceived positively throughout the programme**, the **gaze, social interaction and the compliance to the robot's requests** were **maintained over time** and the **personalisation features were appreciated** by the patients. The patients in **both robot conditions** reported that **working with a robot improved motivation to attend the therapy sessions**. In addition, the **continuous monitoring** was found to **facilitate immediate intervention**

[17] Video presentation of the study is available online: `https://www.youtube.com/watch?v=_g2H1Dk83wQ`

by the medical team **in critical situations** and enable **high-intensity training**. Multimodal user recognition with online learning was found to perform better than the non-adaptive model, when the identifiers are malfunctioning (Chapter 9) (Chapter 9; Lara et al. (2017a,b); Casas et al. (2018b,c,a); Irfan et al. (2020a)).

The work presented in this thesis has been conducted under the EU H2020 Marie Skłodowska-Curie Actions Innovative Training Networks project Applications of Personal Robotics for Interaction and Learning (APRIL), grant 674868.

The experiments on socially assistive robotics are conducted at Fundación Cardioinfantil Instituto de Cardiología (Bogotá, Colombia) in collaboration with Emmanuel Senft and the research group in Colombian School of Engineering Julio Garavito funded by the Royal Academy of Engineering IAPP project Human-Robot Interaction Strategies for Rehabilitation based on Socially Assistive Robotics (grant IAPP/1516/137). The author designed and built the user recognition and personalisation systems used in the *personalised robot* condition, in addition to contributing to the design of the overall study.

The work on barista robot, conducted at Cité Internationale Universitaire de Paris, was done in collaboration with Mehdi Hellou funded by SoftBank Robotics Europe as a research intern under the joint supervision of the author and Alexandre Mazel. The author designed the barista robot and the rule-based dialogue management system (RBDMS) architectures, created the corresponding datasets, designed the experimental procedure and the evaluation methods for the user study and conducted the analysis of the study.

1.4.2 Other Contributions

- We conducted two social facilitation studies on established tasks in the literature to observe the effects of the presence of a robot on task performance and cheating. However, we could not replicate the effects of social facilitation by humans, which supported the "replication crisis" in psychology (Aarts, 2015). Consequently, we concluded that HRI studies based on psychology experiments should be designed carefully by ensuring the replicability of previous studies and avoiding the confounding factors (e.g., Hawthorne effect and demand characteristics). We also highlighted the importance of registering studies, reporting null-results and designing reference tasks for HRI (Irfan et al., 2018a).

- We participated in and won the Social Care Challenge of EPSRC UK Robotics and Autonomous Systems Network (UK-RAS Network). The challenge consisted of designing a project proposal for using Pepper (SoftBank Robotics Europe) or Miro[18] (Consequential Robotics) robots in social care. Our proposal was using a personalised robot for reminiscence dialogues with care home residents. The winners of the challenge collaborated in organising socially assistive robotics demonstrations during UK robotics week (June 2017). The collaboration resulted in UK-RAS Social Care White Paper (Prescott & Caleb-Solly, 2017).

- A personalised robot narrator with non-verbal emotional communication was developed. The work was presented at the 9th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-Epirob 2019) workshop on Personal Robotics and Secure Human-Robot Collaboration[19].

- A number of ASR engines were evaluated for child-robot interaction under a variety of conditions on a NAO robot. The study showed that the current state-of-the-art ASR engines are not reliable in real-world applications with children (Kennedy et al., 2017).

- Chilitags (Bonnard et al., 2013) were adapted to be used with object recognition on a NAO robot in the Second Language Tutoring using Social Robots (L2TOR)[20] project (Wallbridge et al., 2017).

The social facilitation work was done in collaboration with a team of researchers at University of Plymouth. The author equally contributed to the experimental design, execution, data analyses and writing, in addition to leading this research work, conducting the literature review, and presenting the work at an international conference as the first author of the published paper.

In designing the proposal for Social Care Challenge and conducting the demonstrations during the UK robotics week, the author collaborated with Asimina Marmpena and Marta Romeo under the supervision of Prof. Tony Belpaeme and Prof. Ray Jones.

The personalised robot narrator was developed in collaboration with Asimina Marmpena. The author contributed by integrating the user recognition system (MMIBN) and imple-

[18]http://consequentialrobotics.com

[19]Video demonstrating the personalised robot narrator is available online: https://youtu.be/fBIIn0PQGSA

[20]Second Language Tutoring using Social Robots (L2TOR): http://www.l2tor.eu

menting additional personalisation modules for retrieval of user profile, based on the estimated age of the user.

The author contributed to the analysis and writing of the results for the study on child speech recognition.

1.5 Structure

The structure of this thesis is outlined below with an overview of the content and context for each chapter. A summary of key elements and findings are included at the start of Chapters 3 to 9 to clarify their contribution.

- This chapter provided an introduction and motivation for personalisation in long-term human-robot interaction, in addition to the research questions and objectives including the primary thesis, the scope, and the research approaches and contributions of the work presented in later chapters.
- Chapter 2 provides a background for the research, touching upon topics on long-term human-robot interaction, personalisation and long-term memory, user recognition, conversational artificial intelligence and real-world application areas where personalised robots can create an impact in long-term interactions, identifying the key areas for contribution that this thesis aims to address.
- Chapter 3 proposes a novel multi-modal user recognition system, MMIBN, that enables autonomous incremental and online learning in long-term human-robot interactions. This chapter also introduces the *quality of the estimation (*Q*)* and *long-term recognition performance loss (*L*)* parameters for optimising user recognition in long-term interactions.
- Chapter 4 presents the first study evaluating the applicability of MMIBN to real-world interactions. The results show that multi-modal recognition improves the identification rate compared to face recognition and that the system is suitable for real-world HRI. Moreover, it is found that online learning performed worse than using fixed likelihoods for the user recognition system.
- Chapter 5 presents the multi-modal long-term user recognition dataset, which is created to extend the work in Chapter 4 for evaluating MMIBN for a large number of

users. As a result of the study, it is found that MMIBN significantly outperforms and improves the identification rate compared to the face recognition, soft biometrics and a state-of-the-art approach in open world recognition. In addition, it is confirmed that online learning does not provide better performance than fixed likelihoods, but it allows users to be recognised equally well.

- Chapter 6 presents the developed Barista and Personalised Barista datasets that are used to train and evaluate rule-based and data-driven approaches for task-oriented closed-domain dialogue.
- Chapter 7 describes the architecture for a fully autonomous barista robot with user recognition, automatic speech recognition and a rule-based dialogue management system. In addition, it examines the impacts and the challenges of using a barista robot in a real-world study with non-native English speakers.
- Chapter 8 explores the applicability of the state-of-the-art end-to-end data-driven approaches to personalisation in long-term interactions based on the Barista datasets described in Chapter 6.
- Chapter 9 describes the architecture of a personalised socially assistive robot for cardiac rehabilitation, in addition to the user study showing its application in the real-world therapy.
- Chapter 10 draws on the main findings of technical and experimental work from previous chapters, alongside the context of the related work, to present a discussion concerning the research questions and contributions introduced in this chapter. The limitations of the work conducted here are outlined, along with the future directions of research and the potential impact of the designed systems in long-term HRI.

Chapter 2

Background

This thesis focuses on multi-modal personalisation in long-term human-robot interactions, particularly through user recognition and natural language-based interactions, for real-world applications. This chapter provides a theoretical background in these areas.

In Section 2.1, previous work in the domain of long-term human-robot interaction is surveyed to identify the challenges and methodology necessary in interactions that last for extended periods of time. Subsequently, long-term memory systems and personalisation in human-robot interaction are reviewed. In addition, a brief overview of personalisation in real-world applications is laid out, highlighting service robotics and socially assistive robotics, where an impact has been previously observed for long-term interactions.

Following the perspectives of the research in long-term interactions and personalisation, we focus on finding a suitable architecture for user recognition that enables fully autonomous incremental and online learning for real-time interactions on a robot. A broad overview of user recognition is provided in Section 2.2, whereby, the terminology is introduced for biometric systems, and the underlying design choices in this project are described under the topics of multi-modal recognition, soft biometrics, open-world recognition and online learning, in addition to the previous methods in Human-Robot Interaction (HRI), identifying areas for contribution.

As previously established in the scope of this work, we focus on social HRI based on multi-modal bi-directional exchanges of information. Consequently, in Section 2.3, several approaches in Conversational Artificial Intelligence (AI) for the fields of human-robot interaction and chatbots are discussed, followed by a review of personalised conversational agents.

2.1 Long-Term Human-Robot Interaction

In short-term interactions ("the first 10 minutes of HRI" (Bartneck et al., 2019)), users are generally affected by the "novelty-effect", which arise from the unfamiliarity with robots, hence their initial reactions might be quite different from their reactions over a longer period of time. In contrast, long-term interactions consist of multiple interactions with the user over extended periods of time. An interaction is considered "long-term" after the novelty-effect wears off and the user becomes familiarised with the robot, which depends on the number of interactions with a robot, and the context and complexity of the robot's behaviour (Leite et al., 2013). Research in this area enables investigating changes in user behaviour and experiences, and observing the integration of robots into human social contexts over time, as well as facilitate the design of systems for real-world applications (Leite et al., 2013; Bartneck et al., 2019). Here we will explore the studies in a variety of application areas, highlighting the challenges, methodologies and the key conclusions.

2.1.1 Challenges and Design Considerations

There are fewer number of long-term interaction studies in the literature compared to short-term studies, due to the challenges that long-term interactions face, such as lower number of subjects and the required level of autonomy and lack of robust technology, as described in detail below.

2.1.1.1 Subjects

Conducting long-term studies are more labour and time intensive both for the experimenter and the subjects, in comparison to short-term studies (Ganster et al., 2010), especially in the real world. Thus, it is challenging to recruit subjects willing to participate in long-term studies, which limits the number of subjects in these studies (Leite et al., 2013). For example, Huttenrauch & Eklundh (2002) explored the long-term effects of a service robot, which is designed to help motion-impaired people with the transportation of light objects in an office, with only a single participant over a period of 3 months on an average of 3 hours per day. In another study (Wada & Shibata, 2007), which is accepted as one of the landmarks in the field of long-term interaction (Leite et al., 2013), 12 participants in a care house interacted with a PARO robot (a seal shaped robot) over a month for 1-4 hours per week. The robot demonstrated animal-like behaviour, such as responding to touch, sound and lights, and recognised a limited number of words. The experiment showed that the robot strengthens the social ties among the residents of the care house, deducted from the increased duration of interaction between residents. In addition, most residents established moderate or strong ties with the robot.

On the other hand, Gockley et al. (2005) deployed a "roboceptionist" at a university campus with 233 participants over 6 months. The robot could recognise users through their ID cards and conducted limited text-based dialogue to provide directions and talk about its background story. A receptionist role was chosen for the robot to have more frequent interactions, and it was deployed in a public space to maximise the number of participants. Even though many users repeatedly interacted with the robot, after a certain period, only a few of the interactions lasted for more than 30 seconds. Kanda et al. (2004) conducted a field trial with 228 children in a school for two weeks. The robot identified students using RFID tags and interacted with them in English through a limited set of recognised words and uttered sentences. The results showed that the duration of the interaction declined from 3-7 minutes in the first interaction to less than a minute in the later interactions.

These studies show that the duration of the interaction and the number of available participants depend on the capabilities of the robot and the application area.

2.1.1.2 Autonomy

A common technique in HRI is the Wizard-of-Oz (WoZ) methodology (Green et al., 2004), in which a robot is tele-operated by an experimenter away from the view of the user, to make the user believe that the robot is autonomous (Riek, 2012). This method allows researchers to bypass issues that make it difficult to run the system autonomously, such as automatic speech recognition (ASR), natural language understanding (NLU), natural language generation (NLG) and navigation. Additionally, WoZ has been used to gather data to design architectures for autonomous interactions. However, reliance on this methodology may create unrealistic expectations and result in findings that are not grounded in a realistic interaction, which threaten the validity of the studies and applicability of these results to future interactions with fully autonomous robots (Fernaeus et al., 2009; Riek, 2012). In addition, long-term interactions, especially in real-world applications, require a high level of autonomy (Goodrich & Schultz, 2007; Thill et al., 2012; Leite et al., 2013). Human operation does not scale to interactions over extended periods of time or in a variety of places. Especially for assistive robotics, the robots are aimed to help overcome the shortfall of the workforce (Fasola & Matarić, 2013a), hence, having robot operators is not plausible in the real world. However, in order to design and deploy fully autonomous robots to the real world, we need robust and reliable technology. Currently, such a technology does not exist.

Interactions in the real world carry the challenges of complex environments and noisy data. In addition, for long-term interactions, the robots may need to interact in dynamic environments, with a large number of people, for extended periods of time starting with incomplete or incorrect knowledge due to the inaccuracies in the sensors. Kunze et al. (2018) define common necessities and challenge areas in robotics for long-term autonomy to be: navigation and mapping, perception, knowledge representation and reasoning, planning, interaction, and learning. In this thesis, we rely on a static robot to avoid problems with navigation, however, we will examine the other challenges that are of particular importance.

If the users will be encountered repeatedly during long-term applications, autonomous user identification that can adapt to the changes in the appearances of users is essential, which is an open challenge of perception for robots (Kunze et al., 2018). Currently, there is no reliable user recognition system that can identify and learn users autonomously and incrementally for long-term human-robot interactions, as outlined in more detail in Section 2.2.

Autonomous robots in long-term interactions in the real world require knowledge representation and reasoning capabilities to represent various aspects of the world and reason about them, in particular when these aspects change over time. If the environment is not fully known before deployment or new objects may need to be learned, it is known as an *open world* problem. Novelty and anomaly detection, in addition to belief revision through updating the beliefs with new information, e.g., through online learning, are essential in long-term scenarios (Kunze et al., 2018).

Planning and scheduling technologies are fundamental to account for the changes in the environment or task dynamics, and to determine the necessary sequence of actions for achieving a task. These structures allow adapting the behaviour of the robot in an online manner (Ingrand & Ghallab, 2017).

As previously mentioned, long-term HRI studies with fully autonomous social robots are not common in the literature due to the difficulties of the dynamics and non-predictability of an interaction with a human. Additionally, if the interaction is based on natural language, the lack of robust technology might cause challenges in the interaction. Especially in child-robot interaction, the state-of-the-art approaches (e.g., Google Speech-to-Text, Microsoft Speech, CMU PocketSphinx, Nuance NAOqi speech recognition engine) fail to provide reliable results due to a high number of disfluencies and ungrammatical language utterances in child speech (Kennedy et al., 2017). Moreover, unconstrained speech recognition in noisy environments (Shiomi et al., 2008), speech recognition with non-native speakers (Kitashov et al., 2018) or elderly (Young & Mihailidis, 2010) also pose as challenges for natural language interactions.

It is not possible to design and generate every possible response, action or sensor processing structures for human-robot interactions (Goodrich & Schultz, 2007). Hence, learning during deployment rather than during a design phase is crucial in achieving long-term autonomy in open or dynamic worlds to compensate for the lack of complete knowledge of users and the environment at the start of an interaction (Kunze et al., 2018). In addition, techniques that allow robots to continually learn from experience should allow online improvement of capabilities, autonomy, and interaction. However, unsupervised online learning may result in worse performance due to inaccuracies in the data, which can be overcome with "human-in-the-loop" systems to confirm the learned information.

2.1.2 User Expectations and Engagement

Short-term studies benefit from the novelty of the robot, that is, a higher positive user experience is achieved due to it being the first time interaction with a robot. Even though users might initially have high expectations, they might not have enough time to explore the extent of capabilities, or rather the lack of capabilities, of the robot. In contrast, long-term interactions allow users to encounter such limitations, which may result in a decrease of interest and user engagement (Leite et al., 2013). The behaviour of the robot might not be attractive enough to keep user expectations, and the frequency of interacting with the robot and the user interest may decrease over time (Fernaeus et al., 2010; Kanda et al., 2007; Tanaka et al., 2006; Huttenrauch & Eklundh, 2002; Kanda et al., 2004; Salter et al., 2004; Gockley et al., 2005; Sung et al., 2009).

Anthropomorphism might raise false expectations regarding the cognitive and social abilities of a robot, that the robot fails to fulfil (Dautenhahn, 2004). In fact, when the robot looks less human-like and demonstrates matching behaviour, users would expect less cognitive human-likeness (Hayashi et al., 2010). In addition, the theory of the "uncanny valley" (Mori, 1970) suggests that the likeability of a robot increases with more human-like appearance, until a point where subtle deviations from human appearance and behaviour create an unnerving effect, at which point the likeability decreases dramatically. Moore (2012) supports the previous views on user expectations, and suggest that the uncanny valley effect arises because of expectations that increase with human-likeness, such that when a certain aspect of the appearance or behaviour is wrong, the likeability decreases. Hence, one way to prevent users from having high expectations is by using robots that have a simpler appearance.

However, the simpler appearance of a robot does not guarantee low expectations. Forlizzi & DiSalvo (2006) stated that users had lower expectations for the practical functionality of a Roomba robot (a service robot for cleaning the floors) compared to other social or fictional robots. However, the users were disappointed that the robot did not gain knowledge of the environment over time and adapt its behaviour accordingly, based on their expectations that arise from other technological systems (e.g., phones, cars). Nowadays, these expectations have risen due to the false advertisement of robots in the media[1]. The state-of-the-art approaches, especially for verbal skills, do not meet these level of expectations, which causes a sharp decrease in user interest over time (Gockley et al., 2005; Kanda et al., 2007). Hence, it is crucial to well inform participants of the capabilities of the robot and the context of the study.

Another important approach that would facilitate user engagement in long-term interactions and help meet the growing expectations of users is to design robots that can increase their knowledge over time. Remembering previous aspects of an interaction and learning new information through the use of a memory structure may give users the impression of behavioural coherence and plausibility, hence, it might positively influence the perception of intelligence and, in turn, the quality of the interaction with the robot (Lim et al., 2011).

[1]The news article on falsely advertised robots in the media: `https://www.forbes.com/sites/noelsharkey/2018/11/17/mama-mia-its-sophia-a-show-robot-or-dangerous-platform-to-mislead/#160615ca7ac9`

In addition, adapting to users and personalising the experience can help improve user engagement and create a sense of familiarity over time to facilitate establishing rapport and trust between the user and the robot (Dautenhahn, 2004; Sabelli et al., 2011; Leite et al., 2013; Irfan et al., 2019). In the following sections, we will discuss the methodology and effects of long-term memory systems and personalised robots in detail.

2.1.3 Long-Term Memory Systems and Personalisation

As stated in Chapter 1, this research focuses on personalisation in long-term human-robot interactions. Memory is essential for learning, recalling and personalising interactions that last for extended periods of time (Castellano et al., 2008; Wood et al., 2011; Leite et al., 2013; Baxter & Belpaeme, 2014). Several memory models have been developed in HRI for long-term interactions. Here we will focus our attention on long-term memory models for personalised social HRI based on natural language.

The first generally accepted model of human memory is Atkinson-Shiffrin model (Atkinson & Shiffrin, 1968), known as the multi-store model, that consists of *sensory*, *short-term* and *long-term memory*. According to the model, input passes from sensory memory into a short-term storage gated by attention, which holds information for a finite length of time. Through rehearsal, the information is passed to the long-term memory for storage for longer periods of time, in addition to retrieval of information into short-term memory, when necessary. Short-term memory is bounded by the context and task demand, whereas long-term memory is not dependent on specific tasks. *Working memory* is a variant of short-term memory (Baddeley & Hitch, 1974; Cowan, 1988) for a temporary recollection of task-relevant information. Long-term memory can be further categorised into two parts: *procedural* and *declarative memory*. Procedural memory contains non-consciously accessible information such as skilled motor behaviours, habits, and stimulus-response conditioning. In contrast, declarative memory holds consciously accessible information, consisting of *semantic memory* that stores symbolic information about context-free objects, facts, and concepts, and *episodic memory*, which contains "who, what, where, when information" about previous interactions. For a nontrivial level of social HRI, Wood et al. (2011) suggest using a biologically-inspired long-term memory that contains semantic (e.g., user's name and preference), as well as, episodic information, by using contextual, spatial, and temporal information about previous interactions. For example, for a long-term companion robot, Ho et al. (2010) proposed a semantic memory to store user's preferences in the initial interaction and adapt to the changes in the preferences, and episodic memory to retrieve similar events for deciding the correct actions in a currently encountered situation.

Section 1.1.2 highlighted the importance of using incremental learning (i.e., expanding a model for new users or attributes) and online learning (i.e., updating a model sequentially with incoming information) for adaptation in long-term interactions. Previous work in long-term memory systems in HRI have implemented several ways of learning from users during deployment, mostly based on rule-based systems relying on a knowledge-base to extract and update information, or probabilistic methods and "human-in-the-loop" systems that involve human input to learn and validate the information. For example, an interactive manipulator robot arm used a combination of working memory and long-term memory based on rule-based systems for incrementally learning and updating sensory-motor actions through vision and dialogue (Mavridis & Roy, 2006; Mavridis & Petychakis, 2010). Müller et al. (2014) designed a robot for accident prevention and assistance for elderly people in long-term interactions, using a Partially Observable Markov Decision Process (POMDP) that adapts it behaviour online based on user's reactions and explicit rewards. The results of their user study suggested that exploration feature of the architecture was negatively perceived because the users were expecting persistent behaviour. Campos et al. (2018) designed a "conversational memory" for a robot that personalised its interactions by recognising users and revisiting common episodic shared history to maintain a coherent social relationship over time. Situated learning and crowd-sourcing were used to learn from user utterances, generate new utterances and validate continuation sentences and end of a conversation. Their evaluation study over 14 days showed that only on rare occasions, the robot was able to refer back to previously shared history because the users changed the topic or did not understand the intention of the robot. While this memory system allowed users to converse about a various range of topics with the robot, crowd-sourcing is very costly, hence, is not feasible in the real-world applications. On the other hand, data-driven approaches that automatically infer knowledge and strategies from data are rarely used in HRI, due to the amount of data necessary to train these systems, which is especially challenging to obtain in HRI. In fact, there are no studies that explored data-driven approaches in long-term HRI. The studies that have used data-driven approaches for designing social robot behaviours are based on short-term interactions, which were trained on corpora of natural human-human interactions (Liu et al., 2014; Liu et al., 2016, 2018; Liu et al., 2019; Doering et al., 2019a,c,b). In this thesis, we explore rule-based approaches (Chapter 7 and Chapter 9), in addition to data-driven approaches for personalisation in long-term interactions (Chapter 8).

A key problem with long-term memory is that storing all the information obtained in interaction obstructs recalling relevant information in new interactions (Castellano et al., 2008). Therefore, it is essential for the robot to know which information should be recalled based on user inputs and the context of the interaction. The ways to obtain the relevant information from an interaction vary depending on the application area and the complexity of the task. On the one hand, close-ended structured dialogues can be used to obtain pre-determined information from the user, however, this might reduce the adaptability to variations in user responses. On the other hand, extracting the relevant information from data would require an extensive amount of interactions, and might prove counterproductive in some applications due to the noise in the data. Both types of approaches are explored in this thesis, and the corresponding background in conversational agents is detailed further in Section 2.3.

Most research focuses on using a pre-determined set of attributes for adaptation, depending on the task. These attributes and the behaviour of the robot are determined based on the task and the application domain. Previous studies showed the benefits of remembering **user's personal attributes** (e.g., name, gender, age) (Kanda et al., 2004; Kanda et al., 2007, 2010; Sabelli et al., 2011; Gockley et al., 2005; Mutlu et al., 2006; Belpaeme et al., 2013; Kennedy et al., 2015; Leite et al., 2014; Churamani et al., 2017; Campos et al., 2018; Zheng et al., 2019), **preferences** (Belpaeme et al., 2013; Ho et al., 2010; Churamani et al., 2017; Zheng et al., 2019) and **behaviour patterns** (Glas et al., 2017; Zheng et al., 2019), in addition to recalling **previous shared history** (Ho et al., 2010; Belpaeme et al., 2013; Matsumoto et al., 2012; Leite et al., 2014, 2017; Campos et al., 2018; Zheng et al., 2019; Ahmad et al., 2019) for improving user experience in long-term interactions, especially through personalising the interaction. Zheng et al. (2019) compared remembering these four types of information with a personal assistant robot (ERICA robot, by Glas et al. (2016)) that tracked the status of users' tasks and gave health tips. They used a combination of sensory memory, working memory, and long-term memory with a rule-based knowledge-base structure that stored and retrieved memories. Their findings suggested that commenting on observed user behaviour patterns elicits stronger positive feelings, and tracking the progress of user's goals and recalling previous shared history are more effective in building rapport than commenting on semantic information (e.g., personal attributes or preferences). However, their results were based on a study with three users and referring to the behaviour patterns were encountered far less than the other behaviour, hence, a more comprehensive study is necessary for more conclusive results since each of these types of information proved useful in the literature.

2.1.4 Real-World Applications

Conducting longitudinal studies in the real world can reveal how users interact with the robot when the tasks involving them become more of a routine (Sung et al., 2009). However, deploying autonomous robots in the real world creates additional challenges, such as incomplete data and dropouts, thereby, decreasing the success rate of the interaction (Dondrup et al., 2018). However, such challenges allow testing the limits of HRI systems and enable observing how people react to failures in the real world, such that we can design reliable and durable systems.

Robots are beginning to emerge in real-world applications in a variety of areas, such as retail[2] (e.g., promoting sales[3], managing warehouses or being used in delivery), restaurants[4] (e.g., servers, cooks or hostesses), domestic environments (e.g., for cleaning (Prassler et al., 2016) or as a companion) and healthcare[5] (Prescott & Caleb-Solly, 2017) (e.g., exoskeletons, surgery, telepresence and robots in therapy), in addition to industrial and military robots. However, several domestic robots (e.g., Jibo[6], Vector (Anki)[7], Kuri (Mayfield Robotics), Keecker[8]) recently were withdrawn from the market or the companies manufacturing them shut down due to low sales, which is presumed to arise from the lack of a profoundly valuable task to justify the significant purchase cost[9]. In contrast, in retail, restaurants and healthcare, there is a "one-to-many" structure, where one robot can interact with hundreds of people per day or over long periods of time, thereby, making it more likely to justify its purchase value. In addition, some customers may repeatedly visit the same shops or restaurants. Hence, recognising "regular" customers and recalling their preferences for

[2] https://emerj.com/ai-sector-overviews/robots-in-retail-examples/

[3] Pepper robot is used by several companies in retail (Pandey & Gelin, 2018): https://www.softbankrobotics.com/emea/en/industries/retail

[4] https://medium.com/@olivermitchell/the-new-restaurant-experience-robot-servers-cooks-and-hostesses

[5] https://interestingengineering.com/15-medical-robots-that-are-changing-the-world

[6] https://www.jibo.com

[7] https://anki.com/en-us.html

[8] https://www.keecker.com

[9] https://spectrum.ieee.org/automaton/robotics/home-robots/why-the-pursuit-of-a-killer-app-for-home-

personalising their experience, e.g., to recommend new products, may increase rapport with customers and encourage them to return to the shop or the restaurant again (Gwinner et al., 1998; Kanda et al., 2010; Niemelä et al., 2019). While recalling all previous customers is a very difficult task for a human, this can potentially be achieved with a robot (Glas et al., 2017). Concerning healthcare, the number of individuals with long-term conditions, such as stroke, arthritis, heart disease or dementia, is increasing with population growth, whereas there is a growing need of qualified nurses and residential care workers (Fasola & Matarić, 2013a; Prescott & Caleb-Solly, 2017; UK, 2017). Robots in assistive roles can help reduce the workload of carers and medical specialists. Furthermore, personalised robots can help track the progress of patients which would help doctors personalise the therapy to the patient. Hence, we decided to focus our research on customer-oriented service robots and robots in healthcare settings.

2.1.4.1 Service Robotics

Despite the growing use of customer-oriented service robots in the real world, there are relatively few studies that explore their long-term effects. For example, Kanda et al. (2010) used a robot in a shopping mall to give directions and advertise shops. The robot also recognised users through RFID tags and personalised its recommendations based on user preferences over time. The results showed that personalisation resulted in increased familiarity and rapport. In addition, the authors suggested that using more natural user recognition methods (e.g., face recognition) will increase the robot's perceived intelligence. Similarly, a robot that personalised its greetings based on the visiting patterns of the recognised users in a shopping mall, such as frequency, time of the day, walking speed, and the group size accompanying the user, was found to increase familiarity with the robot (Glas et al., 2017). Lee et al. (2012) designed a personalised snack delivery robot for long-term interactions that recalled snack choices, service usage patterns, and the robot's shared history with users (e.g., referring to previous robot failures). User orders were taken and tracked via a website. The results suggested that personalisation reinforced participants' rapport, cooperation, and engagement.

However, none of the previously mentioned service robot studies was fully autonomous. These studies used partial WoZ methods for helping with speech recognition, tracking user preferences, controlling dialogue, specifying user locations or validating user recognition.

2.1.4.2 Socially Assistive Robotics

Socially assistive robotics (SAR) refers to a domain of HRI where robots are used in healthcare and therapies (Tapus et al., 2007). Examples of such studies involve **cognitive and developmental disorders** (Michaud & Théberge-Turmel, 2002; Robins et al., 2005; Scassellati, 2007; Thill et al., 2012; Moro et al., 2018; Rudovic et al., 2018; Scassellati et al., 2018; Cao et al., 2019; Clabaugh et al., 2019), **care for elderly** (Roy et al., 2000; Tapus, 2009; Fasola & Matarić, 2013b; Khosla et al., 2016; Hanheide et al., 2017; Cao et al., 2019; Kachouie et al., 2014), and **rehabilitation** (Eriksson et al., 2005; Tapus et al., 2008; Fasola & Matarić, 2013b; Süssenbach et al., 2014; Schneider et al., 2017; Woodworth et al., 2018; Cao et al., 2019). However, most research in rehabilitation has been carried out in laboratory conditions or during short-term interventions, which restrict the applicability of the results to long-term therapies in real-world applications due to the confounding factors, such as the novelty effect (Gockley et al., 2005) and the adaptation of the technology (Riek, 2017). In fact, similar to the other domains, longitudinal research on SAR (Wada & Shibata, 2007; Kidd & Breazeal, 2008; François et al., 2009; Sabelli et al., 2011; Süssenbach et al., 2014; de Graaf et al., 2015; Scassellati et al., 2018; Clabaugh et al., 2019) is notably less than short-term studies, where some studies report a considerable decrease in user interest and motivation compared to the initial interaction (Kidd & Breazeal, 2008; Süssenbach et al., 2014).

Each area of rehabilitation is specialised in its own requirements, therefore, the role of the robot can change depending on the task to assist with (Duffy, 2003), the user population to work with (Scheeff et al., 2002), and the appearance and behaviour of the robot (Feil-Seifer & Matarić, 2005). However, the applications share common goals, such as providing monitoring, feedback and assistance, increasing user motivation, and improving task performance and progress (Ahmad et al., 2017). Personalisation can provide individualised care in these tasks (Matarić et al., 2007; Matarić & Scassellati, 2016), improve user performance (Tapus et al., 2008; Tapus, 2009), increase perceived familiarity and sociability (Sabelli et al., 2011; Fasola & Matarić, 2013b), and elicit and maintain user engagement over extended durations (Scassellati et al., 2018; Winkle et al., 2018; Clabaugh et al., 2019). Personalisation needs to focus both on the short-term changes that represent individual differences (e.g., name, personality, preferences) and on the long-term changes (e.g., therapy progress) that enable the interaction to continue to be engaging in the long-term through both verbal and non-verbal communication (Tapus et al., 2007). However, due to the diverse individual needs, the noise of real-world environments, and the scale of need in rehabilitation, non-autonomous personalisation of SAR is infeasible.

Most research in personalisation in SAR relies on rule-based approaches that use a predefined set of rules determined by medical staff or based on previous research. Model-based or statistical approaches (e.g., reinforcement learning or Bayesian methods) have also been used in SAR for learning user preferences (Woodworth et al., 2018), adaptation to user profile or states (Conn et al., 2008; Chan & Nejat, 2011; Gordon & Breazeal, 2015; Gordon et al., 2016; Schodde et al., 2017; Rudovic et al., 2018) or personalising instruction and feedback (Clabaugh et al., 2019). However, these methods require learning from a large amount of data (Clabaugh et al., 2019), which is especially lacking in HRI. In addition, these approaches may prove unreliable in providing a structured therapy (Gordon et al., 2016) and may result in incorrect behaviours of the robot, therefore, posing serious health problems, especially in rehabilitation therapy.

In this section, we highlighted the challenges of long-term human-robot interaction, and we identified that common solutions involve autonomous personalisation of the interaction through recognising users and adapting the communication and robot behaviours. In the following sections, we discuss the previous work on user recognition and conversational artificial intelligence to determine the methods appropriate for customer-oriented service robotics and socially assistive robotics.

2.2 User Recognition

User identification is an important step towards achieving and maintaining a personalised long-term interaction with robots. Contrary to the general approaches in biometric recognition, in an HRI scenario, the robot may start from a "tabula rasa" state with no prior knowledge of users. The users would be encountered incrementally, that is, all the users will not be introduced as a "batch". Hence, it is necessary to differentiate "unknown" users from those that are previously enrolled. Ideally, new users should be allowed to enrol in the system at any time for future recognitions. This section introduces the concepts of recognition and details the state-of-the-art approaches in biometric systems and HRI.

Recognition in biometric systems can either refer to *verification* (*authentication*) or *identi-*

fication (Phillips et al., 2011). Verification involves a *one-to-one* match that compares the presented biometric sample, called a *probe*, to the corresponding biometric samples of the claimed identity in the enrolment database (*gallery*). Identification corresponds to *one-to-many* matching that compares the given probe to multiple biometric samples in the gallery to associate the identity of the user to one of those in the database. Moreover, identification can be divided into two categories: *closed-set recognition* and *open-set recognition*. Closed-set identification assumes that the probe should belong to someone in the gallery, that is, all users to be identified are previously enrolled in the system. In open-set identification, the probe does not have to belong to a user within the database. Hence, the system has to decide whether the user is new or a previously enrolled user. Open-set recognition is a well-established area (Scheirer et al., 2013; Jain et al., 2014; Scheirer et al., 2014), but in a real-world setting, these unknown classes might need to be added into the system for future recognitions, extending the open-set recognition problem to include the incremental learning of new classes, known as *open world recognition*. Recognition in human-robot interaction is a type of open world recognition problem where the new users need to be identified as unknown, and these users should be further allowed to enrol in the system.

Face recognition (FR) is the most common method in biometric recognition. Nevertheless, most FR challenges, such as Face Recognition Vendor Tests[10], only evaluate verification algorithms. To this date, Unconstrained Face Detection and Open Set Recognition Challenge[11] remains the only available open-set identification challenge, which show that the best algorithms achieved good identification accuracies at the cost of high false identification rates (Gunther et al., 2017).

FR could be unreliable during a real-time identification, due to several reasons, such as changing facial features, expressions, occlusions and lighting conditions (Wójcik et al., 2016). Another example is the recent release of a smartphone with a built-in FR system, which was reported to fail in distinguishing family members of different genders and ages due to the similarity of their facial features[12]. This issue raised awareness of the security and privacy problems that using FR might cause, as compared to access through a passcode.

[10] https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

[11] http://vast.uccs.edu/Opensetface

[12] https://www.wired.com/story/10-year-old-face-id-unlocks-mothers-iphone-x/

In addition to false identification, in the case that a biometric system cannot obtain meaningful data, Failure to Capture error (FTC) will be encountered (Ross & Jain, 2007). For example, a face may not be detected in the case of a blurry image while the person is in motion. Failure to Enroll error (FTE) denotes the proportion of users that cannot be successfully enrolled in a biometric system for this reason. Moreover, biometric systems that use a single identifier, called a *uni-modal biometric system*, has an upper bound on matching accuracy.

2.2.1 Multi-modal Recognition

Humans use multi-modal information for recognising a person, especially in case of incomplete information, such as using voice in a dark room. Similarly, *multi-modal* biometric systems can improve the matching accuracy by fusing information from multiple biometric identifiers, which allows reducing the effects of noisy data, decreasing FTE error, and eliminating the upper bound set on the accuracy for a better estimation of the identity. Robots are ideal platforms for using multi-modal recognition due to carrying a wide range of sensors.

Several post-classification fusion methods have been proposed for integrating multi-modal information, as shown in Fig. 2.1, which can be classified into three categories: *decision*, *rank*, and *confidence* level fusion (Jain et al., 2005). Decision level fusion (e.g., majority voting, AND/ OR rule) is mainly used for combining individual best matches from each biometric classifier. Rank level methods (e.g., highest rank, logistic regression) are used when the output of each biometric classifier consists of ranked matches.

Confidence level fusion methods are the most common approaches, as they combine the individual scores from multiple biometric sources, which provide more information than the ranks or best matches. There are two approaches to combine the scores for confidence level fusion: *classification* and *combination*. In the classification approach, a feature vector is used to combine the output of individual identifiers, which is then classified into categories. This approach allows combining non-homogeneous data, such as distance or similarity metric and different numerical ranges, therefore, no pre-processing is required. Examples of the classification approach are neural networks, k-nearest neighbors, Support Vector Machine (SVM), and decision trees.

In the combination approach, the individual matching scores from multiple biometric

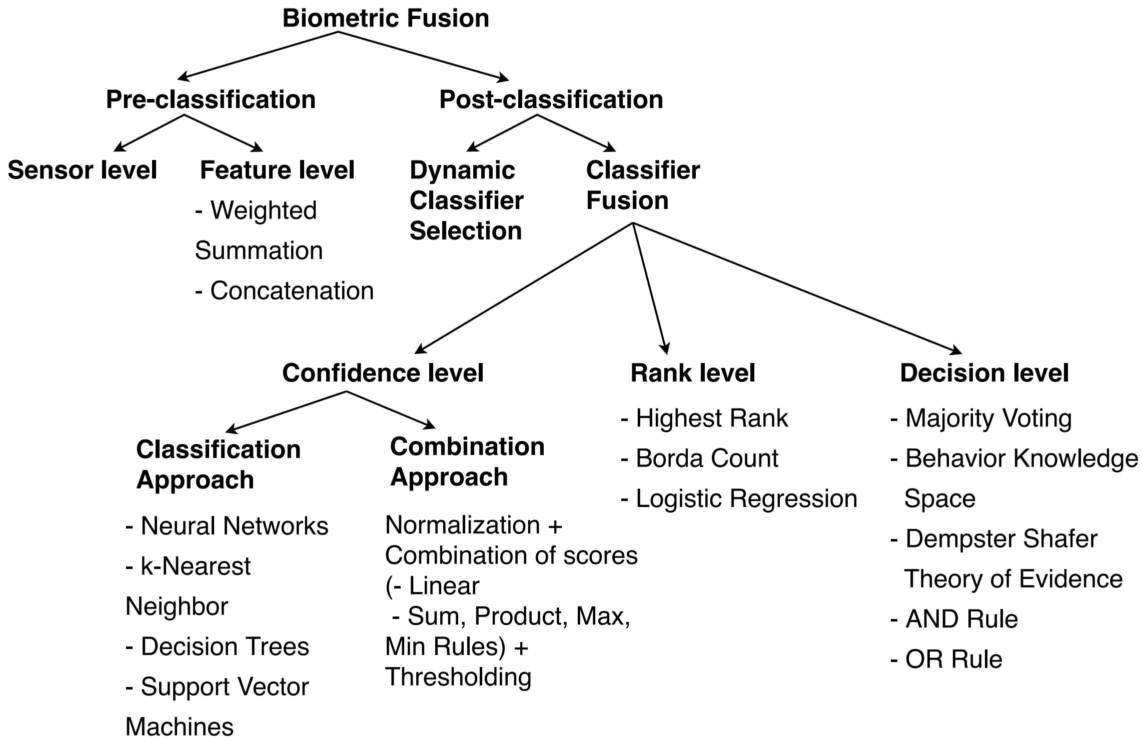


Figure 2.1: Biometric fusion methods (adapted from Jain et al. (2005)).

identifiers are combined into a scalar score, which is used to identify the person. Several combination methods exist in the literature, such as *product rule*, *sum rule*, *min rule*, *max rule*, *median rule*, and *majority voting*. For instance, a Bayesian network (BN) computes the probability of an outcome by using the product rule to combine the probabilities of a set of variables based on their conditional dependencies. Thus, BNs are suitable for multi-modal biometric recognition (Bigün et al., 1997; Verlinde et al., 1999; Jain et al., 2004; Scheirer et al., 2011).

In order to ensure a meaningful combination of scores, the scores must be normalised into a common range through normalisation. A good normalisation method should be *robust* and *efficient* in estimating the *location* (i.e., mean of a normal distribution) and *scale* (i.e., standard deviation) parameters, where *robustness* indicates the insensitivity to the outliers, and *efficiency* refers to the proximity of the optimal distribution to the estimated one when the former is known.

After the combination of the scores, a threshold is used to determine if a matching score corresponds to a genuine user or an impostor, that is, the values below the threshold are regarded as impostor score.

2.2.2 Soft Biometrics

Although most biometric systems utilise *primary biometrics*, such as fingerprint or face, for person recognition, other attributes of an individual such as age, gender, ethnicity, height, eye colour and clothing -referred to as *soft biometrics (SB)*- can provide additional information to improve the recognition performance (Dantcheva et al., 2016). Soft biometric traits are characteristics that provide information about a user that is not capable of uniquely identifying a person on its own.

The primary work (Jain et al., 2004) that proposed fusing soft biometric traits (e.g., gender, ethnicity, and height) with a primary biometric (e.g., fingerprint) used a . They proposed a weighting scheme where the traits with smaller variability and larger distinguishing capability will be given more weight in the computation of the final matching scores. Other notable research combined soft biometric data (ethnicity, hair colour and gender) with contextual information (occupation and location) using a Bayesian attribute network (Scheirer et al., 2011), or with other primary biometrics (Abreu & Fairhurst, 2011; Jain & Park, 2009; Zewail et al., 2004; Park & Jain, 2010).

2.2.3 Open World Recognition

User recognition in HRI, especially within long-term interactions, require identifying new users and enabling them to be enrolled for subsequent interactions, which is an open world recognition problem. Several open-set recognition algorithms exist for detecting novel classes, such as 1-vs-Set machine (Scheirer et al., 2013), Support Vector Machine (SVM) approaches (Scheirer et al., 2014; Jain et al., 2014), and nearest neighbors methods (Fayin Li & Wechsler, 2005; Mendes Júnior et al., 2017). However, these approaches require re-training the entire system in order to add new classes. While this may be feasible for a small amount of classes, the running time increases with the increasing number of classes and data (Bendale & Boult, 2015; Suguna & Thanushkodi, 2010; Wang & Wang, 2007). Thus, they are not suitable for real-time open world recognition. In addition, these approaches require a sufficient amount of data to form meaningful clusters and accuracies.

Despite the fact that most state-of-the-art face recognition methods use deep learning (Taigman et al., 2014; Sun et al., 2014; Parkhi et al., 2015; Schroff et al., 2015), only a few approaches exist for open-set recognition (Bendale & Boult, 2016; Ge et al., 2017; Shu et al.,

2017) and there does not exist one for open world recognition. The reason is most deep learning methods suffer from *catastrophic forgetting* problem, which refers to the drastic loss of performance on previously learned classes when a new class is introduced (McClelland et al., 1995; McCloskey & Cohen, 1989; Parisi et al., 2019). Existing approaches that could help to overcome this problem often require a part of the previous data for re-training, which might not be available. Moreover, similar to SVM and nearest neighbors methods, re-training does not scale well with time.

Bendale & Boult (2015) introduced the first algorithm applied to open world recognition, called Nearest-Non Outlier, based on Nearest Class Mean (Mensink et al., 2013) for open-set classification and incremental learning. Rudd et al. (2018) developed Extreme Value Machine (EVM) using Extreme Value Theory, which selects points and distributions that best summarise each class to form the probabilistic representation of decision boundaries to classify samples and identify unknowns. They used a Weibull distribution over the positive class scores to estimate the unnormalised posterior probability of inclusion for each class, while incrementally adding a batch of new classes using model reduction with a threshold. They showed that EVM is comparable to the W-SVM (Scheirer et al., 2014) in open-set recognition and outperforms Nearest-Non Outlier for open world recognition. Similarly, Fei et al. (2016) proposed a new approach for what they termed as *cumulative learning* which is a type of lifelong learning problem applied to open world recognition, based on a centre-based similarity space learning method and the 1-vs-rest strategy of SVM. However, none of these methods have been evaluated on user recognition. In addition, there are no open world user recognition methods that apply sequential learning of new classes.

Bayesian networks assume *a priori* knowledge of the states of variables and the conditional probabilities of these variables. Their structure, in theory, allows incremental learning to add new states, however, this would require readjusting of the probabilities between conditionally dependent variables, which could be the reason that they have not been applied to open world recognition problems before. Nevertheless, an extension of a BN was proposed to the *open universe* problem Milch & Russell (2010), which is concerned with the uncertainty about which objects exist, and the relations between the variables. However, the open world problem is simpler, in that, only the number of states of a variable, such as the number of users or the number of known faces, can change.

2.2.4 Online Learning

Humans can update their prior beliefs by continuously adapting to changing circumstances and learning effectively from their experience. An algorithm designed for open world recognition may not be sufficient to recognise a person after changes in their appearance, because the model is not updated for known samples. Such a problem, where the training data becomes available sequentially, can be solved by online learning (OL). Online learning allows the model to be updated at each sample, which can improve the performance in recognition (De Rosa et al., 2016).

There are many existing algorithms for online learning for classification. For instance, Lee & Kriegman (2005) proposed an online learning algorithm of probabilistic appearances for video-based recognition and tracking, which is comparable in performance to batch learning, but a prior generic model is necessary for their approach.

De Rosa et al. (2016) used online learning in open world recognition for incremental learning of classification metric and the threshold for novelty detection and local learning for describing the space of classes. Their results showed that online learning increases performance. The approach was applied to object recognition on three existing algorithms, namely, Nearest Class Mean, Nearest-Non Outlier and Nearest Ball Classifier (Rosa et al., 2015), and they showed that their approach performed better than these baselines.

For Bayesian networks, the conditional probability tables, hence, the priors and likelihoods are assigned based on *a priori* knowledge of the data, or they are learned from data (Koller & Friedman, 2009). However, hand-crafted likelihoods can cause incorrect estimations if the set probabilities are not accurate enough. If the data is fully available, Maximum Likelihood estimation is commonly used for learning the likelihoods from data, which relies on counting how many times each of the possible assignments of conditions appears in the training data (Koller & Friedman, 2009). However, in the case of incomplete data, the assumptions or learned likelihoods might not hold valid, which, in turn, would affect the overall estimations of the posterior probabilities. Online learning of the likelihoods can help achieve better performance.

Contrary to online learning of the network structure or parameters, the online learning of the likelihoods or priors has been understudied, as the availability of the conditional probability tables is assumed. If the relation is dependent over adjacent time steps, dynamic Bayesian networks can be used. However, in a recognition scenario, the previously recognised user can be different from the current target user, hence, dynamic Bayesian networks are not suitable. Some online learning approaches involve making assumptions, such as conditional independence on the priors (Oravecz et al., 2016) and Gaussian distribution of parameters (Opper & Winther, 1999), which does not apply to biometric data, or require careful selection of parameters (Honkela & Valpola, 2003) for good performance.

Bauer et al. (1997) first proposed the use of Expectation Maximization (EM) with a constant learning rate, named EM(η), for online learning of likelihoods in Bayesian networks with sequential and incomplete data. Cohen et al. (2001a,b) extended the approach for complete (fully observable) or missing (partially observable) data with an adaptive or constant learning rate, called Voting EM. When the data is fully observable, they suggested using an adaptive learning rate, such that the effect of new samples would decrease with time, based on Maximum Likelihood estimation or through a set of pre-defined values. On the other hand, for missing data, a fixed learning rate can be used, unless if there is hidden nodes, in which case, incremental EM (Neal & Hinton, 1999) can be used.

2.2.5 User Recognition in Human-Robot Interaction

The challenges in user recognition in HRI derive from the need to automatically learn and recognise users without intrusive methods or external devices, such as QR codes or access cards. This requires a real-time biometric system that allows incremental and online learning.

Similar to biometric recognition, the most common approach for user recognition in HRI is through face recognition (FR). The earliest work on incremental user recognition in HRI used FR on the Kismet robot (Aryananda, 2001) with seven subjects in a lab environment and with the Mertz robot (Aryananda, 2009) with 500 subjects in a public environment. User recognition was based on batch clustering of images taken from a video sequence during the interaction with a user. Even though the system was incremental, it required initial training based on 300 images to recognise users, and the users were generally misclassified as a known individual until their clusters were formed. Other classification approaches for open-set recognition recognition in HRI used SVM (Hanheide et al., 2008), Bayesian approaches (Cruz et al., 2008) and a combination of k-nearest neighbor with principal component analysis (Gaisser et al., 2013). However, all of these methods require offline pre-training or re-training in order to achieve autonomous user recognition. In other words, none of these methods allow autonomous open world user recognition starting from a state without any known user.

Robots lend themselves well for multi-modal recognition based on their multiple sensors. However, only a few studies exist that used soft biometrics for recognition in HRI. Martinson et al. (2013) used only soft biometrics, namely, clothing, complexion (nose and forehead) and height, to guess the identity of the individual out of three subjects within a group, with 202 subjects in total. They achieved a 90% correct identification rate, where clothing was the most reliable parameter, and complexion the least. The importance of clothing parameter can be explained by the short duration between the first initial 20-second training phase and the second recognition. However, in a long-term HRI scenario, clothing could be unreliable as the clothes of the users will change frequently. Moreover, the problem is likely to be more complex than a fixed choice task with 3 options, hence, soft biometrics alone would not be sufficient to identify a user.

Ouellet et al. (2014) combined face recognition, speaker identification and human metrology with a weighted sum in closed-set identification. They used pre-training that consisted of 30 seconds of facial images from different angles and facial expressions, reading passages for 2 minutes and 60 seconds for human metrology. They trained the system on 22 participants, but only evaluated on seven people with increasing distance to the camera. Thus, given the required length of the training data and the small number of participants, it is hard to conclude if the method works well in open-set identification. Another study for multi-modal recognition in closed-set identification, proposed combining face, body and speech information, but they have not tested their approach in a real-world HRI scenario (Al-Qaderi & Rad, 2018). However, none of these methods combined soft biometrics with a primary biometric for open world user identification in real-time HRI.

In this section, we have established that there are no incremental user recognition methods suitable for real-world HRI. Correspondingly, in this thesis, we will build a multi-modal user recognition system with incremental and online learning for long-term interactions, without the necessity of any preliminary training (Chapter 3, 4 and 5).

In the next section, we turn our attention to the communication within the interaction, such that we can identify methods that are suitable for personalisation in long-term interactions.

2.3 Conversational Artificial Intelligence

Verbal and non-verbal communication are vital for achieving a natural human-robot interaction with mutual understanding (Mavridis, 2015). Thus, as previously established in Chapter 1, our research focuses on multi-modal communication based on speech, text-based and sensory information. Conversational Artificial Intelligence (AI) is the field that focuses on building speech and text-based solutions to automate communication with conversational agents, such as chatbots, virtual agents and social (or *sociable*) robots. In this section, we will provide an outline of conversational AI to introduce the related concepts and common state-of-the-art architectures, ranging from rule-based methods to data-driven approaches, before focusing on personalised chatbots and robots.

Various approaches exist within conversational AI depending on whether the conversation is *task-oriented* (goal-oriented), such as meeting scheduling, giving directions and booking restaurants, or *general purpose* (e.g., chit-chat), and whether the range of topics is restricted (*closed-domain*) or not (*open-domain*). Our research focuses on closed-domain task-oriented natural language interaction.

Task-oriented dialogue is composed of a sequence of *turns* of *dialogue acts* from the user and the agent, where utterances are actions that can change the (mental) state of both the user and the system (Wittgenstein, 1953; Austin, 1962; Core & Allen, 1997; Traum, 1999). These actions can be used to suggest, plan, acknowledge, inform, request and confirm certain information (Bach & Harnish, 1979). A typical task-oriented dialogue system is composed of four modules (Tur, 2011; Young et al., 2013; Gao et al., 2019), as illustrated in Figure 2.2:

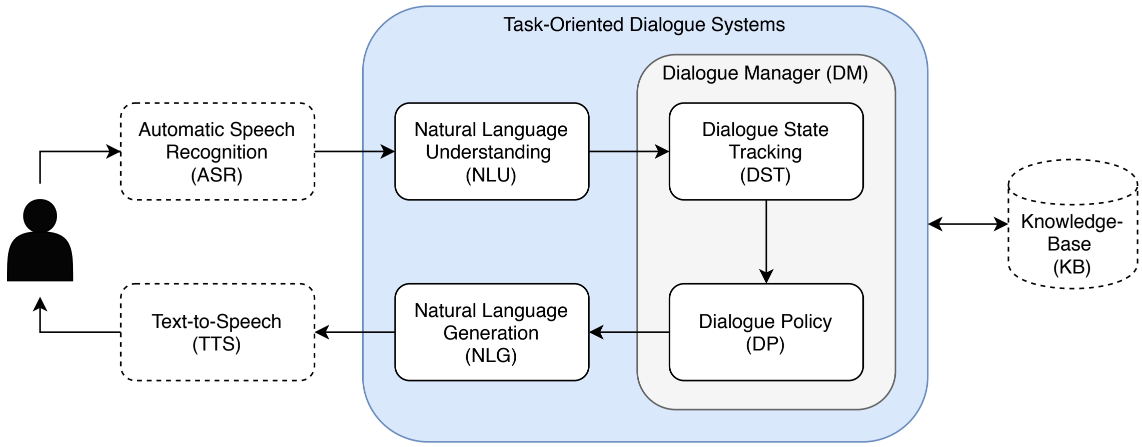


Figure 2.2: Task-oriented dialogue management system architecture, with optional mechanisms for storing and extracting information (KB), analysing speech input (ASR) and producing speech output (TTS).

(1) natural language understanding (NLU) module for identifying conversation domain, user intents and extracting related information through pre-determined templates; (2) dialogue manager (DM) for controlling the dialogue, composed of dialogue state tracking (DST) that determines the (belief) state of the conversation and dialogue policy (DP) that chooses the best action given the current state; (3) natural language generation (NLG) for transforming the agent action into a natural language response. Knowledge-base (KB) is an external database that is often used in task-oriented dialogue systems to inquire about or store information to accomplish the task. A system action can be a user response in the form of dialogue acts, or an internal operation such as a KB lookup or an application program interface (API) call. If a user communicates with the agent through speech instead of text-based input, an automatic speech recognition (ASR) system is necessary to extract corresponding words or sentences from speech signals before it is passed on to the NLU module. If the agent utterance will be delivered via speech, a text-to-speech (TTS) module transforms the text output into speech. Depending on the application domain (e.g., for robots), both the user input and the agent output may include a variety of modalities including gestures, visual displays and haptic feedback.

The dialogue managers in task-oriented dialogue systems can either rely on a set of hand-crafted if-then-structures and templates to track the dialogue state and choose the corresponding dialogue act (i.e., *rule-based* approaches) or model the uncertainty in dialogue states and learn the dialogue policy through *statistical* approaches. The simplest rule-based dialogue management systems (RBDMSs) rely on finite state machines (FSMs), where states represent questions, and links between states correspond to actions depending on the user response (Winograd & Flores, 1986; Goddeau et al., 1996; Stent et al., 1999). These systems are designed for *user-initiative* or *system-initiative* conversations, in which either the user or the system controls the dialogue through prompts or questions. *Template-based* (frame-based) systems aim to fill a set of slots with values extracted from the user's utterances, by matching them to pre-defined structures or keywords (Bobrow et al., 1977; Simpson & Eraser, 1993; Aust et al., 1995; Stent et al., 1999; Pieraccini et al., 1997; Thompson et al., 2004a). Bobrow et al. (1977) proposed the template-based Genial Understander System (GUS) architecture for travel planning that extracted various travel information, such as date of travel, the origin of travel and destination, and booked user's travel plans accordingly. This architecture underlies most (if not all) modern commercial digital assistants, with a *mixed-initiative* structure and nonlinear dialogue flow, in which the system and user may ask and answer questions in any order (Jurafsky & Martin, 2019). Additional rule-based approaches include plan-based systems (Allen et al., 1994; Cohen & Perrault, 1986; Cohen & Levesque, 1990), models of rational interaction (Sadek et al., 1997), and Bayesian approaches (Sun et al., 2014). Most task-oriented dialogue systems in HRI also rely on rule-based approaches (Gockley et al., 2005; Kanda et al., 2007, 2010; Giuliani et al., 2013; Kasap & Magnenat-Thalmann, 2012; Churamani et al., 2017; Williams et al., 2018; Zheng et al., 2019).

A variety of statistical approaches emerged due to the Dialogue State Tracking Challenges[13] (Henderson, 2015; Chen et al., 2017), such conditional random fields (Lee, 2013; Lee & Eskenazi, 2013; Ren et al., 2013), maximum entropy models (Williams, 2013), web-style ranking (Williams, 2014), Partially Observable Markov Decision Process (Young et al., 2013), and several machine learning approaches (see Henderson (2015) for a survey). In addition to state tracking, statistical approaches, such as supervised learning, reinforcement learning and transfer learning, have been applied to policy learning (Gašić et al., 2013; Genevay & Laroche, 2016; Mo et al., 2016). Nowadays, most of the publicly available and commercial task-oriented chatbot systems (e.g., Alexa, Siri, Google Assistant) are often a combination of hand-crafted components, which allow extracting information through common queries, and statistical methods that provide robustness to noise and ambiguity and allow learning through data (Gao et al., 2019).

Recently, data-driven approaches, which rely on extracting and learning the structures and values directly from the training data, have been devised for task-oriented dialogue to reduce the cost of laboriously hand-crafting dialogue managers (see Yan (2018) and Gao et al. (2019) for recent surveys on these methods). Data-driven approaches are categorised based on how the dialogue response is generated: *retrieval-based* (also called ranking or information retrieval) and *generative* models. Retrieval-based models select a dialogue response from a set of predefined responses (*candidates*). These systems can provide syntactically correct and specific responses, however, the responses are limited to those in the candidate set. In contrast, generative models generate a response word-by-word based on the conversation history (*context*), thus, they are prone to grammatical errors. One of the most popular generative models is that of Sutskever et al. (2014), called Sequence-to-Sequence (Seq2Seq). It relies on multi-layered long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) for encoding and decoding. The main advantage of this model is

[13] `https://www.microsoft.com/en-us/research/event/dialog-state-tracking-challenge`

that it requires only a small amount of feature engineering and domain specificity, thus it can be applied to tasks where domain knowledge may not be readily available or where the design rules are too complex to create manually (Vinyals & Le, 2015). In addition, it can generate entirely novel responses, in contrast to retrieval-based model approaches. However, these approaches require a large amount of data to train and are computationally expensive because they need to learn the sentence structure. Furthermore, the vanilla approaches rely only on recent dialogue history without using an external memory (Zhang et al., 2018).

Remembering all the information learned in an interaction, given the expanding volume of data over time, may prevent recalling salient information. One of the solutions to this problem in data-driven approaches is using attention mechanisms, which focus on particular elements of a task to respond to queries. Memory Networks (Weston et al., 2014; Sukhbaatar et al., 2015) combine a long-term memory with an attention mechanism, hence, in theory, they should be suitable and adaptable for long-term interactions. However, similar to other data-driven approaches, they have only been evaluated in single interaction or application domains, such as question-answering (Weston et al., 2014; Sukhbaatar et al., 2015; Chandar et al., 2016; Ganhotra & Polymenakos, 2018), language modelling (Sukhbaatar et al., 2015), task-oriented (Bordes et al., 2016; Joshi et al., 2017) and open-domain dialogue (Dodge et al., 2015; Zhang et al., 2018).

A similar approach to Memory Networks is Dynamic Memory Networks (DMN) (Kumar et al., 2015; Xiong et al., 2016), which uses gated recurrent neural networks (GRU) (Cho et al., 2014; Chung et al., 2014) and word sequence representation instead of sentence representation. DMN models are applicable to various domains such as question answering (QA), text classification for sentiment analysis, part-of-speech tagging and visual QA, however, it has not been evaluated on dialogue. Other similar approaches are Neural Turing Machine (Graves et al., 2014) and the work of Bahdanau et al. (2014), however, similarly to DMN, they have not been applied to task-oriented dialogue.

2.3.1 Personalised Conversational Agents

As we have established in the previous sections, personalisation is important in long-term human-robot interactions to improve user engagement and build rapport with users. In the domain of chatbots, personalisation was also shown to increase the perceived level of social intelligence of the agent, in addition to increased task efficiency and awareness of the situational context of the conversation (Neururer et al., 2018; Kocaballi et al., 2019). In the following sections, we will outline the previous research in personalisation of dialogue with chatbots and social robots to identify the appropriate architectures.

2.3.1.1 Personalised Chatbots

Some studies in the domain of chatbots relate personalisation to the agent's personal qualities, such as personality. For example, Zhang et al. (2018) applied End-to-End Memory Networks (MemN2N) to open-domain conversations with personalities (Persona-Chat dataset) to increase user engagement and create consistent dialogue. They compared several retrieval-based and generative models, which showed that ranking models outperform generative models in choosing the correct response. Nevertheless, generative models offer flexibility in a conversation and handle previously unseen (*out-of-vocabulary*) words better than the ranking models. Based on the Persona-Chat dataset, the Conversational Intelligence Challenge[14] was created under the scope of competitions tracks in NIPS (NeurIPS) conference in 2017 and 2018 (Dinan et al., 2019). This challenge gave rise to many approaches (e.g., Yusupov & Kuratov (2018) and Wolf et al. (2019), the winning systems of those years) that focus on maintaining a personality in open-domain dialogue.

Other studies focused on personalisation for adapting the conversation, content and behaviour of the agent to the preferences and needs of users and situational context. For example, Thompson et al. (2004b) proposed a template-based interactive system with a knowledge-base for restaurant recommendation, which could learn user preferences in a dialogue to improve future conversations. The results showed that the number of turns and time required to find an acceptable restaurant decreased over time with personalisation. Other rule-based approaches (Rich, 1979; Pargellis et al., 2004; Lucas et al., 2009; Tokunaga et al., 2017) used offline user-entered information inserted to the system at the set-up time, however, as we highlighted in previous sections, online learning from users during an interaction is essential for long-term interactions, since behaviours and user preferences can change over time.

Rule-based approaches perform well in simulated text-based datasets for personalisation, but they are not suitable for real-world applications, where handcrafting every possible

[14] http://convai.io

request or utterance in terms of templates is not plausible. Statistical dialogue systems, such as transfer learning (Genevay & Laroche, 2016; Mo et al., 2016), combine the modular architecture of rule-based systems with learning of states and actions from training data for personalising the interaction. However, the performance of these approaches depend on the pre-defined parameters of the reward function and may deteriorate due to the differences between the selected source users and the target user.

Data-driven approaches map the user input to the agent output directly, hence, provide more flexibility without requiring any set of rules. Joshi et al. (2017) created a personalised simulated text-based dataset building upon bAbI dialog dataset (Bordes et al., 2016) for restaurant booking, with recommendations based on dietary preferences and favourite food of the user, and adapted conversation and recommendation styles based on the user's gender and age. They also introduced a split memory architecture based on MemN2N, which allow the model to perform separate attention over the knowledge-base facts (or profile attributes), in order to reduce the confusion during retrieval and generative settings. Such confusion arises while retrieving information from the database about an entity, such as the address of a recommended restaurant, because there is no mechanism to enforce the attention of the network to interpret knowledge about a specified entity or link it to the user's attributes, unlike in a rule-based system. The proposed approach significantly outperformed supervised embeddings (Bordes et al., 2016) baseline, and performed better in recommending the correct restaurant and in full dialogue than MemN2N, however, it performed worse in changes to the requests and responding to user queries. Luo et al. (2019) and Zhang et al. (2019) proposed improvements on the split memory architecture based on this dataset. However, all of these approaches focus on personalising the dialogue based on general attributes (gender and age), instead of adapting to each user. In addition, personalisation was based on single interactions of users, instead of long-term interactions. Moreover, user attributes were pre-defined at the beginning of the conversation, instead of obtained from the interaction. Nevertheless, the split memory architecture may be suitable for personalisation in long-term HRI, since one part of the memory may be used for focusing on individual user attributes.

2.3.1.2 Personalised Dialogue in Human-Robot Interaction

Humans communicate with each other through a variety of ways, such as speech, gestures, expressions, and text and image-based interfaces. Building robots that support multi-modal communication through conversing with humans in natural language, as well as supporting crucial non-verbal communication aspects, will allow a more natural interaction, whereby, ensuring communication and task effectiveness (Mavridis, 2015). In addition, multi-modal interaction is particularly important in personalisation of long-term interactions, in order to achieve user recognition and communication to improve the user experience and engagement, which is a common approach in previous research.

Most of the HRI studies personalise the interaction by using partial or fully Wizard-of-Oz operation in order to choose an appropriate dialogue response, avoid errors in speech recognition or learn from users (e.g., Kanda et al. (2010); Sabelli et al. (2011); Lee et al. (2012); Senft et al. (2015); Leite et al. (2017); Glas et al. (2017); Ahmad et al. (2019); Kennedy et al. (2015)). However, as previously stated in Section 2.1.1, having an operator is not suitable for real-world applications for extended periods of time. Only a few studies (Kasap & Magnenat-Thalmann, 2012; Zheng et al., 2019; Churamani et al., 2017) explored fully autonomous personalisation in dialogue for long-term HRI. However, none of these studies were conducted in the real-world environments.

Kasap & Magnenat-Thalmann (2012) combined short-term and episodic memory in multi-modal personalisation of the teaching style depending on the affective state of the agent and the user, and the agent's relationship with the user over time. They used a rule-based approach based on beliefs, desires (goals) and intentions (actions) (Bratman, 1999) of the agent combined with a hierarchical task network and finite state machine to plan and execute the goals. In addition, a rule-based emotion engine is used to determine and update emotional state and mood of the agent. The users are recognised through face recognition, and user responses are obtained by speech recognition. While this model proves to be a successful implementation of personalisation in increasing user engagement and responsiveness over time, using affect may not be appropriate in some domains (e.g., healthcare).

Also presented in Section 2.1.3, Zheng et al. (2019) used sensory, working and long-term memory based on a rule-based dialogue manager and a knowledge-base to offer personalised activity tracking. Personalisation consisted of learning and recalling the user's name, hobbies, plans and behaviour patterns, and referring to the robot's own failures. A case study of three users was conducted over eight days, which suggested that commenting on meta-behaviour (e.g., user's leaving time from the office) is more successful in inducing positive reactions, followed by recalling shared experience (e.g., robot's failures) and referring to user's previously stated plans, with the least positive reaction being induced by remembering factual information (e.g., hobbies). However, the results cannot be generalised due to the limited number of users, and the unequal and small amount of personalisation patterns.

Churamani et al. (2017) used a rule-based knowledge-base to personalise greetings and conversation by obtaining the user's name and preferences in the initial interaction and recalling it in the subsequent one, during small-talk for an object teaching task. Multi-modal user recognition is used by combining deep learning approaches for speaker identification (Ng et al., 2017) and face recognition (Schroff et al., 2015). However, the user recognition required pre-training. The results showed that personalisation significantly increased the likeability and perceived intelligence and attention of the robot, along with user engagement. In contrast, personalisation decreased the perceived safety and social influence of the robot, and users recommended it less for real-world context, which can be attributed to the complexity of the interaction and noisy environment that caused speech recognition failures. However, the study was only based on two interactions, hence, the novelty effect might have affected the results.

There is no data-driven approach for long-term HRI, due to the necessity for (and the lack of) available corpora to train such systems. In this thesis, we compare and validate the applicability of rule-based systems (Chapter 7 and 9) and data-driven approaches (Chapter 8) to fully autonomous personalisation in dialogue for real-world HRI.

2.4 Summary

This chapter has described the challenges and methodologies in long-term human-robot interactions, user recognition and conversation. Previous literature revealed that personalisation improves user interest and engagement and facilitates building rapport with users. In order to achieve personalisation in real-world interactions over extended periods of time, fully autonomous robots with long-term memory systems that support incremental and online learning are essential. Several long-term studies confirmed these findings through personalising the interaction by recalling user's personal attributes, preferences and behaviour patterns, along with previous shared history with users, particularly in customer-oriented service robotics and socially assistive robotics for healthcare applications in task-oriented interaction. In order to achieve autonomous personalisation of the interaction users should be identified and learned autonomously and incrementally for long-term human-robot interactions, starting from a state without any known users. However, a review on biometric recognition systems showed that there are no user recognition systems developed with such capabilities. In addition, several studies revealed the importance of multi-modal user recognition and natural language-based verbal and non-verbal communication in achieving reliable autonomy and naturalness in the interaction. Research in conversational agents revealed that most task-oriented dialogue systems in HRI use rule-based approaches due to their robustness, however, none has been evaluated autonomously in real-world studies. Additionally, data-driven approaches, which offer flexibility and reduce the costs of laboriously hand-crafting rules, have not been applied to personalisation in long-term interactions with chatbots or robots. Nevertheless, attention mechanisms, particularly End-to-End Memory Networks, show promise in recalling relevant memories for long-term interactions.

Chapter 3

Multi-modal Incremental Bayesian Network with Online Learning for Open World User Identification

Key points:

- Multi-modal Incremental Bayesian Network (MMIBN) is proposed, which is the first method for sequential and incremental learning in open world user recognition that allows starting from a state without any known users. It is also the first multi-modal approach that combines a primary biometric (face recognition) with soft biometrics (gender, age and height and time of interaction) for open world user recognition in human-robot interaction.
- An online learning method for Bayesian networks is proposed based on Voting Expectation Maximization and Maximum Likelihood estimation for modelling noise in modalities and frequency of user appearances.
- Quality of the estimation is introduced to decrease the number of incorrect recognitions that may arise from the combined noise in the identifiers.
- Long-term recognition performance loss is introduced for balancing the trade-off between identifying known users and unknown users for long-term interactions.
- Hybrid normalisation is introduced, which combines the optimal normalisation methods for each parameter in the Bayesian network.

Parts of the work presented in this chapter have been published at the Social Robots in the Wild workshop at 2018 ACM/IEEE International Conference on Human-Robot Interaction (Irfan et al., 2018b)[1] and under review at the ACM Transactions on Human-Robot Interaction journal. The source code for the Multi-modal Incremental Bayesian Network (MMIBN) models is available online[2] for academic use based on the license terms.

[1] Available online at: http://socialrobotsinthewild.org/wp-content/uploads/2018/02/HRI-SRW_2018_paper_6.pdf

[2] https://github.com/birfan/MultimodalRecognition

3.1 Motivation

We explored various architectures suitable for user recognition in long-term Human-Robot Interaction (HRI), as presented in Chapter 2.2, to answer our first research question, *Which user recognition algorithms are applicable to long-term recognition in the real world?*. The core problem that we face within HRI for personalising the interaction is to recognise unknown users and enrol them incrementally and autonomously, which is classified as open world recognition. However, as shown in Chapter 2.2, there exists a limited amount of research on this topic, and none of the available methods is evaluated on user identification. In addition, these methods use batch learning of classes instead of sequential learning, which is unsuitable for HRI, because the users do not present themselves to the robot in batches. In contrast, it is more likely that the same users will be encountered several times before the introduction of another. However, there are no user recognition systems which can support identifying and learning users autonomously and incrementally in long-term interactions, starting from a state without any known users. Thus, this thesis proposes the first open world user recognition system for sequential and incremental learning of users without the necessity of any preliminary training.

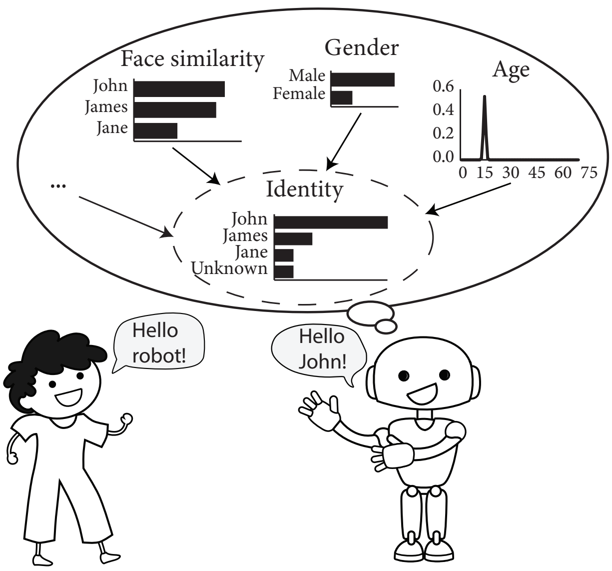


Figure 3.1: Robots can make use of multi-modal information to recognise users more accurately in long-term interactions. *This image contains the artwork of Hoang-Long Cao. Permission was granted for use and modification.*

Uni-modal systems based on primary biometrics, such as face or voice recognition, may not always provide good performance due to noise in the identifiers and the similarities between users. Previous literature revealed that multi-modal biometric recognition helps overcome these problems. Robots are also suitable for multi-modal recognition as they have multiple sensors and perception algorithms (as shown in Figure 3.1), which allow them to recognise users even when data are inaccurate or noisy. Particularly, soft biometrics enables using ancillary physical or behavioural characteristics of a user (e.g., gender and age) to markedly improve the user recognition (Dantcheva et al., 2016; Jain et al., 2011). However, there are no HRI studies that combine a primary biometric with soft biometrics for improving open world user identification in real-time HRI. This thesis explores the use of soft biometrics in user recognition with non-intrusive features, namely, gender, age, height and time of interaction, to combine with a primary biometric, face recognition (FR). While the proposed modalities have shown improvement in recognition when fused separately or with other biometrics (Scheirer et al., 2011; Martinson et al., 2013; Arigbabu et al., 2015; Dantcheva et al., 2016), our approach is the first in combining them.

Adapting to the changes in user appearances can improve user recognition in long-term interactions. In addition, online learning (OL) can help overcome misidentifications due to similarities between users, especially in combination with multi-modal recognition. For instance, a user can be initially mistaken for another in certain circumstances, however, these variations can be learned over time and combined with other modalities to improve recognition where FR fails.

Bayesian approaches are suitable for multi-modal biometric recognition as they allow combining estimates of various identifiers (Bigün et al., 1997; Verlinde et al., 1999; Jain et al., 2004; Scheirer et al., 2011). Moreover, online learning can be applied to Bayesian network (BN)s to adapt the *a priori* likelihoods (Oravecz et al., 2016; Opper & Winther, 1999; Honkela & Valpola, 2003; Bauer et al., 1997; Cohen et al., 2001b; Lim & Cho, 2006; Liu & Liao, 2008), which can be used to learn the similarities between users over time and better adapt the likelihoods according to how identifiers estimate a user's attributes in reality (e.g., age estimation as 20 for a 25 year old user). Accordingly, to accomplish our first research objective (RO1), we propose the Multi-modal Incremental Bayesian Network (MMIBN), which allows incremental and online learning for long-term human-robot interactions[3]. Correspondingly, we propose methods to adapt the Bayesian network to reliably recognise unknown users, extend the architecture for adding new users, combine multiple modalities and apply online learning with an adaptive learning rate for continuous probabilities.

[3] *Continual (or lifelong) learning* in machine learning refers to learning continuously, incrementally and adaptively. However, the definition includes both batch and online learning, whereas, user recognition in long-term HRI should be both incremental and online.

3.2 Methodology

Bayesian network (BN) is a probabilistic graphical model which represents conditional dependencies of a set of variables through a directed acyclic graph. BNs are suitable for combining scores of identifiers with uncertainties when the knowledge of the world is incomplete (Scheirer et al., 2011). The naive Bayes classifier model assumes conditional independence between predictors, which is a reasonable assumption for a multi-modal biometric identifier as the individual identifiers do not affect each other's results.

We propose the Multi-modal Incremental Bayesian Network (MMIBN) that integrates multi-modal biometric information for reliable recognition in open world identification through a naive Bayes model, as shown in Fig. 3.2. The primary biometric in our system is **face recognition** (Face, F), which is fused with soft biometrics (SB), namely, **gender** (G), **age** (A), and **height** (H) estimations and **time of interaction** (Time, T). We hypothesise that the integration of these SB will reduce the effects of noisy data and increase the identification rate. The pyAgrum (Gonzales et al., 2017) library is used for implementing the BN structure.

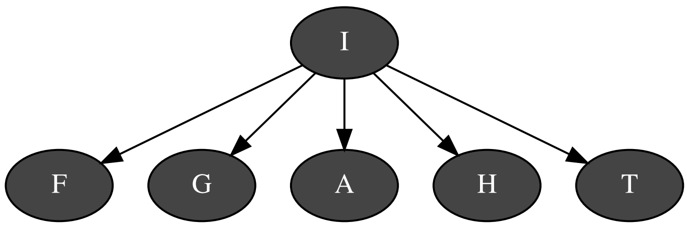


Figure 3.2: The multi-modal incremental Bayesian network model with Identity (I), Face (F), Gender (G), Age (A), Height (H), and Time of interaction (T) nodes for user recognition.

3.2.1 Structure

The number of states for each node depends on the modality: Face and Identity nodes have n_e+1 states, where n_e is the number of enrolled (known) users, and the additional state is the Unknown (U) state; Age and Height nodes are restricted to the available range of the identifier, such as [0,75] for Age and [50,240] for Height; Gender has "female" and "male" states; Time is defined by the day of the week and the time of the day, through time *slots* (e.g., every 30 min is a time slot). An example of the representation within the database is given in Table 3.1.

When a user is encountered, the corresponding multi-modal biometric evidence is collected

Table 3.1: Exemplary database for multi-modal user recognition.

Identity	*Name*	*Gender*	*Age*	*Height (cm)*	*Time(s) of interaction*
1	Jane	Female	25	168	11:10, Tuesday [71st slot]
2	James	Male	27	175	11:20, Wednesday [119th slot]
3	Joe	Male	40	178	11:15, Tuesday [71st slot], 18:40:00, Thursday [182nd slot]

from the identifiers. FR provides similarity scores, which give the percentage of similarity of the user to the known faces in the database. Age, Height, and Time are assumed to be discrete random variables (e.g., age is taken as 26, when it is between 26 and 27) with a discretised and normalised normal distribution of probabilities, $N(\mu, \sigma^2)$, defined by Equation 3.1, where V is the estimated value, Z is the standard score, and C is the confidence of the biometric indicator for the estimated value.

$$\mu = V, \quad P(\frac{-0.5}{\sigma} < Z < \frac{0.5}{\sigma}) = C \tag{3.1}$$

The number of Time states (slots) is determined by the time period (t_p), which can be set depending on the precision required in the application. A smaller time period and standard deviation (σ_t) ensure a higher precision, however, this would increase the complexity of the Bayesian network, thereby, the time to calculate the identity. In addition, a higher precision carries the risks of decreasing the recognition rate, if the users are not encountered near the time slot that they were previously seen. For instance, if users in the application scenario will change every 5 minutes, then $t_p = 5$ min and $\sigma_t = 15$ min would be reasonable. In contrast, in an HRI scenario, $t_p = 30$ min with $\sigma_t = 60$ min can allow better identification because it is less likely to encounter users at the same time every day. Hence, we use the latter in this work.

3.2.2 Weights of the Network

Soft biometric traits are characteristics that are not suitable to identify an individual uniquely. However, previous research shows that they can improve the recognition rate if used in combination with primary biometrics (Dantcheva et al., 2016). Some soft biometric features may contain more information about an individual than others depending on the characteristics of a population. For instance, if there are users that have a particular characteristic (e.g., very tall, or very young), they will be identified more easily within a group. We assume that a large population will have similar characteristics. Nevertheless, some features will still remain more informative, such as age compared to gender. This can be modelled by giving higher weights to the parameters with smaller variability and larger distinguishing capability (Jain et al., 2004).

Similar to the work of Zhou & Huang (2006), we use weights (w_i) as the exponential to the likelihoods of the child nodes (i.e., biometrics, X_i). The posterior probability, $P(I^j|X_1,..,X_n)$, is approximated, as shown in Equation 3.2. I^j stands for the jth user ($I = j$), where I is the Identity node.

$$P(I^j|X_1,..,X_n) \propto \frac{P(I^j)\prod_i P(X_i|I^j)^{w_i}}{P(X_1,..,X_n)} \tag{3.2}$$

We assume that the identifiers perform equally well on all users (Jain et al., 2004), thus, the accuracy of an identifier is independent of the user. Accordingly, equal priors are assumed for each of the identifiers, i.e., $P(X_i^k) = P(X_i^l)$ for all $k \neq l$. The posterior probability simplifies to the equation shown in Equation 3.3.

$$P(I^j|X_1,..,X_n) \propto P(I^j)\prod_i P(X_i|I^j)^{w_i} \tag{3.3}$$

Because the distribution of users over time is not known, one approach for determining $P(I^j)$ is to use adaptive priors using frequencies, as shown in Equation 3.4, where n_{oj} is the number of times user j is observed.

$$P(I^j) = \frac{n_{oj}}{\sum_j n_{oj}} \tag{3.4}$$

However, this can create a bias in the system towards the most frequently observed user as it affects the posterior probability directly, thus, may result in a decrease in the identification rate. Therefore, we assume that the probability of encountering user j is equally likely as encountering user m, hence, we assume equal priors for $P(I)$.

3.2.3 Quality of the Estimation

Algorithms for open-set problems generally use a threshold (e.g., over the highest probability/score) to determine if the user is already enrolled or unknown. However, the resulting posterior probabilities in a BN can be low due to the multiplication of the conditionally independent modalities and vary depending on the number of states. Hence, we use a two-step ad hoc mechanism to transform the BN to allow open-set recognition. (1) Unknown (U) state is used in both Face and Identity nodes. The similarity score of FR for the Unknown state, i.e., the probability of the user being unknown according to the face recognition, is set to the FR threshold (θ_{FR}), such that when normalised, scores below/above the threshold will have lower/higher probabilities than Unknown. This allows maintaining the threshold for the FR system in use. (2) We use the confidence measure called the *quality of the estimation (Q)*. Given the evidence y_t at time t, it compares the highest (winning) posterior probability (P_w) to the second highest (P_s), as shown in Equation 3.5. The difference between the probabilities decreases, as the number of enrolled users (n_e) increases since $\sum_j P(I^j|y_t) = 1.0$. A similar method was used by Filliat (2007) for estimating the quality of localisation based on different images.

$$Q = [P_w(I^j|y_t) - P_s(I^j|y_t)]\, n_e \tag{3.5}$$

Using the quality of the estimation enables decreasing misidentifications. For example, the highest posterior score can be very high, however, if the second highest posterior is very close to it, there are two possible strong candidates for the current user. If the system were to identify the user in this case, the resulting misidentification could cause adverse effects on the current user especially in the case of different genders or age differences between the two users, as well as security issues. Thus, it is preferable to identify the user as unknown, if the quality is below a determined threshold (θ_Q), or if the Unknown state has the highest posterior probability. Otherwise, the identity is estimated with a maximum a posteriori estimation, given in Equation 3.6.

$$j^* = \begin{cases} \mathsf{U}, & \text{if } Q < \theta_Q \text{ or} \\ & P(I^{\mathsf{U}}|y_t) > P(I^j|y_t) \text{ for all } j \\ \arg\max_j P(I^j|y_t), & \text{otherwise} \end{cases} \tag{3.6}$$

3.2.4 Incremental Learning

In HRI scenarios, it is desired to allow the users to enrol in the system, such that they can be recognised at the next encounter. For this, we use an online system, where a user can enrol by entering their name, gender, birth year, and height, and then a photo of the user is taken by the robot. This information is gathered to have the ground truth values for recognition, and for setting the initial likelihoods.

Initially, the system starts as "tabula rasa", where there are no known users. The BN is formed when the first user is enrolled with the Identity node having one state for the new user and one for Unknown. The initial likelihood for the Face node is set to be much higher for the true values as shown in Equation 3.7, where w_F is the weight of the face variable, and n_e is the number of enrolled users. The value was found based on preliminary experiments.

$$P(F^k|I^j) = \begin{cases} 0.9^{w_F}, & \text{if } k = j \\ [0.1/(n_e - 1)]^{w_F}, & \text{otherwise} \end{cases} \tag{3.7}$$

The remaining likelihoods are set using the prior knowledge that the user entered in a similar structure to the evidence for age, height and time variables with a discretised and normalised normal distribution, $N(\mu, \sigma^2)$, where μ is the true value (e.g., age of the person), and σ is the standard deviation of the identifier. Gender is set at $[0.99^{w_G}, 0.01^{w_G}]$ ratio, which is experimentally found.

For the Unknown state, $P(X_i^k|I^u)$ is set to be uniformly distributed, as an unknown user can be of any age, height and be recognised at any time of the day, except for the Face node. Using uniform probabilities or reducing the constant (0.9) in the Face node for the Unknown state resulted in worse recognition performance in preliminary experiments, hence, the likelihood is set as in Equation 3.7.

When a new user is enrolled, BN is expanded by adding a new state to the Identity and Face nodes. For each previous state in the Identity node (including Unknown), $P(F^k|I^j)$ is updated by appending the value corresponding to $k \neq j$ condition in Equation 3.7 ($[0.1/(n_e - 1)]^{w_F}$ with updated n_e) for the new user, and re-normalising the likelihoods. The likelihoods of Gender, Age, Height and Time nodes for the previously enrolled users remain the same. This scalability feature removes the need to retrain the network when a new user is enrolled, thereby, decreasing the time complexity, which can be crucial if the new user is introduced at a later step (e.g., after 1000 users). For instance, if each image corresponding to $\overline{n_o}$ average number of observations per user was to be recognised again after a new user is added to the face database, it would take a significant amount of time to expand the network compared to scaling, since $n_e\ \overline{n_o}\ \mathcal{O}(FR) \gg n_e\ \mathcal{O}(1)$ updates, where $\mathcal{O}(FR)$ is the time complexity of the FR algorithm. In order to allow the network to make meaningful estimations, in the first few recognitions (here, we chose 5 recognitions), the identity is declared as unknown, regardless of the estimated identity.

3.2.5 Online Learning of Likelihoods

BN parameters are generally determined by expert opinion or by learning from data (Koller & Friedman, 2009). The former can cause incorrect estimations if the set probabilities are not accurate enough. The latter, for which Maximum Likelihood estimation is commonly used, is not possible when the BN is constructed with incomplete data. One solution is to use offline batch learning, however, it requires storing data that can cause memory problems in long-term interactions. Another approach is to update the parameters as the data arrive, which is termed online learning. Several approaches exist in the literature for online learning in Bayesian networks, as outlined in Section 2.2.4.

Bauer et al. (1997) proposed using Expectation Maximization (EM) with a learning rate (EM(η)) for online parameter estimation in a Bayesian network, formulated as in Equation 3.8. θ^t_{ijk} represents an entry in the conditional probability table of X_i variable given the value of its parent node Pa_i at time t, that is, the likelihood of variable X_i is $\theta^t_{ijk} = P(X_i = x^k_i | Pa_i = pa^j_i)$. η is the learning rate that controls how much is the past likelihoods is weighed in for the current likelihood. As η approaches 1, the effect of previous data decreases, hence, the update relies more on the present evidence.

$$\theta^{t+1}_{ijk} = \eta\ \frac{P_{\theta^t}(x^k_i, pa^j_i | y_t)}{P_{\theta^t}(pa^j_i)} + (1-\eta)\ \theta^t_{ijk} \tag{3.8}$$

Cohen et al. (2001b) proposed an extension of EM(η) called Voting EM for continuous probabilities and missing (partially observable) data in the evidence (as shown in Equation 3.9) and discrete probabilities and complete (fully observable) data (Equation 3.10)[4].

[4]The equations are presented in the formulation of Bauer et al. (1997) for consistency.

$$\theta_{ijk}^{t+1} = \begin{cases} \eta \, \frac{P_{\theta^t}(x_i^k, pa_i^j | y_t)}{P_{\theta^t}(pa_i^j | y_t)} + (1-\eta) \, \theta_{ijk}^t, & \text{if } P_{\theta^t}(pa_i^j | y_t) \neq 0 \\ \theta_{ijk}^t, & \text{otherwise} \end{cases} \tag{3.9}$$

$$\theta_{ijk}^{t+1} = \begin{cases} \eta + (1-\eta) \, \theta_{ijk}^t & \text{for } P(pa_i^j | y_t) = 1 \text{ and } P(x_i^k | y_t) = 1 \\ (1-\eta) \, \theta_{ijk}^t & \text{for } P(pa_i^j | y_t) = 1 \text{ and } P(x_i^k | y_t) = 0 \\ \theta_{ijk}^t, & \text{otherwise} \end{cases} \tag{3.10}$$

While user recognition is a fully observable problem (i.e., has complete evidence), biometric identifiers may provide varying evidence at each interaction due to noise in the identifiers or the biometric sample, such as varying timing of interactions, lighting conditions, and occlusions. For instance, the estimated age of a person may change from time to time depending on the previously mentioned factors among others, thus, the identifier confidence score may be less than 1. Thus, we cannot assume discrete probabilities, as in Equation 3.10.

The only parent node (Pa_i) in our architecture is the Identity node. Cohen et al. (2001b) suggest that $P_{\theta^t}(pa_i^j | y_t) = 1$ can be assumed with no loss of generality. However, as we discussed before, the noise in the identifiers or the sample causes nonzero values for $P(I^k | y_t)$ for $k \neq j$, where $\sum_k P(I^k | y_t) = 1$. Thus, that assumption does not hold. Similarly, if we use unsupervised learning using $P_{\theta^t}(pa_i^j | y_t)$ term in Equation 3.9, we may update the likelihoods incorrectly. Hence, supervised learning is necessary to achieve accurate online learning. The identity of the user should be known for updating the corresponding likelihoods, which can be achieved in HRI by directly requesting confirmation of the estimated identity, or by obtaining the confirmation implicitly from the dialogue. Thus, $P_{\theta^t}(pa_i^j | y_t)$ becomes $P(I^j) = 1$. Given the evidence at time t and the independently obtained true identity, we can use the Bayes rule to derive, $P_{\theta^t}(x_i^k, pa_i^j | y_t) = P_{\theta^t}(x_i^k | pa_i^j, y_t) \, P_{\theta^t}(pa_i^j | y_t)$, where $P_{\theta^t}(pa_i^j | y_t) = 1$. The resulting formulation of online learning in MMIBN is presented in Equation 3.11. We will refer to the proposed Multi-modal Incremental Bayesian Network with online learning as MMIBN:OL.

$$\theta_{ijk}^{t+1} = \begin{cases} \eta_j \, P_{\theta^t}(x_i^k | I^j, y_t) + (1-\eta_j) \, \theta_{ijk}^t, & \text{if } P(I^j) = 1 \\ \theta_{ijk}^t, & \text{otherwise} \end{cases} \tag{3.11}$$

The learning rate (η) can either be fixed or adaptive, such as using a set of pre-defined values or Maximum Likelihood estimation through dividing by the number of times the parents were equal to their jth configuration. Cohen et al. (2001b) suggest using a fixed learning rate for incomplete data, whereas, an adaptive one for complete data with discrete probabilities. Selecting a fixed learning rate may result in poor performance, if the value is not correctly estimated. In addition, user recognition is a fully observable problem, however, the evidence is noisy and continuous. Hence, we apply Maximum Likelihood estimation to continuous probabilities, in contrast to the work in (Cohen et al., 2001b,a; Liu & Liao, 2008). We use the number of observations of the user j (n_{oj}) to adapt the learning rate, as shown in Equation 3.12. Each observation of the user creates a progressively smaller update on the likelihoods, such that, the effect of a new observation decreases as the number of recognitions of the user increases.

$$\eta_j = \frac{1}{n_{oj} + 1} \tag{3.12}$$

If the user j is not previously enrolled in the system: (1) the Face likelihoods for the Unknown state, $P(F^k|I^u)$, are updated, (2) the user is added to the Bayesian network, as described in Section 3.2.4, (3) online learning is applied to the other states' Face likelihoods $P(F^k|I^j)$, for each user k. Likelihoods of gender, age, height, and time remain the same for Unknown to ensure uniform distribution. If the user is already enrolled, online learning is applied to likelihoods of all modalities for that user.

3.2.6 Long-Term Recognition Performance Loss

The standard metrics for open-set identification are Detection and Identification Rate (DIR) and False Alarm Rate (FAR) (Phillips et al., 2011). DIR is the fraction of correctly classified probes (samples) within the probes of the enrolled users ($\mathscr{P}_{\mathscr{E}}$), given in Equation 3.13. FAR is the fraction of incorrectly classified probes within the probes of unknown users ($\mathscr{P}_{\mathscr{U}}$), given in Equation 3.14. Note that FAR only changes for unknown users, i.e., when a new user is enrolling.

$$\mathrm{DIR} = \frac{|\{\arg\max_j P(I^j|y_t) = j \mid j,\, j \in \mathscr{P}_{\mathscr{E}}\}|}{|\mathscr{P}_{\mathscr{E}}|} \tag{3.13}$$

$$\text{FAR} = \frac{|\{\arg\max_j P(I^j|y_t) = j \mid k,\ j \in \mathscr{P}_{\mathscr{E}},\ k \in \mathscr{P}_{\mathscr{U}}\}|}{|\mathscr{P}_{\mathscr{U}}|} \tag{3.14}$$

In other words, DIR represents the True Positive of enrolled users, in which the current probe (i.e., the multi-modal biometric sample) belongs to a previously enrolled user, and it is identified correctly. FAR serves as the False Positive for unknown users, that is, the probe belongs to an unknown user, but he/she is identified as an enrolled user. However, True Positive and False Positive are notions of *verification* problems, in which the probe is compared against a claimed identity, thus, are generally not applicable to *open-set identification*.

Ideally, DIR= 1.0, when all users are correctly identified, and FAR= 0.0, when all unknown users are identified as such. In reality, there is a trade-off between DIR and FAR, which is determined by the threshold of the identifier, i.e., a higher threshold might decrease FAR, while also reducing DIR. This trade-off is generally represented by a Receiver Operating Characteristic (ROC) curve. The standard practice is to determine the desired FAR, which would then set the identifier threshold and DIR based on this curve.

Depending on the biometric application, the cost of incorrectly identifying a user as known may be very different from the cost of incorrect identification of the enrolled user (Jain et al., 2011). For short-term interactions, in which a user will be encountered one or two times, FAR is as important or more important than DIR. However, for long-term interactions, users will be encountered a greater number of times. Thus, correctly identifying a user (in a closed-set) becomes more important than correctly identifying an unknown user (open-set). Hence, we introduced the *long-term recognition performance loss (*L*)* that creates a balance between DIR and FAR based on the average number of observations per user ($\overline{n_o}$), as presented in Equation 3.15, where α is the ratio of the importance of DIR compared to FAR. Ideally, $L = 0$, when all known and unknown users are correctly identified. This measure was developed due to the necessity to better optimise the quality of the estimation and the weights of MMIBN on the data obtained in Chapter 4.

$$\begin{aligned} L &= \alpha\ (1 - \mathrm{DIR}) + (1 - \alpha)\ \mathrm{FAR} \\ \alpha &= 1 - \frac{1}{\overline{n_o}} \end{aligned} \tag{3.15}$$

3.2.7 Normalisation Methods

The scores from each modality must be normalised into a common range (e.g., [0,1]) to ensure a meaningful combination. It is important to choose a method that is insensitive to outliers and provides a good estimate of the distribution (Jain et al., 2005), such as, normsum (dividing each value by the sum of values), minmax, softmax (Bishop, 2006), and tanh (Hampel et al., 1986), formulated as:

$$\text{normsum}: x_i' = \frac{x_i}{\sum x_i} \tag{3.16}$$

$$\text{minmax}: x_i' = \frac{x_i - \min}{\max - \min} \tag{3.17}$$

$$\text{softmax}: x_i' = \frac{\exp(x_i)}{\sum \exp(x_i)} \tag{3.18}$$

$$\tanh: x_i' = 0.5\left[\tanh\left(0.01\left[\frac{x_i - M}{SD}\right]\right) + 1\right] \tag{3.19}$$

where x_i is the value of the state obtained from FR or from the distribution of probabilities for age, height or time nodes, and x_i' is the normalised probability. M and SD stand for the sample mean and the sample standard deviation of the scores, respectively, instead of the population mean and standard deviation used by Jain et al. (2005).

We initially used normsum method during our user study (in Chapter 4), and then we compared the performance of all the normalisation methods on the data obtained from the study to find the optimal method. However, we noticed that each method performs differently depending on the biometric identifier due to the variations within the data and the information format. Hence, we introduce *hybrid normalisation*, which combines the methods that achieve the lowest long-term recognition performance loss for each modality. In other words, hybrid normalisation uses the best performing normalisation method for each modality.

Extensive tests were made on the dataset obtained in Chapter 4 to get the optimal methods for each modality (Face, Gender, Age, Height and Time). The long-term recognition performance loss was optimised for each combination of the individual modality with face recognition (Face, Face+Gender, Face+Age, Face+Height, Face+Time) while optimising the weights for each of the combinations. The resulting hybrid normalisation uses *normsum* for Face, Gender, and Height; *tanh* for Age; *softmax* for Time of interaction. Subsequently, we optimised the weights of the MMIBN and the quality of the estimation for each normalisation method and compared the results on the multi-modal user recognition dataset that we generated in Chapter 5.

Additionally, BNs use the product rule for combining the results of each node, hence, if a probability of a classifier is zero, it results in an overall zero probability for a class irrespective of the results from other classifier. In our user study, we evaluated the effect of using a small (10^{-6}) cut-off probability threshold (p_t). The effect varied depending on the normalisation method, hence, for our simulated dataset, we set the cut-off threshold equal to the minimum value in the likelihoods obtained from the user study.

3.2.8 Extendability

The presented approach uses only one primary biometric, hence, in the absence of facial information, the image is discarded, and another image is taken, if necessary. It is intended to increase the recognition rate from a single image. However, in order to increase the reliability of the system, multiple images (e.g., 3 images as in Chapter 4) can be taken in succession, and the results can be normalised to estimate the identity of the user. Nonetheless, the system allows extension with other primary biometric traits, such as voice and fingerprint, and other SB, such as eye colour and gait, to improve recognition.

The proposed approach does not require heavy computing, therefore, it is suitable for use on commercially available robots. In addition, it can work with any identifier software on any platform. We employ this system on Pepper (in Chapters 4, 5 and 7) and NAO (in Chapter 9) robots for our experiments. These robots are operated by NAOqi[5] software, which includes different modules that allowed us to extract face similarity scores, gender, height and age estimations from a single image. The estimations from these modalities are fed into the network. The internal states of the proprietary algorithm are inaccessible, hence, we assume that the gender and age estimations are not used to obtain the face similarity scores, and they are conditionally independent of the FR results, even though they are obtained from the 2D image.

In order to benefit research in long-term HRI, we are making the source code of MMIBN and the code to run the system on NAO and Pepper robots available online[6]. We encourage researchers to use MMIBN in their studies and extend it with other biometrics.

[5] `http://doc.aldebaran.com/2-5`

[6] `https://github.com/birfan/MultimodalRecognition`

3.3 Summary

This chapter introduced Multi-modal Incremental Bayesian Network (MMIBN) for fully autonomous user recognition in long-term human-robot interactions, which is the first user recognition method that can continuously and incrementally learn users, without the need for any preliminary training. It is also the first method that combines a primary biometric (face recognition) with weighted soft biometrics (gender, age, height and time of interaction) for improving open world user identification in real-time Human-Robot Interaction. We introduced methods to enable open world recognition in Bayesian networks, in addition to evaluating the quality of the estimation to reduce misidentifications. We proposed an extension of an online learning approach for Bayesian networks that adjusts the learning rate according to the frequency of users and allows working with continuous probabilities arising from uncertainties in the identifiers. We also introduced the long-term recognition performance loss that weighs the importance of correct estimations of known users to the incorrect estimations of unknown users for optimising the parameters of the Bayesian network. The proposed approach can be extended with other biometrics and applied to any commercially available robot due to its computationally lightweight structure.

Chapter 4

Long-Term User Recognition Study

Key points:

- A recognition architecture and study are designed for validating Multi-modal Incremental Bayesian Network (MMIBN) in real-world long-term human-robot interactions and obtaining data to optimise the parameters of the method.
- MMIBN is shown to be suitable for user recognition in the study lasting 4 weeks with 14 participants.
- MMIBN improves the fraction of correctly recognised known users (DIR) compared to face recognition by 1.4% in open-set recognition and 4.4% in closed-set.
- Online learning (OL) is found to decrease the fraction of incorrectly recognised unknown users (FAR), with the cost of decreasing DIR.
- Minmax normalisation method is found to be optimal for non-adaptive MMIBN, whereas, softmax is found to be best for online learning (MMIBN:OL).
- Height is found to be the most effective soft biometric trait, and age is found to be the least.

Parts of the work presented in this chapter have been published in Irfan et al. (2018b). The publication is available online[1] at the Social Robots in the Wild workshop at 2018 ACM/IEEE International Conference on Human-Robot Interaction.

[1] http://socialrobotsinthewild.org/wp-content/uploads/2018/02/HRI-SRW_2018_paper_6.pdf

4.1 Motivation

As revealed in Section 2.2, there are no user recognition systems that allow open world recognition in Human-Robot Interaction (HRI), starting from a state without any known users. In order to address that gap, in Chapter 3, we proposed the Multi-modal Incremental Bayesian Network (MMIBN) as a reliable user identification system for long-term HRI that can identify previously known and new users and incrementally learn new users with multi-modal information. While this system does not need any preliminary training, it is important to find the optimal parameters of the network to ensure reliable identification performance. Consequently, in this chapter, we design a user study to validate that MMIBN is suitable for long-term HRI in the real world and collect data to optimise the system.

Moreover, the proposed system addresses an open challenge in HRI (Kunze et al., 2018) by adapting to the changes in the appearances of users with online learning. However, in order to ensure that the learned biometrics are corresponding to the correct user, supervised learning is necessary. This can be achieved directly by requesting confirmation of identity from the user or indirectly through dialogue. In this study, we choose the former to ensure a reliable collection of data. In addition, new users should be able to enrol themselves to the system, hence, we design a "human-in-the-loop" architecture to enable autonomous identification and enrolment.

4.2 Hypotheses

The primary purpose of the study conducted here is to evaluate whether our proposed approach is suitable in a real-world interaction for long-term HRI. In addition, this study aimed to collect real-world human biometric data to optimise the parameters of the Bayesian network, such as the weights, normalisation method, face recognition threshold and quality of the estimation (Q). Subsequently, the previous literature suggests that multi-modal user recognition with soft biometrics will improve recognition performance (Dantcheva et al., 2016). In addition, online learning has been shown to improve (object) recognition performance (De Rosa et al., 2016). Based on the objective of our study and the findings from previous literature, we derived the following hypotheses for the study, as listed below:

H1 Our proposed Multi-modal Incremental Bayesian Network (MMIBN) is suitable to recognise users autonomously and incrementally for long-term human-robot interactions in the real world.

H2 MMIBN will improve user recognition compared to face recognition (FR), as measured by an increase in the Detection and Identification Rate (DIR) of known users.

H3 Online learning (MMIBN:OL) will improve user recognition over a non-adaptive model (MMIBN), as measured by an increase in DIR.

4.3 Applying MMIBN to Human-Robot Interaction

As stated in the previous chapter, MMIBN can be applied to any platform. In this study, we chose the Pepper robot (SoftBank Robotics Europe) for the following reasons: (1) It is a commercially available robot that shares the same NAOqi[2] software with NAO robot (SoftBank Robotics Europe), removing the need to implement new architectures for each robot that will be used in different projects throughout this thesis. (2) NAOqi has its proprietary identifier algorithms for face recognition, gender, age and height estimation. (3) Pepper and NAO robots are widely used in the literature, hence, it is desired to create an architecture that will benefit the community for future long-term HRI studies[3]. (4) Pepper has a tablet interface, which enables reliable confirmation of identity, as shown in Figure 4.1, which is fundamental to this study.

Figure 4.1: A user is interacting with the Pepper robot during the user study, through its tablet interface to confirm the identity that is estimated.

[2] `http://doc.aldebaran.com/2-5`

[3] We release the code for MMIBN and the recognition architecture in this study on: `https://github.com/birfan/MultimodalRecognition`

In order to enable MMIBN to recognise and learn users autonomously, it is necessary to design an architecture that allows autonomous user detection and enrolment. Correspondingly, we designed the *Recognition Module* (Figure 4.2) and the *Recognition Architecture* (Figure 4.3). User recognition starts when a person is detected by estimating their height and marking the time of interaction. Subsequently, when a face is detected, the user's picture(s) is taken, and it is analysed with face recognition, gender and age estimations. The obtained modalities are input to MMIBN, which estimates the identity as described in Chapter 3. The user is asked to confirm the estimated identity (step 4): if the identity is estimated as Unknown, using the phrase *"I'm sorry, I couldn't recognise who you are! Could you enter your name on the tablet please?"*, or if the user is estimated as a known user, with the utterance *"Hello USER_NAME, it is nice to see you again! Could you confirm that it is you please?"*. If the identity is wrongly estimated, the name of the user is requested. The estimated identity and recognition values are stored for data analysis (step 6). If the user is previously known, online learning of the likelihoods is performed for the known user, as described in Section 3.2.5. Finally, user and face recognition databases are updated with the new information. The video demonstration of the interaction is available online[4].

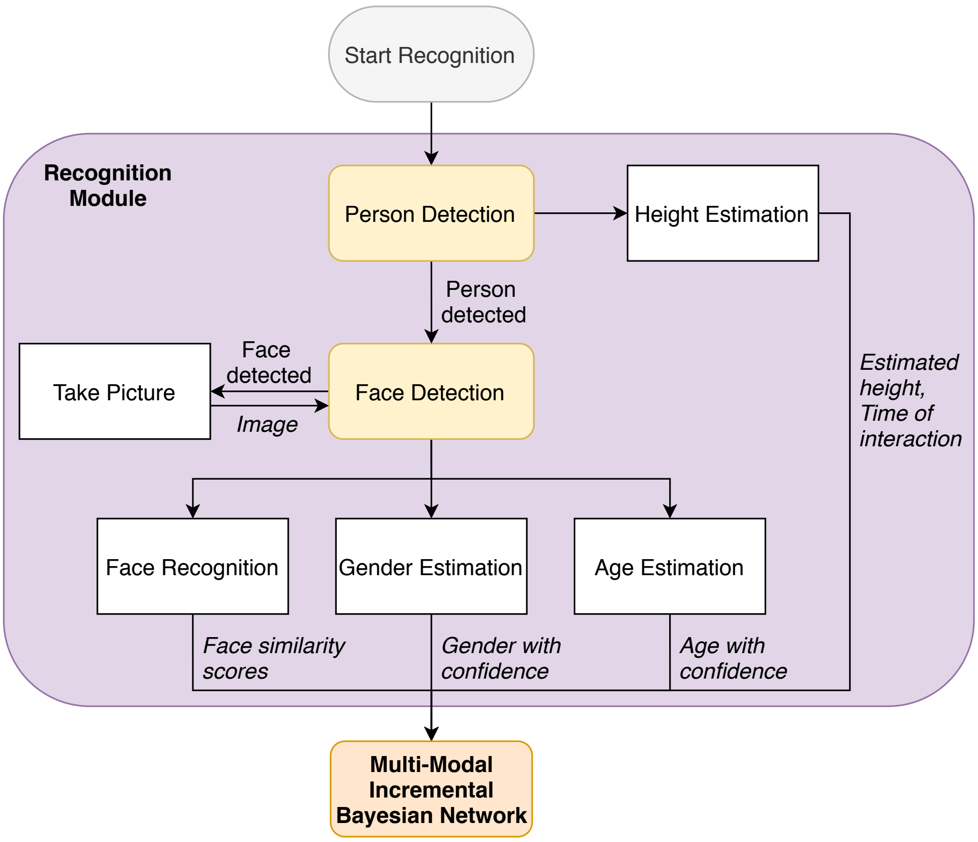


Figure 4.2: Diagram of the Recognition Module. The yellow highlighted modules are proprietary software within NAOqi that are used to obtain the estimated modalities.

[4]Known user interaction: `https://youtu.be/Ix98k6_-2Zc`

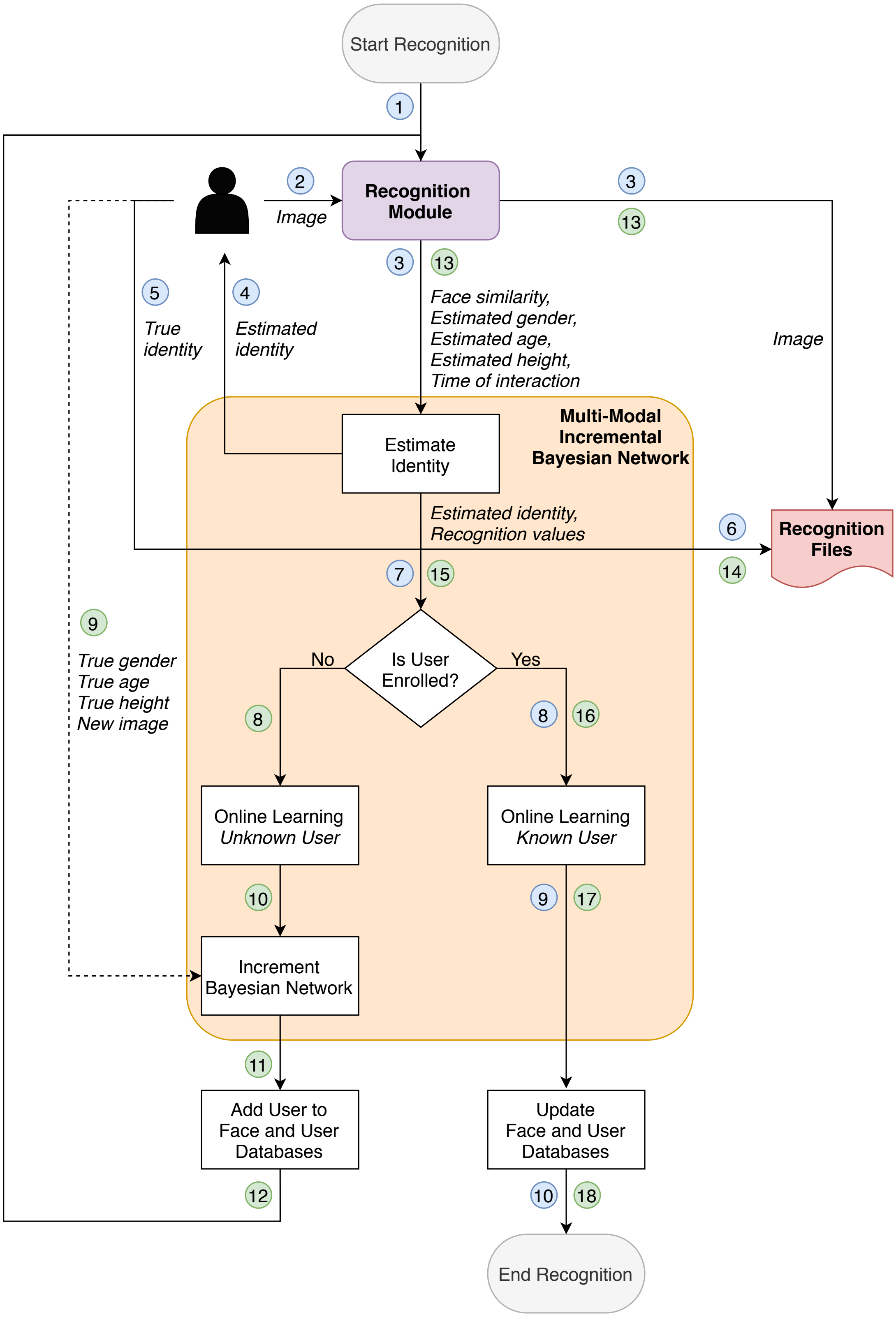


Figure 4.3: Diagram of the Recognition Architecture. First 7 actions are common to both known and new users. Actions 8-18 (in green) are performed for new users, whereas, 8-10 (in blue) are performed for known users. Dashed line shows that ground truth values for gender, age and height are requested from the user and a new image is taken as input when the user is enrolling. The system also allows using estimated values for enrolment.

If the user is new, the Face likelihoods for the Unknown state are updated. Afterwards, in order to obtain reliable ground truth values, the user's gender, age, and height are requested sequentially through Pepper's tablet interface (Figure 4.4), followed by a picture taken by the robot. MMIBN is expanded for the new user, and the user is added to the face recognition and user databases. Subsequently, the identification is performed once again on the new image with the previously described steps for the known user.

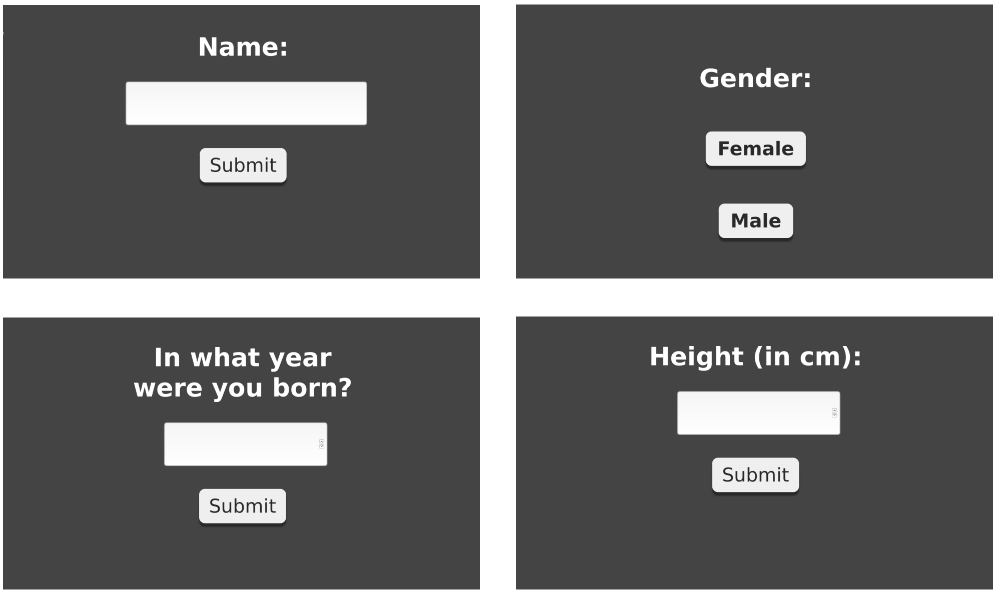


Figure 4.4: Enrolment questions for name, gender, birth year and height.

As previously stated, this user study is aimed to collect real-world data to optimise the parameters of the MMIBN. However, in order to ensure that user recognition is reliable for the study, we conducted preliminary experiments to set the weights and the cut-off threshold. The normalisation method was set as normsum, and the quality of the estimation was set to be 0. The preliminary tests were not enough to accurately determine the face recognition threshold, hence, we used the maximum similarity score per image to set the similarity of Unknown state as 1−max(FR). MMIBN:OL was used to measure the capability to track the times of interaction, learn the similarities between users and determine the identifier estimations of users.

4.4 Experimental Procedure

In order to validate that our proposed user recognition approach is suitable for real-world interactions, we designed a long-term HRI study of 4 weeks in an office of the Centre for Robotics and Neural Systems (CRNS) at the University of Plymouth. In order to ensure a natural level of interaction that frequently occurs throughout the day, we placed the robot at the office kitchen. When students are taking a coffee or lunch break, the robot would autonomously recognise them and request the confirmation of the estimated identity or enrol them into the system, if they have not been previously introduced. Moreover, the participants were encouraged to greet the robot when they arrive and before they leave, but no specific times were enforced to interact with the robot, that is, the participants interacted with the robot whenever they wished. The experimenter was present in the office for the first week of the study, in case of any operational errors, but did not interfere with the experiment.

4.4.1 Participants

The study involved 14 participants (10 males, 4 females) with an age range of 24-40 (M = 29.6, SD = 4.5). The participants were PhD students or researchers that are either working in the office or visiting it frequently. Each participant enrolled in the system within the first week of the study, however, they were not enrolled consecutively, which resembled a real-world situation.

The participants signed an information and consent form under the University of Plymouth ethical approval for image collection and sharing (see Appendix A).

4.4.2 Behaviour of the Robot

While the participants only interacted with the robot through the tablet interface or touch, the robot communicated its requests and feedback verbally to capture the participant's attention. In order to encourage the participants to interact with the robot frequently and achieve a more natural interaction, we used positive responses to the confirmations. The robot replied with a phrase randomly chosen from a list of positive sentences after the correct recognition of a person, such as *"You look very good today"*, *"I feel much better every time I see you"* or *"I knew it was you, just wanted to be sure!"*. If the recognition was incorrect, the robot would apologise for its mistake and either accept its mistake in a positive manner, state that the person is looking different or say that it was a joke.

In order to reduce the effort required by the participants, the robot did not interact with the same user within 30 minutes period, unless the user voluntarily touches the head of the robot, which triggers the recognition. Moreover, if the user does not interact with the robot for 2 minutes after it asks for confirmation or requests the name, the collected images would be discarded and the robot would return to idle mode. This "reset" feature could also be achieved by the experimenter by touching the robot's bottom bumper in case of any system failures, but this feature was not used by (or told to) the participants.

The robot stayed in a fixed position before the interaction to achieve good quality images. When a user was identified, multiple pictures (here, 3) are taken consecutively. If a face is detected in an image, it would be analysed for identity, and the results of each image will be combined to estimate the identity. If a face is not detected in the taken image, it will be discarded. In order to request the confirmation of the estimated identity, the robot became animate to ensure a natural interaction.

The robot was not shut down throughout the duration of the study, but it was put to sleep mode in the evenings in order to release the tension in the motors. The experimenter or the users could wake the robot up by touching the back of its (left) hand. However, the robot could only be put to sleep by the experimenter via remote connection.

4.4.3 Measures

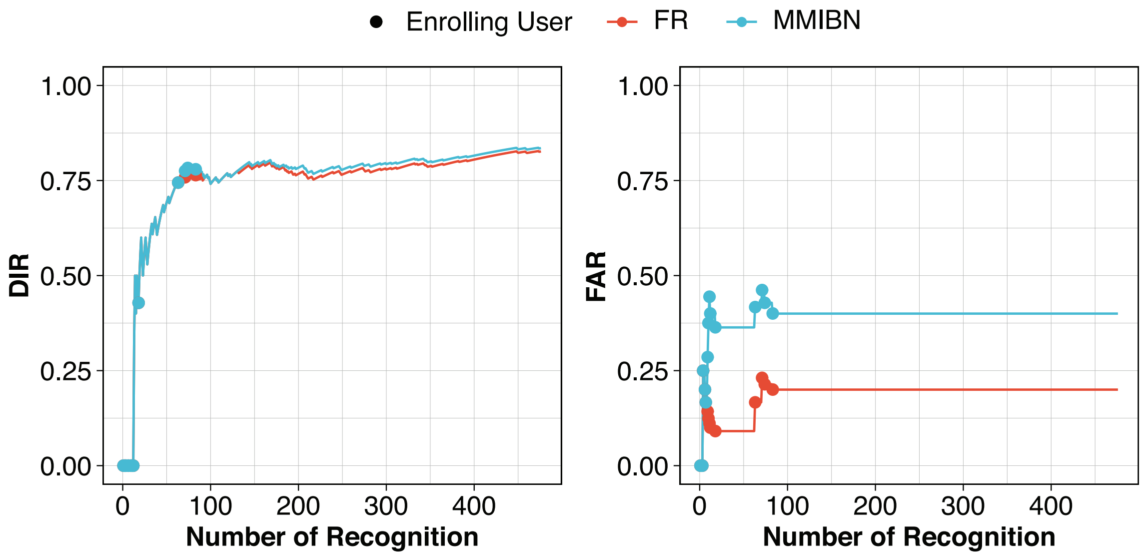
As described in Section 3.2.6, Detection and Identification Rate (DIR) (Equation 3.13) and False Alarm Rate (FAR) (Equation 3.14) are reported along with receiver operating characteristic (ROC) curves, which are the performance measures for the open-set identification problem (Phillips et al., 2011). In addition, Failure to Enroll error (FTE) is reported, which corresponds to the fraction of images where a face cannot be detected.

4.5 Results

4.5.1 User Study

The study was successfully run fully autonomously (with only a few minor interventions for tablet crashes or heated motor alerts) for four weeks in a real-world environment. We reached a total of 476 recognitions, with users interacting with the robot between 24 to 62 times over four weeks period, corresponding to a total of 66 to 175 images per user. DIR improved with the increasing number of recognitions as shown in Figure 4.5 (filled circles correspond to a new user enrolling), and reached 0.835 for Multi-modal Incremental

Bayesian Network with Online Learning (MMIBN:OL), offering a slight improvement over FR (0.826). On the other hand, FAR was lower for FR (0.2) than MMIBN:OL (0.4), which is due to the consecutive enrolment of a few users with similar characteristics. Nevertheless, these findings show that MMIBN satisfies our first research objective and supports our hypothesis (**H1**), that is, the proposed approach is suitable for autonomous user recognition for long-term human-robot interactions in the real world.



(a) (b)

Figure 4.5: The change of (a) Detection and Identification Rate (DIR) and (b) False Alarm Rate (FAR) with the proposed approach (MMIBN) and the face recognition (FR) for the increasing number of recognitions. Ideally DIR= 1.0 and FAR= 0.0. FAR only changes for unknown users, i.e., when a new user is enrolling (depicted with filled circles). The results show that MMIBN:OL offers a slight improvement over FR in DIR in the expense of increased FAR.

Online learning was successful in learning the times that the users interacted with the robot. Figure 4.6 shows the time of interaction probabilities derived from the Time likelihoods in MMIBN for the first week of the study, which reveals that users interacted with the robot at various times during the day. Each likelihood is normalised within itself to sum to 1. Thus, while some users (e.g., the user represented with a green line and triangles) have peaks in the data because they only appeared during certain times or were encountered less, the others (e.g., for the user corresponding to the light blue line) are distributed evenly because they have been encountered more frequently throughout the day. One interesting observation resulting from this graph is that although the workday lasts typically from 9 am to 5 pm (17:00), some users stayed late (until 8-10 pm) in order to finish their daily tasks (e.g., experiments, writing papers). In fact, one user is seen around 1 am on Friday. The users were not encountered on the weekend, as expected. On another note, after the

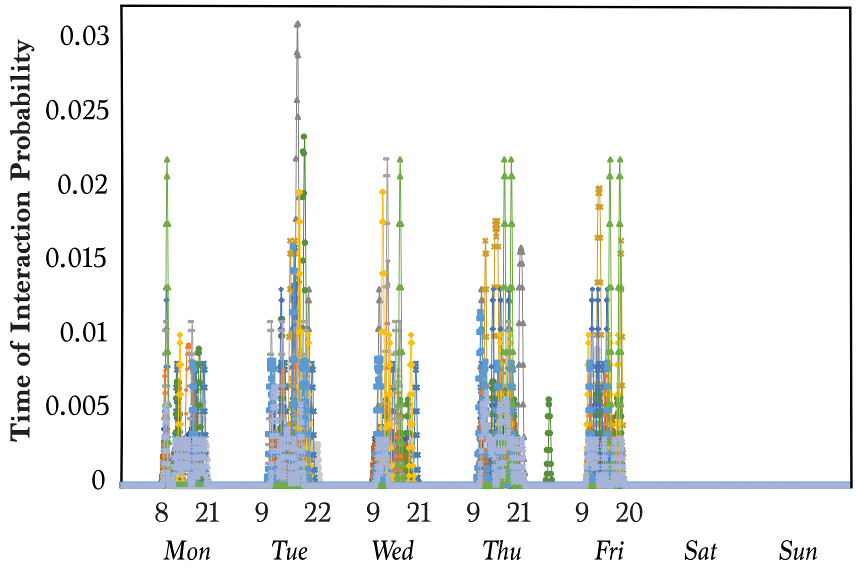
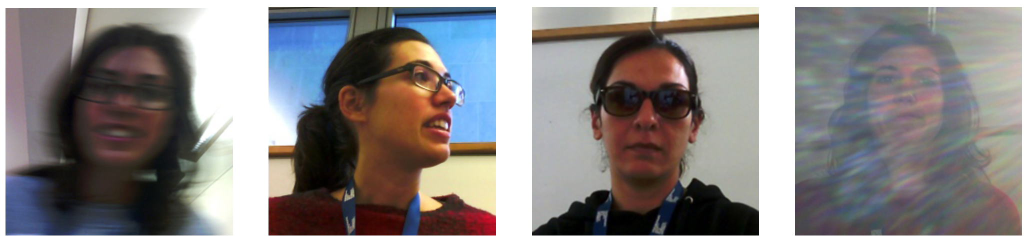


Figure 4.6: Time of interaction probabilities for 14 users derived from the Time likelihoods in MMIBN at the end of the first week of the user recognition study.

study, two participants reported that they frequently interacted with the robot throughout the day, because they enjoyed hearing its positive comments (e.g., "You look very good today!"), which indicates the importance of positive feedback in maintaining user interest in long-term interactions.

The recognition process took approximately 5 seconds: ~2-3s for user detection, ~1.5s (0.5s each per image) for image capture, ~1s to load the network parameters, ~0.6s (0.2s each) for recognition from modalities, ~0.9s (0.3s each) for estimation of the identity using MMIBN. Since the robot did not notify participants when taking images, some of the captured images include people looking sideways, partially covering their faces or moving (see Figure 4.7). In addition, occlusions (e.g., sunglasses) and lighting conditions, such as direct light from the windows, caused problems in FR, resulting in a mean FTE of $M = 0.214$ ($SD = 0.008$). The identity was not estimated by the MMIBN in those cases because the only primary biometric in our system is FR and soft biometrics do not have the deterministic characteristic to estimate the identity on their own.



(a) (b) (c) (d)

Figure 4.7: Examples of unreliable face recognition from the user study: (a) a blurry image; (b) an oblique viewing angle; (c) occlusions, e.g., sunglasses; (d) bad lighting condition.

4.5.2 Optimisation of Parameters

While the proposed approach showed improvement over FR with the preliminary parameters, optimising these parameters would help improve the identification performance further. Hence, we performed optimisation for the network parameters by evaluating each normalisation method with varying weights, learning and cut-off settings on 5-fold cross-validation. We ranged the weights from 0 to 1 with 0.1 increments, used either no cut-off threshold or used a small value (10^{-6}) and set the face recognition threshold $\theta_{FR} = 0.3$, based on the user study results. Cross-validation was performed on the same amount of images per person (65 images per user) collected from the user study. We divided the images into five bins and used four bins as the training set (open-set recognition) and one bin as the test set (closed-set recognition) for each fold, where the test bin was different at each fold. We replaced the "human-in-the-loop" with an offline system that would simulate an interaction by feeding images, time of interactions and estimated heights (from the user study) to the Recognition Module (Figure 4.2) and the true identity to the Recognition Architecture (Figure 4.3). For each fold, we randomised the ordering of users to reduce the bias due to order. Each image contained only a single user corresponding to the true identity and accounted for a single recognition instead of multiple images per recognition. The resulting mean cross-validation DIR and FAR for the optimised weights are given in Table 4.1 (on the next page).

Note that the DIR and FAR of FR is different (higher) than the results in the user study because cross-validation is applied on the partial data from the user study, the data is cleaned of any incorrect identities prior to cross-validation, a face recognition threshold is used, and the ordering of users has an effect on the results, which were averaged to remove the bias in the analysis.

The cross-validation results show that our proposed MMIBN can improve the recognition performance, by a maximum of 1.8% increase in DIR for open-set recognition and 2.2% in closed-set recognition, supporting our hypothesis (**H2**). However, the results show that online learning does not improve performance, thus, **H3** is not supported. While normsum and minmax methods provide good results for a non-adaptive MMIBN, the identification rate drops below FR with learning, whereas, softmax and tanh methods are not markedly affected.

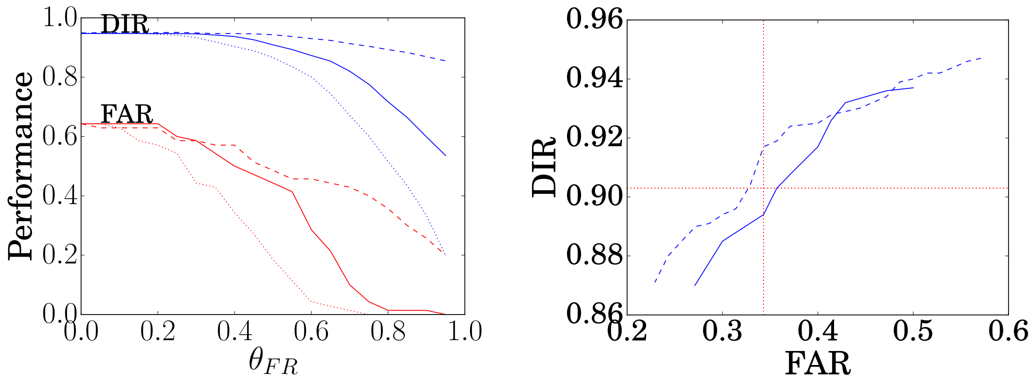
FAR of MMIBN or MMIBN:OL with any normalisation method is greater than the FAR of

Table 4.1: The mean results of the 5-fold cross-validation: optimised weights, DIR for training and closed-set test sets and FAR, for each normalisation method with varying models and cut-off threshold settings. Highlights in blue show the best values, and highlights in red show the chosen methods.

Model	*Normalisation*	*Cut-off Threshold* (p_t)	*FAR*	*DIR*$_1$ *(Training)*	*DIR*$_1$ *(Test)*	w_G	w_A	w_H	w_T
FR	none	none	0.443 (0.078)	0.933 (0.004)	0.945 (0.015)				
MMIBN	normsum	none	0.629 (0.032)	0.951 (0.004)	0.967 (0.013)	0	0	0.1	0
MMIBN	normsum	10^{-6}	0.529 (0.081)	0.943 (0.003)	0.956 (0.015)	0	0	0.1	0.1
MMIBN	minmax	none	0.629 (0.032)	0.951 (0.005)	0.965 (0.015)	0.2	0	0.1	0
MMIBN	minmax	10^{-6}	0.586 (0.060)	0.949 (0.005)	0.965 (0.014)	0.2	0	0.1	0
MMIBN	softmax	none	0.571 (0.072)	0.947 (0.004)	0.965 (0.014)	0.1	0	0.6	0
MMIBN	softmax	10^{-6}	0.571 (0.072)	0.946 (0.003)	0.959 (0.015)	0.1	0.1	0.1	0.1
MMIBN	tanh	none	0.571 (0.051)	0.942 (0.005)	0.955 (0.012)	0	0	0.1	0
MMIBN	tanh	10^{-6}	0.543 (0.039)	0.942 (0.003)	0.957 (0.013)	0	0	0.3	0.1
MMIBN:OL	normsum	none	0.629 (0.032)	0.782 (0.063)	0.694 (0.093)	0.1	0	0.1	0
MMIBN:OL	normsum	10^{-6}	0.571 (0.072)	0.75 (0.082)	0.632 (0.127)	0.1	0	0.1	0
MMIBN:OL	minmax	none	0.629 (0.032)	0.776 (0.064)	0.692 (0.090)	0	0	0.1	0
MMIBN:OL	minmax	10^{-6}	0.643 (0.0)	0.776 (0.061)	0.697 (0.089)	0	0	0.1	0
MMIBN:OL	softmax	none	0.586 (0.060)	0.946 (0.005)	0.961 (0.017)	0.1	0	0.6	0
MMIBN:OL	softmax	10^{-6}	0.586 (0.060)	0.946 (0.005)	0.954 (0.025)	0.1	0	0.6	0
MMIBN:OL	tanh	none	0.571 (0.051)	0.943 (0.007)	0.955 (0.012)	0	0	0.1	0
MMIBN:OL	tanh	10^{-6}	0.543 (0.039)	0.943 (0.006)	0.961 (0.019)	0	0	0.3	0

FR. This is caused by the combination of multi-modal data. For example, if the highest face similarity score is below the threshold, the FR reports the user as "unknown". The network, on the other hand, will still try to identify the user based on other sensor input, where errors might increase FAR.

In order to compare the effects of learning, we chose the minmax method without learning and with a cut-off threshold ($\text{MMIBN}_{\text{minmax}}$) and the softmax method with learning and no cut-off threshold ($\text{MMIBN:OL}_{\text{softmax}}$), because the former provides the second highest DIR but with lower FAR than that of the best methods (highlighted in blue), and the latter provides the best DIR in learning in both training and test sets. In order to determine the ideal face recognition threshold (θ_{FR}), we conducted several cross-validations with varying thresholds. The results are presented in Figure 4.8a.



(a) (b)

Figure 4.8: ROC curves, where dotted lines represent FR results, dashed line is $\text{MMIBN}_{\text{minmax}}$, solid line is $\text{MMIBN:OL}_{\text{softmax}}$: (a) Performance measures, DIR (in blue) and FAR (in red), for varying θ_{FR}; (b) ROC curve for varying Q values for $\theta_{FR} = 0.4$. $\text{MMIBN}_{\text{minmax}}$ is able to perform in the top left zone, where DIR is higher than FR and FAR is lower.

The trade-off between DIR and FAR is apparent from the results. The ideal FR threshold (θ_{FR}) should maintain a low FAR with a good DIR. For example, at $\theta_{FR} = 0.7$, the FAR is very low for both FR and $\text{MMIBN:OL}_{\text{softmax}}$, however, the DIR has also decreased substantially. Hence, we compare the results in the range where $\text{FAR}_{\text{FR}} \leqslant 0.5$ (corresponding to $\theta_{FR} = 0.3$) and $\text{DIR}_{\text{FR}} \geqslant 0.8$ (corresponding to $\theta_{FR} = 0.6$). $\text{MMIBN}_{\text{minmax}}$ is better in identification, by providing DIR in 0.93-0.949 range, whereas $\text{MMIBN:OL}_{\text{softmax}}$ provides between 0.873-0.946, both higher than that of FR (0.801-0.933). However, online learning allows lower FAR (0.286-0.543) than $\text{MMIBN}_{\text{minmax}}$ (0.457-0.571). Within this range, the least FAR for the highest DIR for both proposed methods is obtained at $\theta_{FR} = 0.4$.

Based on the established FR threshold, we compared the effects of the quality of the estimation of estimation (Q) (see Figure 4.8b). As the quality of the estimation increases, the DIR and FAR decreases. If the problem is treated as a closed-set problem (where all the users are enrolled and FAR is not relevant), the DIR of $\mathrm{MMIBN_{minmax}}$ is 4.4% higher than that of FR ($Q = 0$). For the open-set identification problem, the "area of improvement" is where $\mathrm{FAR} \leqslant \mathrm{FAR_{FR}}$ and $\mathrm{DIR} \geqslant \mathrm{DIR_{FR}}$, corresponding to the upper left quartile. Where the FAR of both models are equal ($Q = 0.31$), the DIR of $\mathrm{MMIBN_{minmax}}$ is 1.4% higher than that of FR. If we continue in this range until the DIR of both models are equal ($Q = 0.41$), then $\mathrm{MMIBN_{minmax}}$ achieves 1.4% lower FAR than that of FR. $\mathrm{MMIBN{:}OL_{softmax}}$ does not provide a value in the area of improvement, hence, we can conclude that the proposed online learning model performs worse than the non-adaptive model.

4.6 Discussion

While the proposed approach and the Recognition Architecture allowed real-time recognition in long-term HRI requiring only a low time complexity (0.3 seconds), the time it takes to calculate the estimated identity increases with the number of recognitions and multiple images may result in a noticeable amount of delays in the recognition.

Our user study and cross-validation on the data showed that the FR method used as our primary biometric could provide a good DIR ($88 - 93\%$), and MMIBN improves it further, even for a large number of recognitions. However, we had a small population in the experiment, hence, the face recognition became better at identifying the individuals. In order to ensure that FR and MMIBN are suitable for user recognition for real-world applications, these methods should be evaluated on a larger population, which is difficult to obtain in a real-world long-term HRI study as mentioned in Section 2.1.1.

While previous literature suggests that online learning can improve open world recognition for objects (De Rosa et al., 2016), our results demonstrated otherwise for user identification for known users. On the other hand, online learning is shown to decrease the FAR. In this study, we concluded that MMIBN is more appropriate for long-term interactions because DIR is more valuable than FAR due to long-term interactions involving more known user interactions than new ones. However, the trade-off between DIR and FAR makes it difficult to choose the best set of parameters.

Although online learning was successful in learning the interaction patterns of users, we can also observe (from Figure 4.6) that some of the users interact with the robot at similar times, which makes it difficult for time to be a distinguishing feature for identification. This is also supported by the optimised weights in Table 4.1, which is mostly zero for Time. However, Time can be an important variable in identifying users when users are interacted in a timely fashion, such as recurring weekly appointments in a hospital.

For the remaining optimised weights, we can see that age is the least effective soft biometric in determining the identity (because it is mostly zero), whereas height is the most effective one. However, this might be due to the characteristics of the population in the study, as the participants' ages are close to each other. Another important factor is the reliability of the age recognition software. The standard deviation of the estimated age of a user on average was found to be $\sigma_{\mathtt{Age}} = 9.3$. Hence, we cannot conclude that age should not be used to supplement the FR in general, but if used, the accuracy of the software used should be high, especially in a population with a narrow age range. On the other hand, the effectiveness of the height can also be explained by the nature of the population, because we had 3 relatively tall ($>$ 180 cm) and 2 relatively short ($<$ 160 cm) users, albeit the average standard deviation of $\sigma_{\mathtt{Height}} = 6.3$ cm. We believe that a larger and more balanced dataset would allow observing the true effects of these parameters.

4.7 Summary

The real-world user study showed that the proposed Multi-modal Incremental Bayesian Network (MMIBN) and the designed recognition architecture are suitable for real-time user recognition in long-term human-robot interactions, as they allow fully autonomous user enrolment and recognition, in addition to improving the identification rate, confirming two of our hypotheses. Contrary to our initial hypothesis, while online learning was able to learn the behaviour patterns of the users correctly, it did not improve the known user identification rate (DIR), however, it was shown to decrease the incorrect identifications for new users (FAR). The real-world user study also enabled us to optimise the parameters of the MMIBN using 5-fold cross-validation. Minmax was found to perform best as the normalisation method for the non-adaptive MMIBN, whereas, softmax was the most suitable for online learning. Overall, height was found to be the most effective soft biometric, whereas, age the least. However, the results might be biased due to the small population size and the characteristics of the population.

The additional evaluations using varying face recognition threshold and the quality of the estimation showed that increasing either of these parameters can decrease the FAR, resulting in a decrease the DIR as well. The proposed MMIBN models performed best at 0.4 as the face recognition threshold. MMIBN increased the DIR by 4.4% for closed-set recognition and 1.4% for open-set recognition compared to face recognition.

Chapter 5

Multi-modal Long-Term User Recognition Dataset

Key points:

- Multi-modal Long-Term User Recognition Dataset created with 200 users of varying characteristics, based on the IMDB-WIKI dataset (Rothe et al., 2015, 2018) and artificially generated height and interaction times, for evaluating user recognition models with a large number of users.
- Multi-modal Incremental Bayesian Network (MMIBN) models are shown to perform equally well for uniform or patterned timing, equal or varying frequency of appearance, learning users sequentially or at random intervals, and across training and open-set datasets, and between closed-sets.
- Hybrid normalisation outperforms the individual normalisation methods.
- MMIBN models are shown to significantly outperform face recognition, soft biometrics and a state-of-the-art open world recognition algorithm, by providing lower long-term recognition performance loss and higher identification rate.
- Online learning does not outperform a non-adaptive MMIBN, but decreases the bias in recognition performance between users.

Parts of the work are under review at the ACM Transactions on Human-Robot Interaction journal. The Multi-modal Long-Term User Recognition Dataset[1] and the source code for the Multi-modal Incremental Bayesian Network (MMIBN) models[2] are available online for academic use based on the license terms.

[1] https://github.com/birfan/MultimodalRecognitionDataset
[2] https://github.com/birfan/MultimodalRecognition

5.1 Motivation

We evaluated the proposed Multi-modal Incremental Bayesian Network (MMIBN) in a real-world long-term Human-Robot Interaction (HRI) study in Chapter 4. The results showed that: (1) the proposed model is suitable for user recognition in real-time for long-term HRI, (2) MMIBN with optimised parameters improves recognition performance by increasing Detection and Identification Rate (DIR) and decreasing False Alarm Rate (FAR) compared to face recognition (FR), (3) Multi-modal Incremental Bayesian Network with Online Learning (MMIBN:OL) decreases FAR at the cost of decreasing DIR, hence, a non-adaptive model is found to be better for user recognition, (4) height is the most effective soft biometric in identifying users, whereas age is the least. However, the limited population size (14 users) and the narrow age range (24-40) of the users in that experiment prevented us from claiming that the results can be generalised for application in larger populations.

Obtaining a dataset which encapsulates a diverse set of characteristics for a large number of users over long-term interactions is a laborious task in HRI, as discussed in Section 2.1.1. Thus, this chapter describes Multi-modal Long-Term User Recognition Dataset that we created from images of 200 celebrities obtained from the IMDB-WIKI dataset (Rothe et al., 2015, 2018), in combination with artificially generated height and time of interactions to simulate a long-term HRI scenario similar to the one in our earlier work. The IMDB-WIKI dataset contains cropped faces from images taken at events or still frames from movies. We used proprietary algorithms of the Pepper robot to obtain multi-modal biometric information from these images (face, gender and age), similar to the simulated cross-validations in the previous chapter. On this dataset, we evaluate our proposed MMIBN in comparison to online learning (OL), face recognition (FR), soft biometrics (SB), and a state-of-the-art open world recognition algorithm Extreme Value Machine (EVM).

There were several other challenges and confounding factors in the user study that are analysed further in this chapter. In the interest of overcoming the trade-off between DIR and FAR, we base our comparison on the proposed long-term recognition performance loss (L), presented in Section 3.2.6. While minmax and softmax were found to be the best normalisation methods for non-adaptive and online learning in the previous chapter, further analysis of the results suggested that the optimal normalisation method may vary depending on the biometric trait, hence, we analyse the effects of combining normalisation methods under hybrid normalisation introduced in Chapter 3. We investigate the effect of timing patterns in interactions on the user recognition performance that could vary depending on the HRI application, as time was perceived to be an influencing factor in the previous chapter. Furthermore, we examine the frequency of user appearances, based on users that appear at different frequencies, similar to the user study in the previous chapter, in comparison to all users appearing with the same number of times, similar to the previous cross-validation evaluations. In addition, we evaluate the stability of the proposed approach by comparing enrolling sequentially, which is similar to batch learning, to having repeated interactions of previously enrolled users before the introduction of a new user, similar to the previous user study and cross-validation evaluations.

5.2 Multi-modal Long-Term User Recognition Dataset

To the best of our knowledge, the only publicly available dataset that contains the soft biometrics used in our system (except interaction time) with a dataset of images is BioSoft (Sadhya et al., 2017). However, due to the low number of subjects (75), and the lack of numeric height values, we decided to create Multi-modal Long-Term User Recognition Dataset.

IMDB-WIKI dataset (Rothe et al., 2015, 2018) is chosen as the image resource for our dataset because it contains more than 500k images of celebrities with gender and age labels. 200 celebrities are randomly sampled out of 20k celebrities, choosing only celebrities which have more than 10 images each corresponding to the same age. The resulting dataset contains 101 females, 98 males and one transgender person. In the dataset, each image of the user was chosen from the same year in order to simulate an open world HRI scenario, where the users will be met in consecutive days or weeks. The images that correspond to an age that is within the five most common ages in the set were randomly rejected during the selection to ensure a diverse set of ages. The resulting age range is 10 to 63, with the mean age of 33.04 ($SD = 9.28$). We assume that IMDB-WIKI dataset offers a diverse set of characteristics and soft biometrics.

In the scope of this work, only one user is assumed to be present in each image. Hence, the cropped faces of IMDB-WIKI dataset is used, and the dataset is cleaned in three steps: by removing (1) images with a resolution lower than 150x150, (2) images without a face detected by NAOqi, (3) images that erroneously correspond to another person. Figure 5.1

Figure 5.1: Samples of images from IMDB-WIKI dataset (Rothe et al., 2015, 2018), used in creating the Multi-modal Long-Term User Recognition Dataset.

shows samples of images used from the IMDB-WIKI dataset to create the Multi-modal Long-Term User Recognition Dataset. Because the images come from movies, TV series and events, they may contain bad lighting conditions, occlusions, oblique viewing angles, a variety of facial expressions, partial faces of other people, face paint and disguise, and black and white images, as can be observed from the samples provided.

We artificially created height data for each user, since height was found to be the most important soft biometric in determining the identity in Chapter 4. To keep the data realistic and model the differences between the estimated heights, Gaussian noise with $\sigma = 6.3$ cm found in Chapter 4 was added to the heights of the users obtained from the web.

We assume that the average number of times a user will be observed is $\overline{n_o} \geqslant 10$, which is a reasonable assumption for long-term HRI. Hence, we created two datasets: (1) ten samples dataset (D-Ten), where each user is observed precisely ten times, e.g. ten return visits to a robot therapist, and (2) all samples dataset (D-All), in which each user is encountered a different amount of times (10 to 41 times). Two types of distribution are considered for the time of interaction: (1) patterned interaction times in a week modelled through a Gaussian mixture model (G), where the user will be encountered certain times on specific days, which applies to HRI in rehabilitation and education areas, and (2) random interaction times represented by a uniform distribution (U), such as in domestic applications with companion robots, where the user can be seen at any time of the day in the week, resulting in a total of four datasets (D-Ten$_{\text{Uniform}}$, D-Ten$_{\text{Gaussian}}$, D-All$_{\text{Uniform}}$, D-All$_{\text{Gaussian}}$). The clean datasets of images and the resulting datasets are available online[3].

[3]https://github.com/birfan/MultimodalRecognitionDataset

5.3 Hypotheses

H1 Our proposed Multi-modal Incremental Bayesian Network (MMIBN) will improve user recognition compared to face recognition (FR) alone, as measured by a decrease in the long-term recognition performance loss (L) and increase in the identification rate of known users (DIR).

H2 Online learning (MMIBN:OL) will improve user recognition over a non-adaptive model, as measured by a decrease in L and an increase in DIR.

H3 Hybrid normalisation will outperform the individual normalisation methods.

H4 When assumptions are made about the temporal interaction pattern of the user (i.e., for Gaussian patterned timing of interaction), recognition will improve. When the time of interaction is uniformly distributed, the loss L will be higher.

5.4 Experimental Procedure

5.4.1 Cross Validation

We evaluated the stability and performance of the models using repeated k-fold cross-validation, as described in Algorithm 1. Two different methods are used for creating validation folds, namely, **OrderedKFold** and **ShuffledKFold**. OrderedKFold is the case where users are introduced one by one to the system without any repetitions of previous users during the enrolment (step 4 in Algorithm 1). The order of repeated interactions is random after the enrolment. In ShuffledKFold, there can be repetitions of the previous user(s) before another user is introduced, because the order of overall samples is random (step 10). OrderedKFold is similar to batch learning in an incremental learning sense, whereas, the iteration (repeat) created by ShuffledKFold is more similar to a real-world scenario, as experienced in the user study in the previous chapter. Our aim is to evaluate if there are any performance differences between the two cases and to prove that the model is stable across several repeats. We chose K= 5 folds and R= 11 repeats.

Algorithm 1 Repeated K-Fold Cross-Validation Generation

```
1:  function OrderedKFold(K, U)    ▷ K is the number of folds, U is the samples for each user
2:      k ← 1
3:      while k <= K do                                   ▷ Create initial cross-validation set
4:          SU ← shuffle order of U               ▷ Enrolment order is different across each bin
5:          B[k] ← SU[i][j : j+length(SU[i])/K]      ▷ Divide user samples equally to each bin
6:          V[k] ← stratified randomise order B  ▷ Initial and final bins are different per fold
7:          k ← k + 1
8:      return V                                                           ▷ Validation set
9:  function ShuffledKFold(K, P)                      ▷ P is the (previous) validation set
10:     SP ← shuffle P        ▷ Shuffle the order of the user samples in previous validation set
11:     k ← 1
12:     while k <= K do                                   ▷ Create initial cross-validation set
13:         B[k] ← SP[j : j+length(SP)/K]      ▷ Divide shuffled validation set across each bin
14:         V[k] ← stratified randomise order B  ▷ Initial and final bins are different per fold
15:     return V                                                           ▷ Validation set
16: procedure RepeatedKFold(R, K, U)                        ▷ R is the number of repeats
17:     C[1] ← OrderedKFold(K, U)
18:     r ← 2
19:     while r <= R do                ▷ Create cross-validation set for number of repeats
20:         C[r] ← ShuffledKFold(K, C[r − 1])
21:         r ← r + 1
22:     return C                                        ▷ Repeated K-Fold Cross-Validation
```

Each dataset (D-Ten and D-All) is divided into two with 100 users each. The first set is then divided through cross-validation procedure with 80% of the data for the *training set* (first four bins, corresponding to 800 samples in D-Ten and 2308 in D-All) and 20% of the data for the test set, *closed-set (training)* (final bin, corresponding to 200 samples in D-Ten, 620 in D-All). The *open-set* is created from the remaining 100 users (800 samples in D-Ten, 2280 in D-All). The *closed-set (open)* is similar to *closed-set (training)*, which corresponds to the final bin in each fold (200 in D-Ten, 570 in D-All). The open-set evaluation is made by introducing the open-set samples after the training set, that is, 100 users are enrolled in the system, and recognised multiple times before the introduction of 100 new users. However, the results for the open-set do not include the results for training.

A stratified random bin order (step 6 and 14) is used for having a different initial bin and final bin in each fold to ensure a different enrolment order of users and a different test set, respectively.

The only difference between Gaussian and uniform datasets is the time of the interaction for each sample; that is, the order of the samples is the same.

For online learning, the likelihoods are learned during the training phase (*training* and *open-set*), and the learned likelihoods are used without online learning for the *closed-set*.

5.4.2 Variables and Measures of the Study

Given our datasets and the parameters of our model, we have four independent variables and three dependent variables for analysing the results on the evaluation sets: training, open-set, closed-set (training), closed-set (open). The dependent variables are DIR in Equation 3.13, FAR in Equation 3.14 and long-term recognition performance loss (shortly, loss) in Equation 3.15, which are used to measure the performance of the models. The independent variables are as follows:

1. **Dataset size**: ten samples per user (D-Ten), random amount of samples (D-All)
2. **Timing of interaction**: patterned interaction times (gaussian), random interaction times (uniform)
3. **Model**: non-adaptive (fixed) likelihoods (MMIBN), online learning (MMIBN:OL)
4. **Normalisation method**: softmax, minmax, tanh, normsum and hybrid

5.5 Results

5.5.1 Optimisation of Parameters

In this section, we present our empirical evaluations to obtain the optimised parameters of our system on the described datasets. Bayesian optimisation[4] is used to optimise the weights of the network and the threshold for the quality of the estimation (Q) (θ_Q). A total of 303 iterations is used for 5-fold cross-validation for each combination of the independent variables (for 40 conditions). The parameters are optimised by minimising the loss on the training set. By using the optimised parameters, 11 repeats of 5-fold cross-validation are conducted for each of the conditions to evaluate the effects of the independent variables on the open-set.

The loss parameter α defined in Equation 3.15 should be set to find the optimum face recognition threshold (θ_{FR}) and optimise the parameters in our network. As α increases, the fraction of correct recognitions of enrolled users (DIR) increases, but the fraction of the incorrect recognitions of unknown users (FAR) will increase. Based on our average

[4]`https://thuijskens.github.io/2016/12/29/bayesian-optimisation/`

number of observations assumption $\overline{n_o} = 10$ for long-term interaction, α becomes 0.9. For applications with fewer observations per user, α can be set accordingly.

5.5.1.1 Face Recognition Threshold

In face recognition, if the highest similarity score is below θ_{FR}, the identity is classified as unknown. We examined how θ_{FR} influences the long-term recognition performance loss for the NAOqi FR, and noticed a decrease in performance (i.e. increase in loss) for $\theta_{FR} > 0.4$, as shown in Figure 5.2. Hence, we chose $\theta_{FR} = 0.4$ because it is the highest threshold giving the lowest loss to decrease FAR in our model, in agreement with our previous work in Chapter 4.

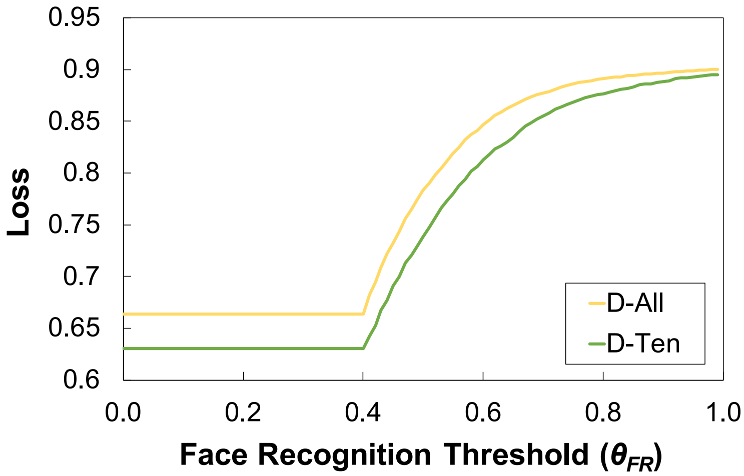


Figure 5.2: The change in long-term recognition performance loss for NAOqi face recognition (FR) given the FR threshold (θ_{FR}), in D-Ten (ten samples dataset) and D-All (all samples) datasets. The results show that 0.4 is the ideal θ_{FR}, because it is the highest threshold giving the lowest loss to decrease FAR.

5.5.1.2 Analysis of Variance of Independent Variables

For clarity of the presentation of results, we will initially analyse the results for 11-repeats of 5-fold cross-validation, before presenting the optimised parameters from Bayesian optimisation. This would allow us to later analyse only the optimisation parameters for the best performing normalisation method.

Levene's test on the loss reveals ($F(10, 2189) = 0.026, p = 1$) that there is no significant difference in variances between the repeats, which, indicates that our Bayesian network models are stable across repeats. Analysis of variance (ANOVA) (Type-I) supports that there is no significant difference between repeats ($F(10, 2189) = 0.044, p = 1$), which shows that there is no significant difference between the OrderedKFold cross-validation and the ShuffledKFold, indicating that the model performs equally well for learning new users incrementally sequentially (similar to batch learning) and at random intervals (similar to a real-world scenario). Hence, we will only analyse the results of a single randomly selected repeat of 5-fold cross-validation. Since the model is stable across repeats, using a single repeat of cross-fold validation instead of independent test sets does not violate ANOVA assumption (Beleites & Salzer, 2008).

Due to the linear relation of loss with DIR and FAR in Equation 3.15, there will be a correlation between the parameters. Pearson's product-moment partial correlation coefficient was computed to assess their relationships. The results show that there is a negative correlation between loss and FAR ($r(200) = -0.18, p = .009$) and a positive correlation between loss and DIR ($r(200) = 0.99, p = 9.9 \times 10^{-195}$). However, no significant correlation is found between FAR and DIR ($r(200) = 0.08, p = .25$). Hence, we will only report the significance analysis of loss.

A factorial ANOVA is conducted for analysing the primary and interaction effects of our independent variables. The results show that there are no significant primary effects for the model ($F(1, 160) = 1.50, p = .22$), and no significant interaction effects are found between the dataset size, timing of interaction, and model combination ($F(1, 160) = 0.01, p = .91$). Every other independent variable and their interactions are found to be significant ($p < 0.001$ level). This shows that the size of the dataset, timing of interaction and normalisation method have significant effects on the performance of the model, but online learning by itself does not provide significant improvement.

There are 200 data points (40 conditions across 5-folds) evaluated for ANOVA, so central limit theorem takes effect for the normality assumption. Inspection of the Q-Q plot shows that most of the data points lie on a linear axis. Residual plots do not show a particular pattern, thus the equality of variances can be inferred. The Cook's distance for high leverage points is less than 0.25, so there is no need to remove them as outliers. Moreover, the dataset is balanced, that is, sample sizes are equal for each condition. Thus, ANOVA results are valid.

5.5.1.3 Normalisation Methods

We conducted a post-hoc analysis using Tukey's Honestly Significant Differences (HSD) test on the cross-validations. Throughout this thesis, we adopt a letter-representation for Tukey's HSD test plots. Levels that are not significantly different from each other at 0.95 confidence level ($p < 0.05$) are represented with the same letter over all the conditions, that is, each method is compared to all the other methods in different conditions. In other words, if two methods do not share a common letter, then there is a significant difference in performance between them. Multiple letters mean that the method is at the same significance level with multiple other methods. Additionally, for clarity of results, D-All and D-Ten datasets have been analysed separately, however, the results show similar patterns in both datasets.

Figure 5.3 shows the resulting boxplot for D-All from Tukey's HSD separated by the learning method and timing of interaction. Tukey's test results for D-Ten show a similar pattern, presented in Appendix B.1.

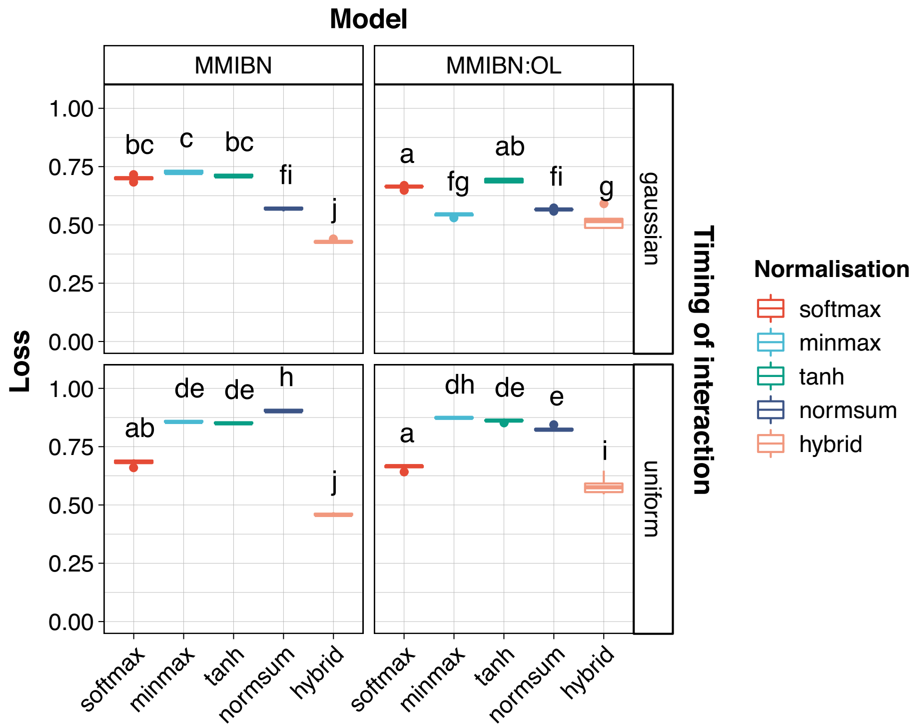


Figure 5.3: Results of Tukey's HSD test of loss in the open-set for normalisation methods with optimised weights for all samples (D-All) dataset: softmax, minmax, tanh, normsum, and hybrid. Lower loss is better. Levels that are not significantly different from each other at 0.95 confidence level are represented with the same letter within all separations. The results show that hybrid normalisation significantly outperforms the other normalisation methods in all datasets and models.

In both of D-All and D-Ten datasets, hybrid normalisation provides significantly lower loss ($p < 0.05$) in all conditions except for online learning in Gaussian timing for D-All ($p = 0.78$ in D-All$_{\text{Gaussian}}$), in which case it still provides the lowest mean for loss. Hence, our hypothesis **H3** is strongly supported, and hybrid normalisation method is chosen for the remaining analyses.

Even though we did not find any significant differences in the primary effect of the learning method, Figure 5.3 shows that there are significant differences between online learning and fixed likelihoods for hybrid normalisation. Online learning results in a higher loss for both datasets, which is in contrast with our hypothesis **H2**. The other methods do not show a stable pattern across conditions or datasets.

Most methods perform significantly worse in uniform timing of interaction (random interaction times), as compared to patterned interactions (Gaussian times), supporting our hypothesis **H4**. Softmax performs equally well on both models for D-All but performs worse in uniform timing for D-Ten. Hybrid normalisation performs equally well for MMIBN in D-All but performs significantly worse in other conditions.

Hybrid normalisation performs better in all conditions and shows stability across varying conditions compared to the other methods. It achieves lower loss in D-All than in D-Ten, as a result of a higher number of samples in D-All (2280 in open-set) as compared to D-Ten (800 samples), which shows that the proposed model gets better with the increasing number of recognitions.

5.5.1.4 Weights and Quality of the Estimation

It seems to be self-evident that in the case of the uniformly distributed time of interaction, online learning would provide worse results because the information provided by time will be unreliable. Hence, the optimisation should find a lower weight for the time parameter. The parameters corresponding to the optimum loss in Figure 5.4 show otherwise. Weight for the uniform time is higher than that of the Gaussian for online learning in both datasets.

In general, based on the relatively high weights, age seems to be the most important parameter, and height the least. This is in contrast with the findings in Chapter 4.

The optimised threshold for the quality of the estimation (θ_Q) was found to be less than 0.1 in each condition. The underlying reason is the disagreement of the modalities, which can decrease the differences in posterior probabilities because the results are combined through the product rule in the BN. When the modalities agree with high confidences (probabilities), the quality can be very high, such as $Q = 7.44$, as shown in Figure 5.12 in Section 5.5.2 for the probe of the second user.

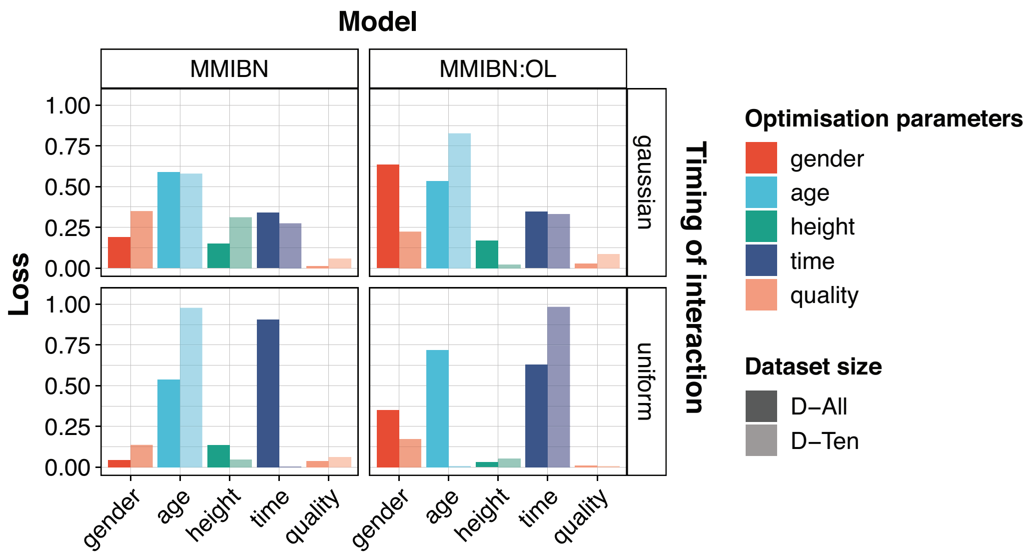


Figure 5.4: Values of the optimised parameters through Bayesian optimisation of 303 iterations over 5-fold cross-validation for hybrid normalisation for all samples (D-All) and ten samples (D-Ten) training sets: the weights for gender, age, height, and time of interaction and the threshold of the quality of the estimation (θ_Q). In general, age seems to be the most important parameter, and height the least.

5.5.2 Comparison to Baselines

On the grounds that the optimised parameters of our proposed MMIBN are found, we can compare its results to FR and SB. FR results are obtained from the NAOqi estimations by setting FR threshold (θ_{FR}) to 0.4, as found in Section 5.5.1.1. SB results are obtained by giving zero weight to FR, that is, only gender, age, height estimates from NAOqi and time of interaction are used for identifying a user. The weights of these modalities in SB are the same as MMIBN, as shown in Figure 5.4. Similarly, the weights of SB:OL are the same as those of MMIBN:OL.

We transformed a state-of-the-art open world recognition method, Extreme Value Machine (Rudd et al., 2018) (EVM), as described in Appendix B.3, to accept sequential and incremental data for online learning by adjusting its hyperparameters to use it as a baseline. In the original work, batch learning of 50 classes was used with an average of 63806 data points at each update, instead of a single data point that we used in this work. We compared our methods with the performance of two EVM models: (a) EVM:FR, using NAOqi face recognition similarity scores as data, (b) Extreme Value Machine trained with multi-modal data (EVM:MM) in the same format as it is used for our methods.

Section 5.5.2.1 compares the long-term recognition performance loss (shortly, loss) between the models. There is a significant correlation between FAR, DIR and loss, as found in Section 5.5.1.2, which can be observed from the various results during the Bayesian optimisation of weights and quality of the estimation shown in Figure 5.5. Hence, the analysis of loss is sufficient to determine how the model performs in comparison to others. Nevertheless, we will report the results of FAR and DIR of the models in Section 5.5.2.2 to further observe how the open-set recognition metrics are affected.

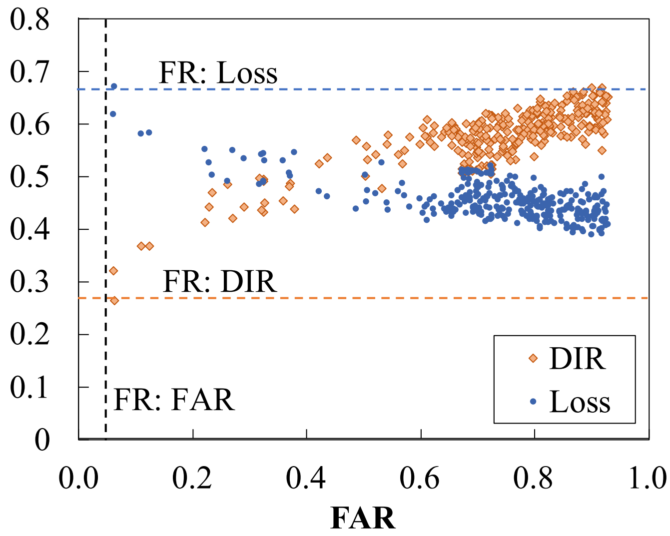


Figure 5.5: ROC curve with long-term recognition performance loss and DIR for varying FAR, based on the Bayesian optimisation of MMIBN with hybrid normalisation for varying weights and quality of the estimation for 303 iterations over 5-fold cross-validation, in the all samples dataset with Gaussian times (D-All$_{\text{Gaussian}}$). FR values are given for comparison. As DIR increases, loss decreases but FAR increases. The loss parameter (α) can be adjusted, or a FAR can be set to obtain a different set of weights.

5.5.2.1 Long-term Recognition Performance Loss

As previously mentioned, the proposed models perform better in terms of loss in D-All than in D-Ten, however, the results for D-Ten datasets show similar patterns to that of D-All. Taken the same number of recognitions for both D-All and D-Ten, that is equal to the number of samples in D-Ten for all evaluation sets, ANOVA shows that there is no significant difference in the sample size ($F(1, 192) = 0.179, p = .67$) as the models perform equally well for D-All and D-Ten for the same number of samples. In other words, it does not matter if each user is observed the same number of times or not. This also supports that a higher number of samples increases the performance of the models. Hence, the following analysis will only be focused on D-All, but any differences in performance between the two datasets will be noted wherever necessary.

Figure 5.6 presents Tukey's HSD test results on the training, open-set, closed-set (training), closed-set (open) for D-All datasets with Gaussian and uniform timing of interaction.

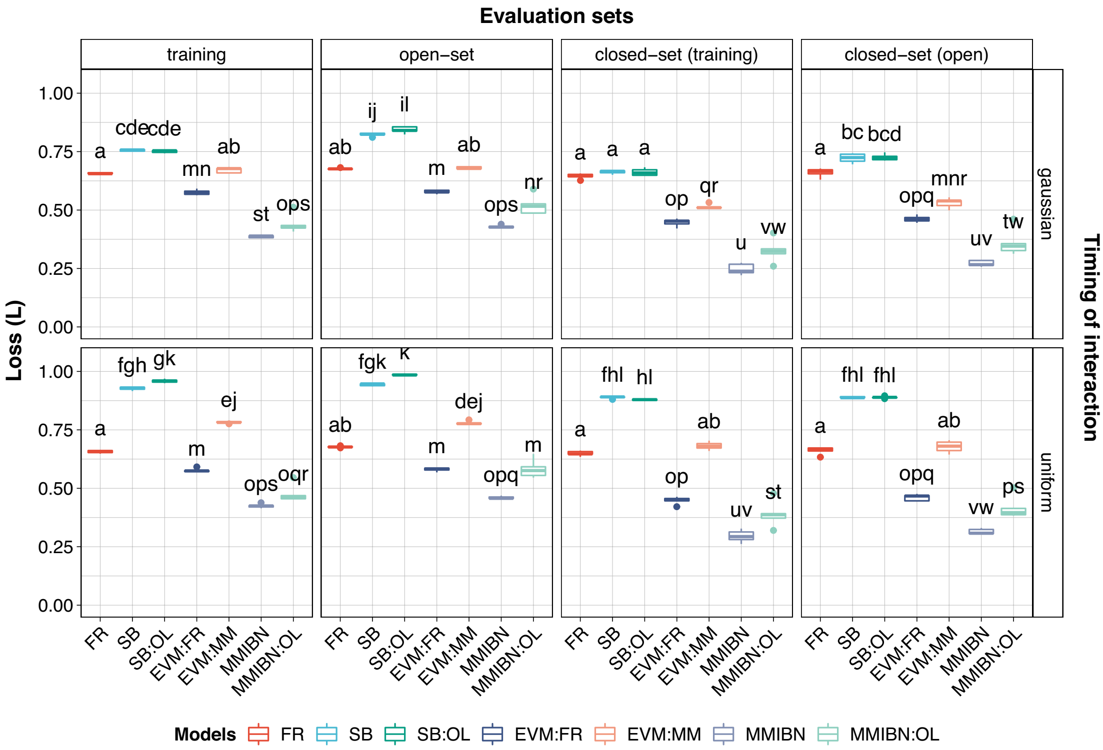


Figure 5.6: Comparison of Tukey's HSD test results on loss for the proposed Multi-modal Incremental Bayesian Network (MMIBN), face recognition (FR), soft biometrics (SB) with online learning condition (:OL), Extreme Value Machine trained with face recognition data (EVM:FR) and with multi-modal data (EVM:MM). The results are presented for training (100 users), open-set test (200 users), closed-set (training) (100 users) and closed-set (open) (200 users) for all samples dataset (D-All) for Gaussian and uniform timing of interaction. Lower loss is better. Levels that are not significantly different from each other at 0.95 confidence level are represented with the same letter within all separations. The results show that our proposed approaches significantly outperform FR, SB and EVM in all sets. The non-adaptive MMIBN performs slightly better than online learning (MMIBN:OL). MMIBN performs equally well between Gaussian and uniform timing, between training and open-set cases, and closed-sets, which shows that the model is not significantly affected by the timing and scales well for an increase in users.

The results show that the proposed approaches (MMIBN and MMIBN:OL) decrease the long-term recognition performance loss significantly ($p = 0$) and substantially compared to FR, supporting the first part of our hypothesis **H1**. This finding is valid across all datasets (D-Ten and D-All for Gaussian and uniform times).

MMIBN performs equally well between Gaussian and uniform timing for D-All evaluation sets (i.e. no significant difference, but slightly worse in uniform), whereas, it does not perform at the same significance in D-Ten evaluation sets (performs significantly worse).

MMIBN:OL performance changes depending on the dataset size and the evaluation set (performs equally well only in closed-sets in D-Ten, and for training and closed-set open in D-All). Nevertheless, the models have slightly or significantly higher loss in uniform timing as compared to Gaussian.

Online learning does not perform better than MMIBN because it increases the loss at all conditions. In fact, except for training set in D-All and D-Ten and closed-sets in D-Ten for uniform timing where MMIBN and MMIBN:OL perform at the same significance level, online learning is significantly worse, which is in contrast with our hypothesis **H2**.

Furthermore, the results show that soft biometric features (SB and SB:OL) are not able to identify a user on their own. In general, they perform significantly worse than FR. However, when the interaction is time patterned (Gaussian), SB performs better and closer to FR as compared to uniform timing. Especially for closed-set training in D-All, it is remarkable that SB features identify the user with the same significance level performance as FR. SB and SB:OL perform mostly equally well in D-All datasets, but SB:OL performs significantly worse in several evaluation sets in D-Ten.

EVM:FR performs significantly better ($p < .005$) than FR across all conditions. EVM:MM is significantly worse than EVM:FR ($p < .01$), and it does not perform better than FR in most conditions. This shows that although EVM is a good method for clustering face recognition data, but it does not perform well with multi-modal data.

MMIBN significantly outperforms ($p < 2 \times 10^{-10}$) both EVM models across all conditions in both D-All and D-Ten. This proves that our proposed approach is significantly better than the state-of-the-art method for incremental open world recognition with multi-modal biometric information. However, EVM models use online learning instead of fixed learning rates, which could potentially lead to worse performance as observed for our model. Nevertheless, comparing EVM models to MMIBN:OL shows that MMIBN:OL significantly outperforms EVM models ($p < .05$ to $p = .0$) in most cases, except for uniform timing for open-set and closed-set (open) in D-All and open-set in D-Ten, in which, it performs equally well with EVM:FR.

MMIBN performs equally well between training and open-set cases as well as between closed-sets, which shows that the model scales well for an increase in users (from 100 to 200 users), suggesting that the proposed approach and the optimised weights can generalise. Similar to the results in (Rudd et al., 2018), EVM performs equally well between those sets, showing that the change in the model from batch updates to incremental updates have not changed its structure for scaling well. The models perform significantly better in closed-sets as compared to training or open-set due to the lack of unknown users in closed-sets where loss only depends on DIR (FAR= 0.0).

The model performance improves with the increasing number of recognitions and stabilises towards the end (around 2000), as can be observed in Figure 5.7. This supports our initial finding of performance difference between D-All and D-Ten, given that they perform equally well for the same number of recognitions. Initially, loss increases with increasing FAR, when the users are introduced to the system (represented by dots in the plot). As the number of recognitions increases, the introduction of a new user does not notably increase the loss as can be observed by the final three new users in the training set. Even though MMIBN models get better over time, they start performing consistently better than both FR and EVM models throughout both training and closed-set after only a small number of recognitions (15 to 48 in training, 1 to 6 in closed-set).

The sudden change at the beginning for the training set is due to the sequential calculation of loss for time plots: a previously enrolled person, which has not been identified correctly for the first time, changes DIR from 1.0 to 0.5 (one out of two enrolled users was incorrectly identified). Note that the introduction of new users is at random order due to ShuffledKFold function described in Section 5.4.1. The results for the open-set, as given in Appendix B.2, show a similar pattern of loss between open-set and closed-set (of the open-set cross-validation).

5.5.2.2 Open-Set Identification Metrics: DIR and FAR

The previously presented results confirm our claims that our proposed multi-modal Bayesian networks perform significantly better than FR, SB and EVM in long-term interactions, because of the significantly high correlation between loss with DIR and FAR. However, analysing the open-set identification metrics allows us to understand how the models perform for enrolled and unknown users through DIR and FAR, respectively.

Tukey's HSD test results for DIR are presented in Figure 5.8. This plot resembles highly that of Figure 5.6 in a reversed direction, because of α $(1 - \mathrm{DIR})$ component of loss, whereby, $\alpha = 0.9$. The increase in DIR is significant ($p = 6.1 \times 10^{-15}$) and drastic, from 0.268 of FR to 0.657 with MMIBN and 0.561 with MMIBN:OL averaging over all the conditions in D-All

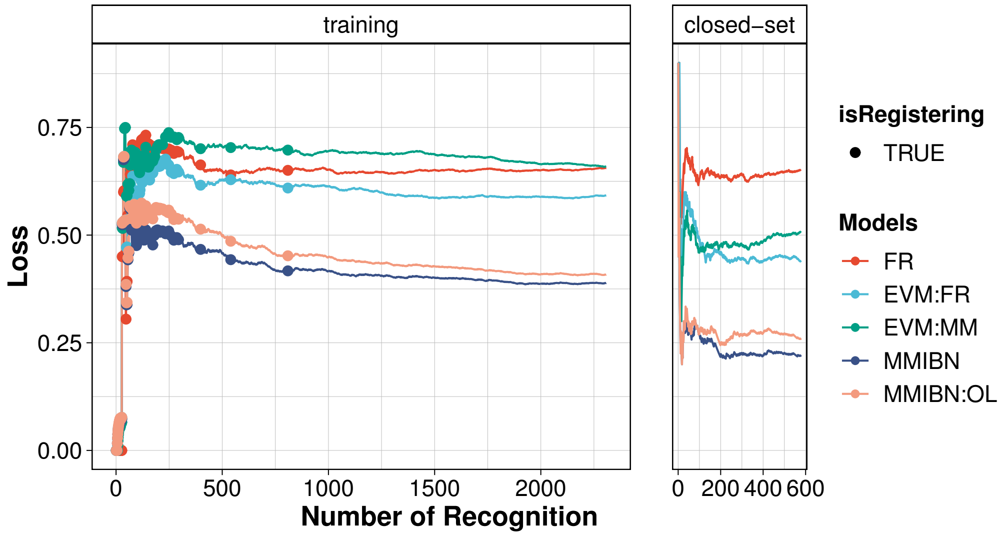


Figure 5.7: The change of loss with the increasing number of recognitions for all samples dataset with Gaussian times (D-All$_{\text{Gaussian}}$) for training and closed-set (training). The loss decreases with the increasing number of recognitions.

(timing of interaction and evaluation set). That is 38.9% increase in identifying the users correctly by using MMIBN, no matter the condition, which is more than double what FR is capable of providing. Hence, our hypothesis **H1** that the loss will be reduced and DIR will be increased using our proposed models as compared to FR alone is fully and strongly supported.

It should be noted that the increase in DIR provided by our network is significantly higher ($p < 2 \times 10^{-16}$) than DIR of soft biometrics (0.226 on average for Gaussian timing in D-All). This shows that soft biometric data are not sufficient to identify an individual, yet when combined with the primary biometric, they improve the identification rate significantly (38.9% in D-All, and 31.8% in D-Ten). This conclusion is supported by the datasets where the time of interaction is uniformly distributed (DIR of SB is 0.013 in average), that is, due to the high variability of time, the identification rate of SB is close to zero. Nevertheless, MMIBN performs equally well in Gaussian, and uniform timing within all evaluation sets in D-All, and MMIBN:OL performs equally well in D-Ten. As previously noted in **H4**, the loss is (slightly or significantly) higher, and DIR is (slightly or significantly) lower for all datasets and MMIBN models between Gaussian and uniform timing.

MMIBN significantly outperforms both EVM methods in DIR in all datasets ($p = 0.0$). EVM:FR has significantly higher DIR than FR and EVM:MM ($p < 1 \times 10^{-9}$). EVM:FR performs equally well between uniform and Gaussian timing in all datasets because it is trained only on FR data. DIR of EVM:MM drops below that of FR for uniform timing

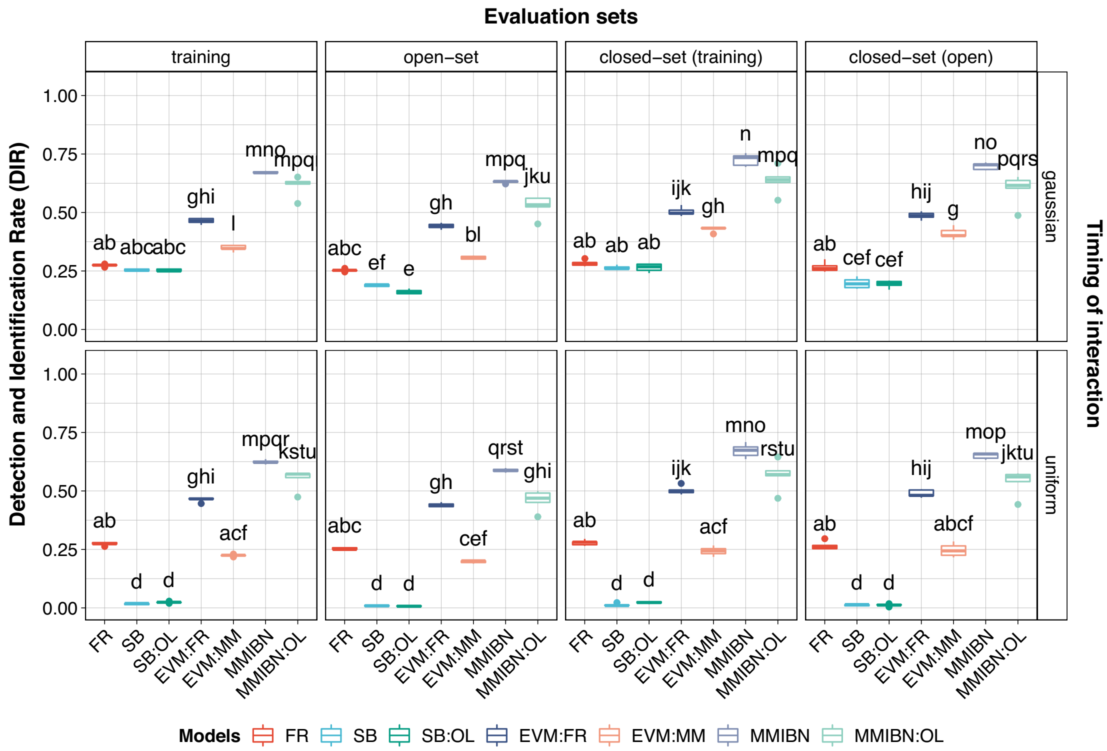


Figure 5.8: Tukey's HSD test results for Detection and Identification Rate (DIR) of all models for D-All datasets. Higher DIR is better. MMIBN outperforms all other methods in all sets and timing.

for both D-All and D-Ten, which shows that EVM is not a model to be used with time information, since the pattern of interaction with the user might not be known beforehand. Similarly, MMIBN:OL provides worse performance for uniform timing in D-All, but it always performs significantly better than or equally well with EVM:FR.

FR performs similarly in open and closed-sets in terms of loss, because it has significantly low FAR compared to MMIBN models, as can be observed in Figure 5.9. Even though low FAR is a desirable feature, the underlying reason of low FAR is that FR has very poor recognition performance on larger datasets and fails to recognise the users, because the highest similarity score returned by the identifier is lower than the threshold ($\theta_{FR} = 0.4$). However, as described in Section 5.5.1.1, this threshold ensures the lowest loss for FR.

FAR of the proposed models is high because of the combination of all modalities, which increase the probability of mixing the unknown user with an enrolled user. Possible solutions to this problem will be proposed in Section 5.6. For our proposed models, FAR in the training set is generally slightly less than that of open-set, because of the higher number of users enrolled, but there are no significant differences across the datasets for MMIBN, supporting that the model scales well to a larger dataset without a significant decrease in performance.

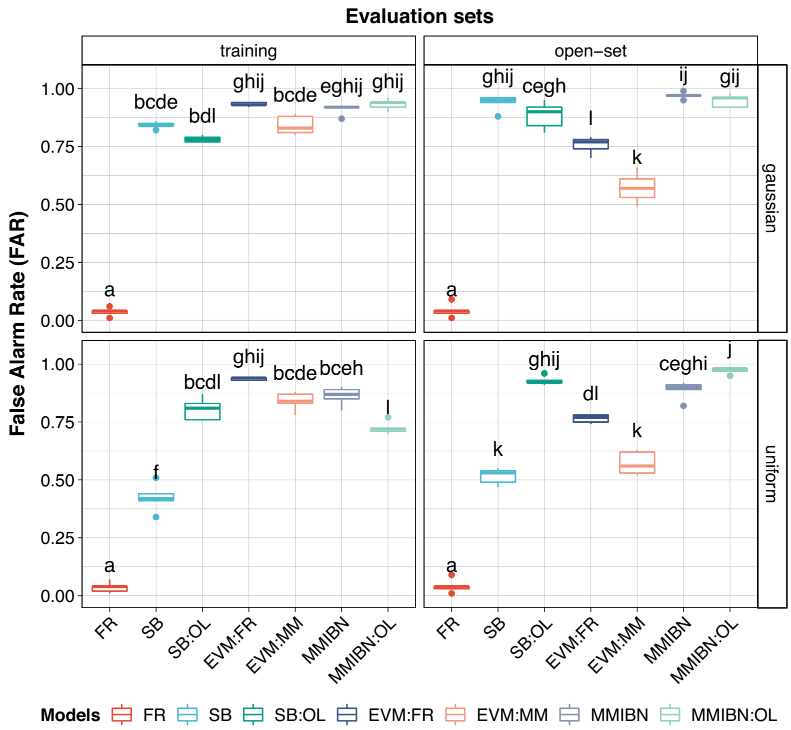


Figure 5.9: Tukey's HSD test results for False Alarm Rate (FAR) of all models for D-All datasets. Lower FAR is better. FR outperforms other methods in FAR, mainly due to identifying most users as unknown.

In the training set, there is no significant difference between MMIBN and EVM models, and MMIBN:OL performs significantly better than EVM models for uniform timing. In contrast to MMIBN, EVM provides significantly lower FAR in open-sets than in training sets. The authors state in (Rudd et al., 2018) that this is due to its ability to tightly bound class hypotheses by their support.

5.5.2.3 User-Specific Analysis

Confusion matrices presented in Figure 5.10 show how users were identified throughout the training set in D-All for a fold of the cross-validation, with 0 as the ID of the unknown user and the remaining numbers corresponding to IDs of the users. The heat map represents the percentage of identification of the user as the estimated user. Ideally, the diagonal should be all dark red if users are correctly identified. However, FR (item a) mostly identifies the users as unknown, resulting in the corresponding vertical axis of 0 to be mostly red and in a low FAR and a low DIR. MMIBN (item b) has mostly red coloured dots on the diagonal but has mixed users with other enrolled users as can be seen from light blue dots all over the matrix. MMIBN:OL shows a similar pattern with slight deviations.

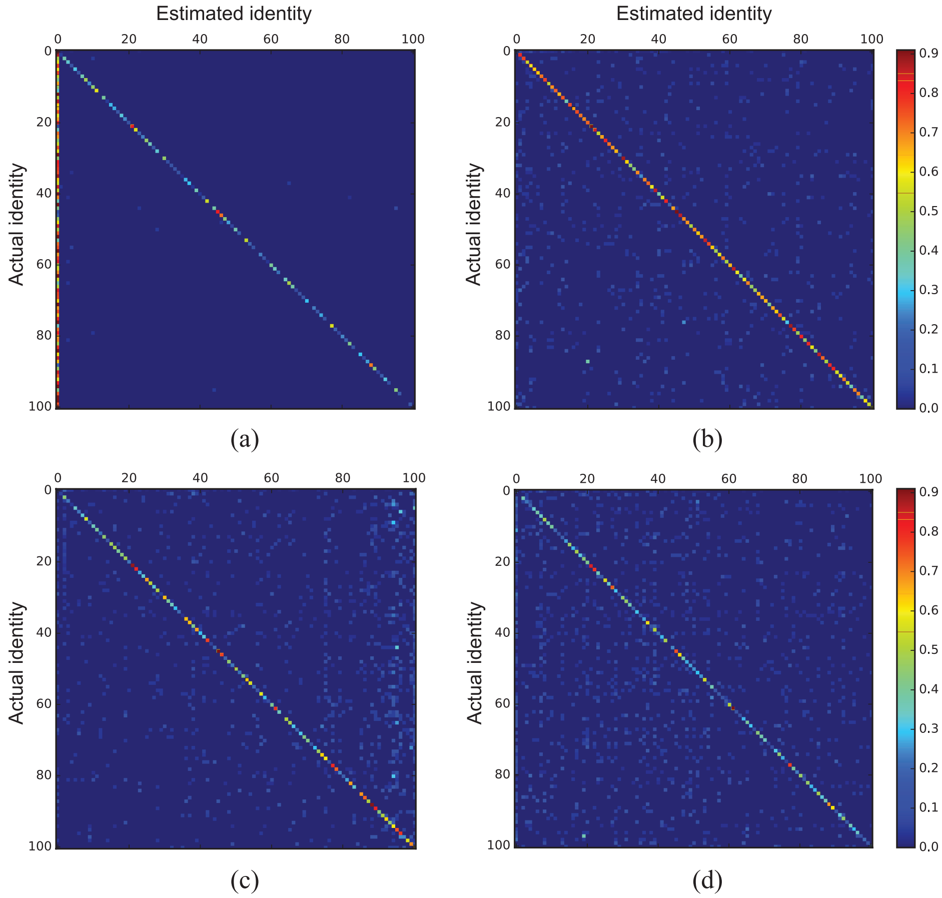


Figure 5.10: Confusion matrices of user identification for second fold of cross-validation on D-All$_{\text{Gaussian}}$: **(a)** face recognition (FR), **(b)** proposed model (MMIBN), **(c)** incremental Extreme Value Machine trained with face recognition data (EVM:FR), **(d)** incremental Extreme Value Machine trained with multi-modal data (EVM:MM). The heat map represents the percentage of identification as the estimated user. Ideally, if all users are correctly identified, the diagonal should be dark red, and the remaining of the matrix should be dark blue.

Even though EVM:FR (item c) only uses FR information, its confusion matrix is different from that of FR. The misidentifications are highly concentrated on the final ten users, suggesting that either FR or EVM might be subject to the catastrophic forgetting problem. Using multi-modal data overcomes that problem, as can be seen for EVM:MM (item d) as misclassifications are evenly distributed, similar to MMIBN. However, the diagonals in EVM models have notably fewer reds than MMIBN.

The significant differences of identification of users over the 5-folds of cross-validation, as revealed by Figure 5.11, shows another striking result. FR (item a in the figure) does not perform equally well amongst the users in that there are significant differences of identification (represented by darker blue colours). Our proposed approach MMIBN (item b) balances the performance amongst users, thereby, reducing any biases in the system while improving the performance of the overall system significantly as compared to FR.

Online learning (MMIBN:OL in item c and EVM:FR[5] in item d) balances the performance further, in contrast to the decrease in performance compared to MMIBN.

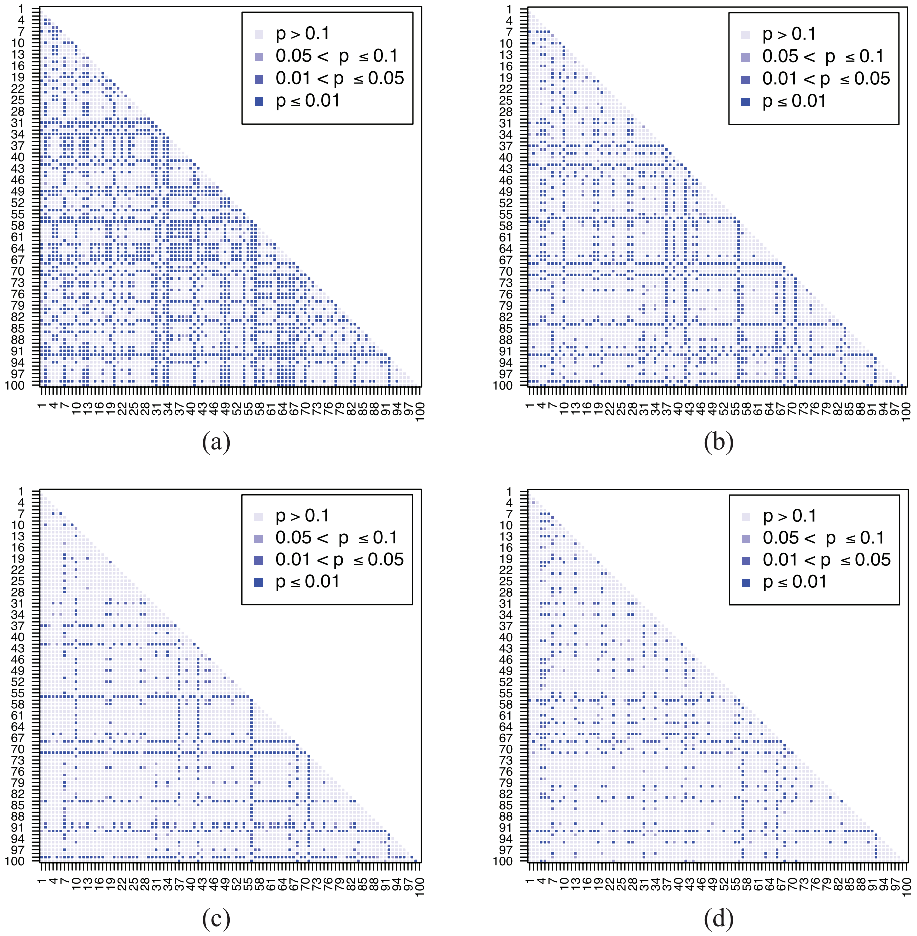


Figure 5.11: Tukey's HSD test results for significant differences of user-based identification over 5-fold cross-validation on D-All$_{\text{Gaussian}}$: **(a)** face recognition (FR), **(b)** proposed model (MMIBN), **(c)** proposed model with online learning (MMIBN:OL), **(d)** Extreme Value Machine trained with face recognition data. The darker blue colours represent significant differences, whereas lighter blue colours means that the users are identified equally well. The results show that our approach and online learning decrease the bias and balance the performance between users.

Figure 5.12 demonstrates examples from D-All$_{\text{Gaussian}}$ where face recognition fails to recognise the user due to the low similarity score ($< \theta_{FR} = 0.4$), whereas, our proposed model identifies the user correctly based on soft biometric information. The quality of the estimation (Q) varies depending on the highest FR similarity score, as well as the disagreement between modalities. For example, for the third user (Sandra Oh), the highest FR similarity score (rank 1) is very low, corresponding to David Schwimmer who is 28 years old in the dataset, has a height of 185 with the enrolment time of interaction on Tuesday at 18:16. Age did not provide information to differentiate the user from the incorrect estimation, whereas, height and time of interaction increased the probability

[5] EVM:MM shows a similar pattern.

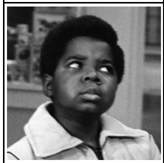
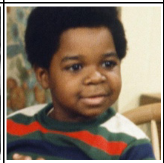
	Enrolment	*Probe: 11*		*Enrolment*	*Probe: 2*		*Enrolment*	*Probe: 8*	
									
	True Value	**Estimated Value**		**True Value**	**Estimated Value**		**True Value**	**Estimated Value**	
		FR	*BN*		*FR*	*BN*		*FR*	*BN*
ID	135	0 **[0.70]**	**135 [0.70]**	129	0 **[0.79]**	**129 [7.44]**	77	0 **[0.83]**	**77 [0.35]**
Name	Emilia Clarke	*FR (rank 1)*: Angelina Jolie **[23.3%]**		Gary Coleman	*FR (rank 1)*: Gary Coleman **[36.9%]**		Sandra Oh	*FR (rank 1)*: David Schwimmer **[13.9%]**	
Gender	Female	Female **[72.7%]**		Male	Female **[88.1%]**		Female	Male **[66.3%]**	
Age	24	38 **[50%]**		10	7 **[100%]**		33	28 **[40%]**	
Height	157	152 **[8%]**		142	154.5 **[8%]**		168	172.7 **[8%]**	
Time	Saturday 10:15	Wednesday 17:35		Wednesday 13:41	Wednesday 13:53		Thursday 08:14	Thursday 07:57	

Figure 5.12: Examples of true values and estimated values of modalities from our Multi-modal Long-Term User Recognition Dataset with Gaussian times (confidence values are given in brackets) using proposed multi-modal Bayesian network with hybrid normalisation (referred to as BN in the figure). Highlights in red show the incorrect detection values. Face recognition was unable to recognise the users (0 represents unknown user) because the similarity scores were below the threshold (40%). Our proposed BN was successful (highlighted in green) in correctly identifying the users with varying quality of estimations (shown in brackets underneath the ID) as a result of the information gathered from soft biometrics highlighted in blue. 8% confidence value of height corresponds to the $\sigma = 6.3$ cm in NAOqi. Images are taken from IMDB-WIKI dataset (Rothe et al., 2015, 2018).

that the user is Sandra Oh, resulting in a correct estimation, but with a low quality score ($0.35 > \theta_Q = 0.013$). The second user (Gary Coleman) was identified correctly by FR with the highest similarity score close to, but slightly lower than θ_{FR}. This was enforced by the age estimation, and the time of interaction, which compensated for the incorrect recognitions of gender and height, to get a high quality score (7.44).

5.6 Discussion

Our findings showed that from our initial hypotheses **H1** and **H3** are fully supported, **H4** is supported for hybrid normalisation, and **H2** is rejected. In this section, we will discuss the implications of our results, validate our initial assumptions, and offer other approaches for our models.

5.6.1 Baselines and Time Complexity

Our proposed multi-modal models perform significantly better than NAOqi face recognition and soft biometrics and Extreme Value Machine (EVM) models, which is a state-of-the-art open world recognition method. We have not used a state-of-the-art deep learning face recognition algorithm (such as Dlib (King, 2009)) for comparison, because those methods are not optimised for low computational power systems, hence, they may require a vast amount of time for encoding images, recognition and re-training[6], which makes them unsuitable for real-time open world user recognition on a robot. The proposed models can run on a commercial robot with low-computational power (on a single CPU of Pepper robot), and only require a small amount of time for execution. After the recognition user study presented in Chapter 4, we optimised the code for the Recognition Architecture and the MMIBN for decreasing the time it takes for calculating the estimated identity and saving and loading the model. Consequently, in addition to the time required from FR and other modalities ($M = 0.14$ s, $SD = 0.001$), MMIBN models take 0.01 second for recognition (compared to 0.3 second in user recognition study), significantly outperforming ($p < 1 \times 10^{-11}$) both EVM:FR and EVM:MM, which require 0.32 and 0.34, respectively[7]. For enrolling new users, MMIBN requires a significantly lower amount of time equal to 0.39 s ($p = 0.002$) for scaling the Bayesian network compared to all models. MMIBN:OL takes 0.54 s, for which 0.17 s is due to online learning. There is no significant difference between MMIBN:OL and EVM models for enrolling, EVM:FR takes 0.48 and EVM:MM takes 0.52 s, with 0.20 and 0.23 s for online learning, respectively. The higher amount of time required for EVM:MM compared to EVM:FR shows that online learning takes longer time when there is more information to be learned per user.

Moreover, in contrast to deep learning approaches, which require "big data" to be pre-trained, our proposed models are able to start from a state of no enrolled users, learn users continuously and incrementally, and improve performance compared to FR after a small number of recognitions (e.g. 48 for Figure 5.7).

[6]An implementation of Dlib for open world recognition using re-training on a dataset with a small number of users is explained in this link, which shows that the recognition can take 6-7 seconds on a single CPU system for a small number of users: https://www.pyimagesearch.com/2018/06/18/face-recognition-with-opencv-python-and-deep-learning/

[7]The results are given for D-All with Gaussian timing on the open-set.

5.6.2 Dataset Size

In general, FAR and DIR are higher, and loss is lower in D-All than in D-Ten. The increase in DIR and the decrease in loss, is explained by the higher number of recognitions, which increases the performance over time. The increase in FAR can be due to different optimised weights for each dataset (see Figure 5.4). However, both datasets show similar patterns in differences between FR, SB and MMIBN models. Even though the number of samples per each user is not the same in D-All, the fact that it performs equally well as D-Ten for the same number of recognitions shows that our equal priors ($P(I)$) assumption, which states that each user is equally likely to be seen, does not have any adverse effect on our proposed models. We suggest using the optimisation parameters (weights and quality threshold) that are optimised for D-All datasets since they are optimised using more samples on this dataset. If the application is based on users to come at specified times during a week (e.g. in a hospital), the optimised parameters for D-All$_{\text{Gaussian}}$ should be used; otherwise, it is better to use that of D-All$_{\text{Uniform}}$ (e.g. for companion robots). These optimised parameters generally perform significantly equivalent in both timing conditions in D-All for both models, as shown in Figure 5.6, even though the timing of interaction does not provide enough information in the uniform timing case.

5.6.3 Effects of the Loss Parameter and the Frequency of Appearance

High FAR of the models is due to the trade-off between identifying known users and spotting unknown people, which is visible in Figure 5.5. The value of α determines the importance of this trade-off in the loss function to ensure a higher number of correct recognitions in a long-term interaction. We found $\alpha = 0.9$ based on our assumption that the average number of interactions is 10. Nonetheless, using a varying amount of samples (D-All) did not change the overall performance in terms of the long-term recognition loss, when we compared D-All and D-Ten at the same number of total samples (800 for training and open-set and 200 for closed-sets). In Figure 5.7, 71% of the users had less than 10 recognitions, and 20% had more than 10, before 800th recognition in D-All dataset. This finding shows that our choice of α did not negatively affect the results.

5.6.4 Multi-modal Approach

As previously noted, the increase in DIR for MMIBN models (0.328 to 0.479) is higher than DIR provided by SB models (0.007 to 0.277), especially for uniform timing where SB provides close to zero DIR. This fact along with the non-zero optimised weights support that the inclusion of age, gender and height modalities increases the identification rates, suggesting that the visual modalities contain additional information to FR, and confirming our initial assumption of conditional independence.

In contrast, EVM:MM performed significantly worse than EVM with only FR information. Moreover, EVM:MM performance was below FR performance for uniform timing of interaction, which was not the case for MMIBN:OL. These findings show that our proposed models are more suitable for multi-modal biometric information than EVM.

5.6.5 Online Learning

We initially assumed that all identifiers work equally well on all users based on the work in (Jain et al., 2004). However, there can be changes in a person's appearance, the similarity between users, as well as changes in the time of interaction, which could negatively affect the visual identifiers and the time component of our models, respectively. We claimed that our online learning approach would adjust to these changes and perform better than the fixed learning rates (**H2**). The second part of the hypothesis was not supported because online learning (MMIBN:OL) performed significantly worse or at the same significance as fixed learning rates (MMIBN). The underlying reason might be the accumulating noise in the identifiers. Nevertheless, the average learned likelihoods (for 200 users) in online learning showed that the initial assumptions in Equation 3.7 hold valid. The mean for face node was 0.913 (compared to the initial assumption of 0.9), with SD $= 0.126$. For the gender likelihood, M $= 0.978$ (the initial assumption was 0.99), SD $= 0.058$.

FR does not perform equally well on users, as shown in Figure 5.11. Our proposed MMIBN models decrease the bias in the system using multi-modal information. This finding is also confirmed for uniform timing of interaction. Moreover, the first part of our hypothesis that online learning will adjust to these changes is supported, which allowed decreasing the bias of FR further. We can conclude that for long-term recognition, our multi-modal incremental Bayesian networks not only perform better than FR alone in all datasets but also increases performance on each user to identify them equally well.

5.6.6 Comparison to Previous Work and Real-World Applications

Our previous work, described in Chapter 4, showed that our proposed approach enables and facilitates incremental identification in a real-world HRI scenario. Our current results show that the model scales well to a larger dataset and provides significantly more reliable identification than face recognition alone. The hybrid normalisation method proposed here performs significantly better in terms of loss as compared to the previous methods used in earlier work (**H3**). Moreover, the Multi-modal Long-Term User Recognition Dataset described in this work contains a higher variability of subject age and heights, as well as patterned and uniform timing of interaction. Both D-Ten and D-All datasets show that the models perform significantly better than FR even though there are only 10 samples per user in D-Ten and a varying number of samples (10 to 41) in D-All, as compared to 66 samples per user in the previous work. Based on these conclusions, we can conclude that the optimised parameters presented here are more suitable than the ones previously found to use in real-world deployments for a large number of users.

5.7 Summary

In this chapter, we presented our Multi-modal Long-Term User Recognition Dataset that we created to simulate long-term human-robot interaction, allowing us to evaluate and optimise the parameters of Multi-modal Incremental Bayesian Network (MMIBN) on a large number of users (200) with varying characteristics. Within this dataset, we generated data with same or varying frequency of appearance of users, and patterned (Gaussian) interaction times, similar to educational or rehabilitation scenarios, and random (uniform) interaction times, similar to domestic interactions with a companion robot. We used simulated height estimations based on the identifier specifications found in the previous chapter. The remaining identifier estimations were obtained by feeding images of simulated users, obtained from IMDB-WIKI dataset, to the Pepper robot's proprietary algorithms, thereby providing real signals to our Bayesian Network.

We compared our proposed methods to face recognition (FR) and soft biometrics (SB), as well as a state-of-the-art open world recognition method, Extreme Value Machine (EVM).

The results show that the proposed MMIBN models with hybrid normalisation decrease the long-term recognition performance loss significantly and improve the identification rate significantly and substantially compared to all the baselines, in exchange for a higher number of incorrect estimations of new users.

Our MMIBN models generally perform significantly equivalent for both random and patterned timing, even though the time of interaction does not provide reliable information in the uniform timing case. Moreover, the results revealed that the proposed models perform equally well for learning new users incrementally sequentially (similar to batch learning) and at random intervals (similar to a real-world scenario). In addition, they perform equally well across training and open-set, and between closed-sets, even though the number of enrolled users is doubled, suggesting that the model scales well to larger datasets. While the models perform equally well for the same and varying frequency of appearance of users at the same number of recognitions, the models perform better with the increasing number of recognitions. Nevertheless, they can perform better than FR and EVM models after only a small number of recognitions.

Similar to the results in the previous chapter, online learning either decreases recognition performance or provides the same significance level as the non-adaptive model (MMIBN), which could be due to the accumulating noise in the network. On the other hand, the user-based analysis showed that online learning equalises the performance between users more than the fixed likelihoods, thereby, decreasing the biases in the system caused by FR.

In comparison to the user study presented in the previous chapter, this study showed that our proposed models scale well to a larger dataset with higher variability of subject age and heights, in addition to providing significantly more reliable user identification in comparison to face recognition and a state-of-the-art open world recognition algorithm. Hence, the optimised parameters in this chapter are more suitable to be used for real-world applications. Furthermore, our optimised proposed models take significantly lower amount of times in comparison to the state-of-the-art approaches, hence, MMIBN models are more suitable to be applied to robots for real-world long-term human-robot interactions as an initial step towards personalising the interaction.

Chapter 6

Task-Oriented Dialogue in a Coffee Shop: Barista Datasets

Key points:

- A text-based Barista Dataset is introduced, which simulates the interactions between a barista and various customers at a coffee shop.
- The Personalised Barista Dataset is introduced that extends the Barista Dataset with personalised subsequent interactions, in which the barista recalls and suggests customers' most frequent (preferred) orders. This dataset also includes user recognition information to identify the users. In order to simulate a real-world scenario with a robot, recognition errors, incorrect recalls and changes to the customer preferences are part of tasks in this dataset.
- The Personalised Barista with Preferences Information Dataset is introduced that contains the information of the most frequent order of the customer along with the user recognition information to simulate a knowledge-base extraction.

The Barista and Personalised Barista Datasets presented in this chapter have been used in the study described in Irfan et al. (2020b) and in the evaluation of data-driven approaches for long-term interactions, which will be presented in the next two chapters, respectively. The datasets are available online[1] for academic use based on the license terms.

[1] https://github.com/birfan/BaristaDatasets

Parts of the work are under review at the Frontiers in Robotics and AI journal.

6.1 Motivation

In Chapters 3 to 5, we have described a multi-modal user recognition system that is suitable for real-world long-term Human-Robot Interaction (HRI). From this chapter onwards, we will focus on recalling elements of previous exchanges, such as user attributes, preferences and behaviour patterns, for personalisation of the current interaction. We will use various memory systems in combination with our multi-modal user recognition system for designing fully autonomous personalised robots for real-world applications.

Previous research, presented in Chapter 2, shows that personalisation is vital for long-term interactions. It accommodates differences between individuals to maintain a positive user experience and facilitates user engagement and responsiveness after the novelty effect wears off. In addition, incrementally learning aspects about a user gives the impression of behavioural coherence and plausibility, which positively influences the perception of intelligence, thus, meeting user expectations and improving the quality of interaction. Based on our research question (RQ5), we identified the real-world application areas where personalisation can make an impact on long-term interactions in Section 2.1.4 to be the customer-oriented service and the healthcare domains. For instance, personalisation may facilitate familiarity, trust and rapport with users that encourage them to visit the same shop or restaurant again (Gwinner et al., 1998; Kanda et al., 2010; Niemelä et al., 2019). Moreover, personalisation may increase task efficiency and awareness of the situational context of the conversation (Neururer et al., 2018; Kocaballi et al., 2019).

Hundreds of customers visit a coffee shop every day. It would be highly ambitious and demanding for a barista to recognise "regular" customers and recall their preferences. On the other hand, robots can more easily recall a high number of customers more accurately. This would result in an improved user experience and decreased waiting times due to faster order taking process. Hence, deployment of personalised robots in coffee shops are desirable and beneficial for both the customers and the businesses. Consequently, in this thesis, we designed a personalised order-taking barista robot, as our second research objective (RO2).

Dialogue in a coffee shop is task-oriented, mostly sequential (i.e., requesting drink, followed by size and snacks), and limited to a specific vocabulary, such as the list of items in the menu, the available sizes, and the phrases to take and deliver an order. Hence, these interactions can be categorised under closed-domain task-oriented dialogue. Our review on the appropriate dialogue architectures for long-term interactions (RQ3) in Section 2.3 showed that most research in HRI uses rule-based approaches and knowledge-bases to obtain, store and recall pre-determined information through close-ended structured dialogues (Gockley et al., 2005; Kanda et al., 2007, 2010; Giuliani et al., 2013; Kasap & Magnenat-Thalmann, 2012; Churamani et al., 2017; Williams et al., 2018; Zheng et al., 2019). However, rule-based approaches might reduce the potential to adapt to variations in user responses. On the other hand, data-driven approaches offer more flexibility, however, they require a vast amount of data for training to learn the relevant information and how to use it. In addition, there are no studies that show how data-driven approaches perform on personalisation in long-term interactions. In this thesis, we explore both the potential of a rule-based approach (Chapter 7) and data-driven approaches (Chapter 8) for personalisation of task-oriented dialogue FOR long-term interactions.

In order to design a dialogue manager, we need to either know the rules that apply to order taking in a coffee shop or have a corpus of barista dialogues. While there are available corpora for restaurant booking (Henderson et al., 2014; Bordes et al., 2016; Joshi et al., 2017) or travel booking (Hemphill et al., 1990; Bennett & Rudnicky, 2002; El Asri et al., 2017) based on Wizard-of-Oz (WoZ), human-machine interactions, or simulated datasets (see Serban et al. (2015) for a recent survey on available corpora), there was no publicly available corpus on barista or personalised barista dialogues with customers at the time of conducting this work.

There are only two publicly available datasets that evaluate personalisation in any domain: Persona-Chat (Zhang et al., 2018) and Personalized bAbI dialog (Joshi et al., 2017) datasets. Persona-Chat dataset contains text-based open-domain conversations of crowdsourced workers that received a set of sentences determining their personality for the dialogue. The Conversational Intelligence Challenge[2] was created based on this dataset (Dinan et al., 2019). The Personalized bAbI dialog dataset is a simulated text-based personalised dataset built upon the bAbI dialog (Bordes et al., 2016) dataset for restaurant booking. The dataset focuses on adapting conversation and recommendation styles based on the user's gender and age, in addition to recommending restaurants based on the dietary preferences and favourite food item of the user. However, this dataset focuses on personalising the dialogue based on general attributes (gender and age), instead of adapting to each user.

[2] `http://convai.io`

Moreover, user attributes are pre-defined at the beginning of each dialogue, instead of obtained from the interaction. Both of these datasets consider only a single user interaction, instead of long-term interactions.

Previous research in HRI debates against using WoZ as a data collection method, because it may create unrealistic expectations and may result in findings that are not grounded in realistic interactions (Fernaeus et al., 2009; Riek, 2012). In addition, a vast amount of data is necessary to train and evaluate data-driven approaches, and recruiting a high number of subjects for several subsequent interactions is a challenge of long-term HRI, as we have previously discussed in Section 2.1.1. Hence, the only remaining option is to generate simulated datasets based on anticipated phrases in the interaction. Consequently, we created the Barista and Personalised Barista Datasets described in this chapter. Personalised Barista datasets are the first released[3] datasets to explore user-specific personalisation in task-oriented long-term interactions.

6.2 Barista Dataset

The Barista Dataset is designed to model a real-world barista[4] who: (1) greets and requests the drink order, (2) size, and (3) snack, (4) confirms the order, (5) changes the order if necessary, (6) takes the customer's name, (7) notes the order pick up location, (8) says goodbye. Typically, a customer can ask for the order in one sentence, removing the need of (2) and (3), however, we separated these steps to reduce the errors in rule-based (e.g., template matching) or data-driven approaches, and to aid speech recognition for the robot.

This dataset is used for the non-personalised robot in Chapter 7 and to evaluate the performance of data-driven approaches on non-personalised tasks in Chapter 8. The Personalised Barista Datasets described in the next sections contain the interactions in this dataset for new customers (users) and personalise the interaction for known customers on top of this structure.

Similar to the bAbI dialog (Bordes et al., 2016) and the Personalized bAbI dialog (Joshi et al., 2017) datasets, we identified the barista dialogue tasks based on the sequential interactions described above. However, contrary to the bAbI datasets, which structure

[3]For academic use only: `https://github.com/birfan/BaristaDatasets`

[4]The author of this thesis used her previous experience in interacting with baristas at coffee shops to design the phrases in the Barista Datasets.

tasks to evaluate model's capability to do application program interface (API) calls or use knowledge-base (KB) facts, we structure our tasks based on the different interaction types of dialogues (e.g., ordering only a drink, making changes to an order), which can occur in a typical scenario, in increasing difficulty of the interaction. In addition, we initially leave out the "greetings" of the agent, because using the name obtained during a conversation may decrease the performance for data-driven approaches, as we will investigate in Chapter 8. Hence, we define the tasks in the Barista Dataset as follows:

- **Task 1 (B1):** *Greetings*. This task evaluates greeting and requesting the drink (1), taking the customer's name (6), noting the order pick up location (7) and saying goodbye (8) to the customer. Some of the utterances in this task may not occur in real-world barista interactions. For instance, the customer can start with the phrase involving the drink order, without the necessity of the barista requesting their order. Moreover, in many coffee shops, the order is delivered by the same barista, hence there is no need to take the name or note the order location. In addition, using the customer's name obtained from the dialogue may decrease the performance for data-driven approaches. Hence, we separately analyse the performance on this task. No order is made.
- **Task 2 (B2):** *Order drink (without greetings)*. This task evaluates ordering a drink.
- **Task 3 (B3):** *Order drink with changes*. This task evaluates ordering a drink and changing the order (up to two changes) during the interaction.
- **Task 4 (B4):** *Order drink and snack*. This task evaluates ordering a drink and a snack.
- **Task 5 (B5):** *Order drink and snack with changes*. This task evaluates ordering a drink and a snack and changing the ordered items (up to two changes) during the interaction.
- **Task 6 (B6):** *Order drink and snack with greetings*. This task is the combination of tasks 1 and 4.
- **Task 7 (B7):** *Order drink and snack with changes and greetings*. This task is the combination of tasks 1 and 5, and contains interaction types from all tasks.

A dialogue example is given in Figure 6.1 for **Task 7**. Each interaction type is colour-coded, and the legend explains the colours corresponding to the task number for clarity. A snack

Figure 6.1: A dialogue example in the Barista Dataset Task 7, showing all the tasks in the dataset.

Table 6.1: Number of phrases for customer and bot (multi-phrases) per each utterance in the Barista Dataset. Note that in the dataset used in Chapter 8, the bot has a single phrase per utterance. The utterances are colour-coded according to the interaction type, corresponding to Figure 6.1.

Customer			Bot (Multi-Phrases)		
Greeting		4	Greeting and request drink		16
Order drink		16	Request size		3
State size		5	Request snack		4
No snack		9	Confirm order without snack		2
Order snack		17	Confirm order with snack		2
Confirm order		8	Request name		4
Change order	with snack	22	Confirm change	with snack	6
Give name		2	Note order location	with name	3
Farewell		10	Farewell		5

order or a change in the order is not compulsory, and each has a probability (for occurring) of 0.5, sampled from a uniform distribution. Note that for tasks 2 and 3, instead of *"Hmm, a veggie tortilla wrap."*, the customer says *"No."* (or a variant). *api_call getCustomerName* refers to using an API call to another resource (e.g., tablet) or a script to obtain the customer's name, and it is not said to the customer. The number of conversation turns (i.e., user-bot utterance pairs) range from 7 (no changes in the order) to 9 (two changes in the order) in the tasks involving orders.

Table 6.1 shows the number of customer utterances for each type of bot utterance. We initially designed a dataset with various multiple bot phrases (i.e., *bot (multi-phrases)*) along with multiple customer phrases per utterance, however, the data-driven approaches used in Chapter 8 are not designed to learn multiple phrases of the bot for the same customer utterance, and preliminary experiments showed very low performance, hence, we used the same phrase for each bot utterance. However, in long-term HRI, it is common to select a phrase from a variety of phrases to avoid repetition and sound more natural, thus, we believe that data-driven approaches in dialogue should also give more importance to this topic for long-term interactions. Hence, we release an additional set of the datasets with multiple bot phrases per utterance. This additional dataset was used for the user study presented in Chapter 7, since using various phrases does not affect the system performance for a rule-based dialogue manager, but the single bot phrase dataset was used in Chapter 8.

Similar to bAbI dialog datasets, we divided our datasets into *training*, *validation* (*development*), *test* and *out-of-vocabulary (OOV)* sets. The *training*, *validation* and *test* sets use the same drink, size, and snack types. *OOV* set contains different drink, size, and snack types that are not part of (i.e., "seen in") the other sets. Regardless, all the items come from the Starbucks menu[5] and there are 20 drink types, 3 sizes, and 20 snacks that the customer can order from in each set. We used a smaller menu based on the store-bought items (11 drinks, 6 snacks) in Chapter 7. The customer and bot phrases used in the *OOV* set is the same as the other sets.

Because our datasets focus on the long-term interactions, the identity of the customer is important, especially for personalisation. In addition, in some coffee shops, the name is asked to separate the customer orders and for announcing the order when it is ready. Unlike those coffee shops, we use the full name of a person for the identity, because using the first name may cause mixtures of orders in a real-world HRI scenario because the verification of the identity is based only on the customer's name. We used 100 names (let's call it, *customer-base A*) from the Multi-modal Long-Term User Recognition Dataset (in Chapter 5) for the *training*, *validation* (*development*) and *test* sets in the Barista Dataset. We extracted an additional 100 names from the IMDB-WIKI (Rothe et al., 2015, 2018) dataset (*customer-base* C) for the *OOV* set.

We used two *dataset sizes* in order to evaluate the task performance depending on the training and evaluation size: 1,000 dialogues (similar to bAbI datasets) and 10,000 dialogues to account for the increased difficulty of the tasks arising from the various names in the dialogues. Table 6.2 presents the *task size* (i.e., the number of customer-bot utterance pairs in the task) and the size of the *vocabulary* (i.e., the unique words in a task). Similar to the bAbI dialog datasets, the *candidate set* contains the unique utterances of the bot in all tasks and sets and is equal to 4,149 in 1,000 dialogues and 5,207 in 10,000 dialogues.

The task difficulty increases when there is a piece of personal information (e.g., customer name) or order details of the customer in the bot utterance, as the dialogue architecture should extract this information from the previous exchanges and use it to respond. Hence, we present the percentage of *personal(ised)*, *order details* and *other (remaining)* phrase types for each task in Table 6.3 for the *test* set, such that we can evaluate the performance of the data-driven approaches in this perspective in Chapter 8. The phrase type information for the remaining sets is presented in Appendix D.1.

We created the dataset in the format suitable for ParlAI[6] (Miller et al., 2017) (i.e., line

[5] `https://www.starbucks.com/menu`

[6] `https://parl.ai/`

Table 6.2: The task sizes (i.e., number of customer-bot utterance pairs) of *training*, *validation*, *test* and *OOV* sets, and the vocabulary (i.e., number of unique words in a task) size for 1,000 and 10,000 dialogues Barista Datasets. The vocabulary size for both dataset sizes is the same. Note that the task size increases in proportion to the number of dialogues, and differs according to the task.

	Dataset Size	B1	B2	B3	B4	B5	B6	B7
Training Task Size	1,000	4,000	4,000	4,767	4,000	4,734	7,000	7,764
	10,000	40,000	40,000	47,399	40,000	47,378	70,000	77,491
Validation Task Size	1,000	4,000	4,000	4,752	4,000	4,724	7,000	7,740
	10,000	40,000	40,000	47,489	40,000	47,486	70,000	77,582
Test Task Size	1,000	4,000	4,000	4,757	4,000	4,760	7,000	7,784
	10,000	40,000	40,000	47,467	40,000	47,506	70,000	77,490
OOV Task Size	1,000	4,000	4,000	4,772	4,000	4,718	7,000	7,746
	10,000	40,000	40,000	47,547	40,000	47,538	70,000	77,541
Vocabulary Size		432	331	350	333	351	446	463

numbers for each dialogue, and tab-separated customer and bot utterances) platform in order to evaluate tasks using available models in that platform and to submit the Barista and Personalised Barista datasets to ParlAI for contributing to the research community in evaluating their algorithms for personalisation in long-term interactions.

While we created the Barista Dataset due to the lack of an available dataset with barista and customer interactions at the time (January 2019) and for the corresponding real-world study (Chapter 7, August 2019), we would like to note that very recently (in October 2019) Google Research released a crowdsourced dataset called the Taskmaster[7] (Byrne et al., 2019) that was obtained from conversations with a personal digital assistant through WoZ or by "self-dialog" (i.e., crowdsourced workers imagined having a dialogue with a personal digital assistant and write the interaction for both sides). This dataset contains

Table 6.3: The percentage of *personal(ised)* (i.e., containing user name), *order details* (i.e., containing an item from the order) and *other (remaining)* phrase types in the bot utterances for the tasks of 1,000 and 10,000 dialogue Barista *test* set.

Dataset Size	Phrases	B1	B2	B3	B4	B5	B6	B7
1,000	Personal	25	0	0	0	0	14.29	12.85
	Order	0	25	36.94	25	36.97	14.29	22.92
	Other	75	75	63.06	75	63.03	71.43	64.23
10,000	Personal	25	0	0	0	0	14.29	12.9
	Order	0	25	36.8	25	36.85	14.29	22.57
	Other	75	75	63.2	75	63.15	71.43	64.52

[7] https://github.com/google-research-datasets/Taskmaster

conversations for ordering a pizza, creating auto repair appointments, setting up a ride service, ordering movie tickets, making restaurant reservations and ordering coffee drinks for pick up at a store. However, the Taskmaster only includes drink orders, hence, excluding snacks, which is common to order at a coffee shop. In addition, the changes to the drink orders only occur if the drink is not available, which is a rare case in real-world scenarios, whereas we believe it is more reasonable to assume that the customer can change their mind while ordering. Most importantly, Taskmaster does not contain customer names or personalised subsequent interactions to evaluate personalisation in long-term interactions that we present in the next sections.

6.3 Personalised Barista Dataset

As we have previously discussed, recognising "regular" customers and recalling their preferences are important for long-term deployment of robots in the customer-oriented service domain. However, *which type of information should be recalled for personalisation?* (RQ4). Previous research presented in Chapter 2 revealed that user's attributes (e.g., name), preferences, behaviour patterns and previous shared history are commonly used in HRI to personalise the interaction. From these aspects, the name and the preferences (i.e., the most common order) of a user are suitable to recall in a barista interaction. However, in a text-based interaction (e.g., for chatbots), the user can log in with their information in a system, hence, the identity of the user would be already available, whereas, for HRI in a real-world scenario, the person should be autonomously recognised. In the previous chapters, we described a user recognition system that is suitable for this interaction. In order to integrate this information into the text-based dataset, we can use the type of information obtained from the user recognition system: (1) whether the user is known (i.e., *False* if new, *True* if enrolled), (2) the identity number (ID) of the user (i.e., *0* if the user is new, otherwise, the ID of the user which is given in the order of meeting users), and (3) the name of the user. These three types of user profile information will be sufficient to recall the favourite orders of the user for a rule-based approach with a knowledge-base, as used in Chapter 7. On the other hand, while there is no research on personalisation in long-term interactions using data-driven approaches, the structure of current approaches seems suitable to track the previous conversations using this information. Consequently, we extend the Barista Dataset with personalised subsequent interactions and user recognition information to create the Personalised Barista Dataset.

In a real-world human-robot interaction scenario, a lot of things can go wrong. For example, automatic speech recognition (ASR) may not perform well, which may cause the dialogue manager to receive incomplete or incorrect information. In addition, as we have seen in the previous chapters, user recognition may not be fully reliable for a high number of users or due to noisy data. Moreover, especially for data-driven approaches, there may be incorrect recalls of user information or knowledge-base entities (Bordes et al., 2016). Therefore, it is important to account for these errors, and know how to recover from them. Correspondingly, we defined the tasks as follows:

- **Personalised Task 0 (PB0):** *Confirmed personalised order suggestion for new customers*. This task learns and evaluates the recall of the preferences of new customers. The most common or the most recent (in the case of ties for the most common order) drink and snack of the customer are suggested, and the customer accepts the suggestion. This task assumes perfect recognition and recall, and no changes are made by the customers to their previous preference. An example is given in Figure 6.2.

  The *training* set has 100 users from the *customer-base A* as described in the previous section, but the *validation* and *test* sets have 100 **different** users (*customer-base B*) from the Multi-modal Long-Term User Recognition Dataset. The *OOV* set has 100 users from the *customer-base C*, corresponding to the customer-base in the *OOV* set of the Barista Dataset. In other words, this task aims to evaluate the performance of learning the preferences of a different set of customers than the ones in the training set. The reason we separate this task (and call it Task 0) is that it evaluates learning new users incrementally without any prior information (i.e., *0* previous data samples, known as *zero-shot learning*) of those users.

- **Personalised Task 1 (PB1):** *Confirmed personalised order suggestion for previous and new customers*. This task has the same type of dialogue interactions as PB0, but the *validation* and *test* sets also have customers from the *customer-base A*, i.e., a total of 200 customers in both sets. This task requires remembering the orders of the previous "regular" customers and incrementally learning the preferences of the new ones. In other words, this task measures the ability of the data-driven approaches to deal with *catastrophic forgetting*, which refers to the tendency to forget previously learned information upon learning new information (McClelland et al., 1995; McCloskey &

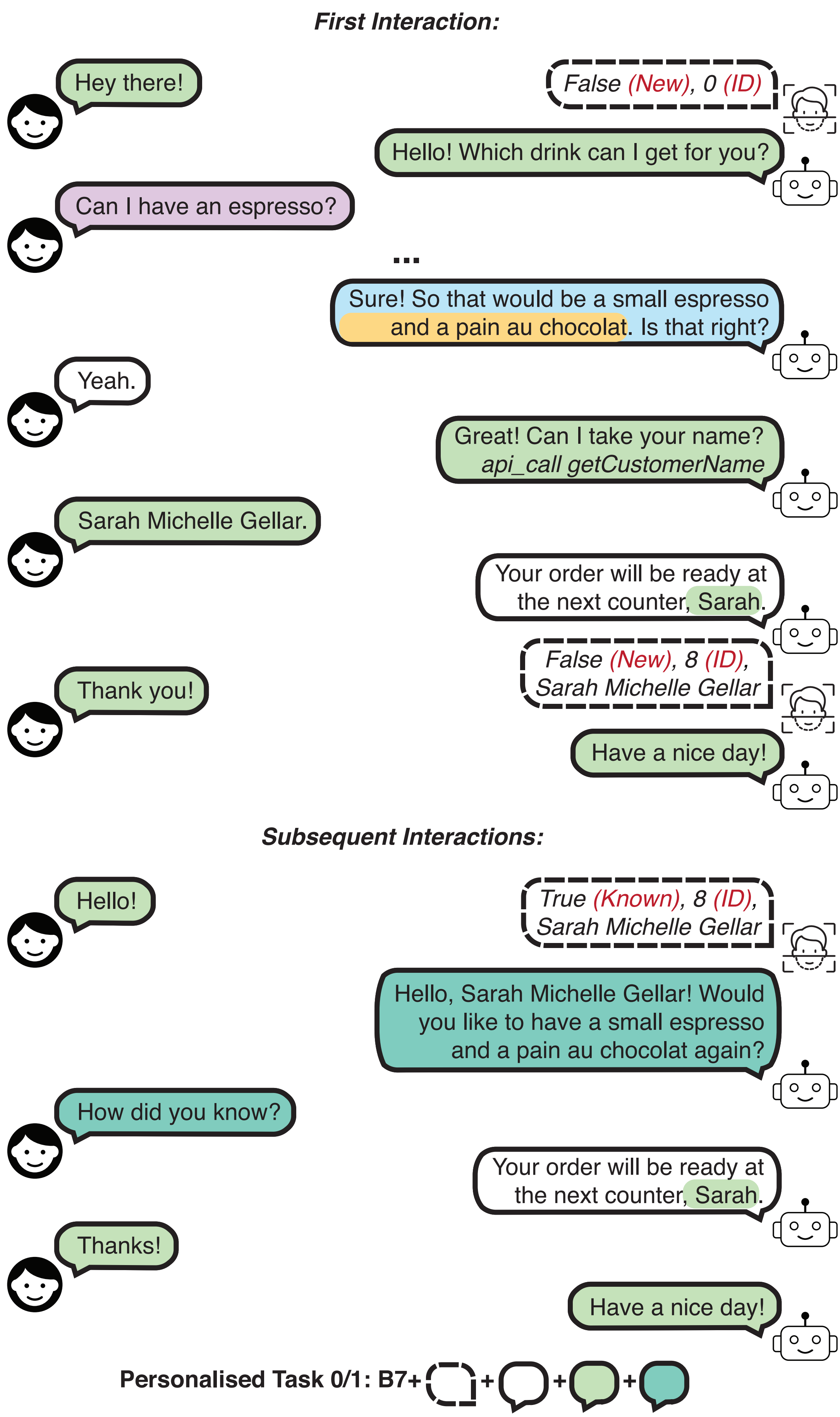


Figure 6.2: A dialogue example in the Personalised Barista Dataset Task 0 and 1 (confirmed personalised order suggestion).

Cohen, 1989; Parisi et al., 2019). In other words, this task evaluates the performance of the data-driven model in learning information continuously, incrementally and adaptively, while remembering previously learned data, known as *continual (or lifelong) learning* in machine learning. The following tasks build upon this task.

- **Personalised Task 2 (PB2):** *Recognition error*. This task evaluates the bot to correct itself after incorrect recognitions in open world recognition. As Chapters 4 and 5 showed, the following type of incorrect recognitions can occur:

  – *Customer is known, but confused with another customer*. An example is shown in Figure 6.3.

  – *Customer is known, but not recognised*. The dialogue is conducted as if a new customer is encountered.

  – *Customer is new, but confused with another customer*. The dialogue is similar to a new customer interaction after the recognition correction.

  We used Detection and Identification Rate (DIR) of 0.9 and False Alarm Rate (FAR) of 0.1. In other words, 90% of the enrolled customers are correctly recognised and 10% of the new customers are confused with an enrolled customer. Hence, this task contains incorrect recognitions as well as correct ones. The reason we used a high DIR and a relatively low FAR is to evaluate if the data-driven approaches can learn to respond to these errors when there is only a small number of erroneous recognitions in the *training* set.

- **Personalised Task 3 (PB3):** *Incorrect recall*. This task teaches and evaluates the bot to correct itself after an incorrect recall of the preferences of the customer. A dialogue example is shown in Figure 6.4. An incorrect memory rate of 0.3 is used, that is 30% of the dialogues contain incorrect recalls of the preferences of known customers. In other words, a random order is suggested, and we ensured that it does not correspond to the preferred order of the customer. Because the customer's order is taken after the incorrect recall, this type of dialogue has phrases from the B7 for the first and subsequent interactions, denoted as $B7^2$. Note that the correct labels for the incorrect recall errors may cause the system to reject some of the correct recall of preferences during training and evaluations for data-driven approaches, which may reduce the model performances, however, it is important to train the models for handling incorrect recalls.

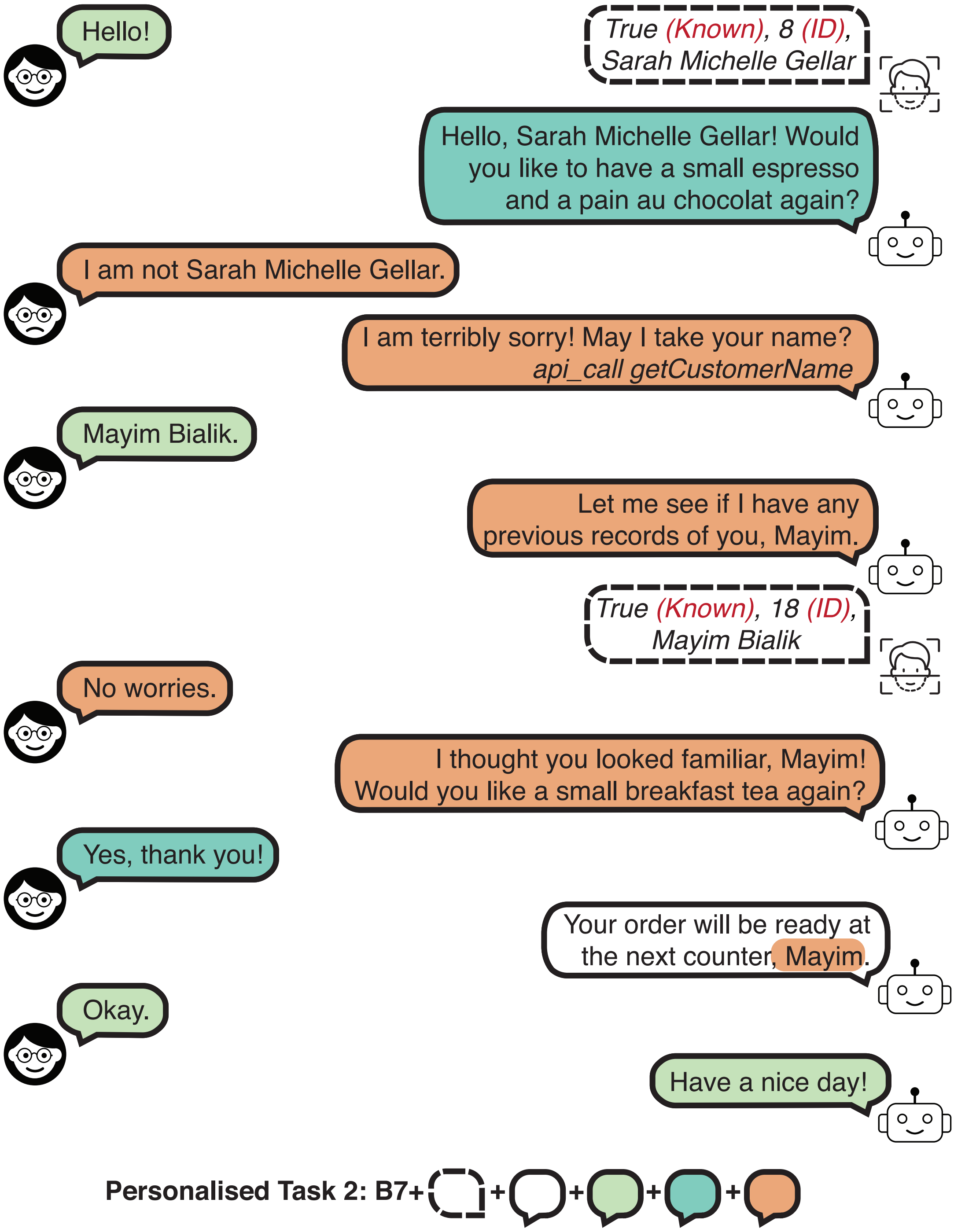


Figure 6.3: A dialogue example in the Personalised Barista Dataset Task 2 (recognition error).

Figure 6.4: A dialogue example in the Personalised Barista Dataset Task 3 (incorrect recall).

- **Personalised Task 4 (PB4):** *Changes to preference.* This task acknowledges that customers may change their preferences. A dialogue example is presented in Figure 6.5. In any interaction, the customer may want to have their most common order (as in PB0) or can ask for some other drink or snack (i.e., including phrases with B7^2 and the specific interaction phrase for the customer asking for a different drink). A change in preference has a probability of 0.5, sampled from a uniform distribution.

Figure 6.5: A dialogue example in the Personalised Barista Dataset Task 4 (changes to preference).

- **Personalised Task 5 (PB5):** *Recognition error and incorrect recall*. This task is the combination of tasks 2 and 3, which can occur within the same dialogue or separately.

- **Personalised Task 6 (PB6):** *Recognition error and changes to preference*. This task is the combination of tasks 2 and 4, which can occur within the same dialogue or separately.

- **Personalised Task 7 (PB7):** *Incorrect recall and changes to preference*. This task is the combination of tasks 3 and 4, which can occur within the same dialogue or separately.

- **Personalised Task 8 (PB8):** *All tasks*. This task is the combination of tasks 2, 3 and 4. This task evaluates all the scenarios that can occur in a personalised barista interaction.

Personalisation in a real-world HRI scenario involves incrementally and adaptively learning users, as we emphasised in the previous chapters. Hence, it is required to learn users' preferences from a single data point, referred to as *few-shot learning*. While this problem is a trivial one for rule-based approaches with a knowledge-base, this poses a very difficult problem for data-driven approaches, as they require a vast amount of data for training (Triantafillou et al., 2017). Consequently, in addition to the 1,000 and 10,000 dialogues datasets, we designed the *Second Interaction* dataset. We use two samples of data (i.e., the first and second interaction) in the *training* set for task 0 (i.e., for new users), and three samples of data (i.e., the first, second and third interaction) in the remaining tasks to account for learning to suggest the previous order of a new user, and learning to count the most common or recent order of a previous user. The number of dialogues per task is presented in Table 6.4 in the format *the number of dialogues in PB0 - the number of dialogues in other tasks*, along with the number of customer-bot utterance pairs (i.e., task size) and the number of unique words in a task (i.e., vocabulary size). The reason why PB0 has 200 dialogues is that it only has 100 users in the training/evaluation sets, in other words, previous users are not evaluated in that task. In the *validation* and *test* sets, the previous user (from *customer-base A*) in the *training* set will be seen for the fourth or fifth time, whereas, the new user (from *customer-base B*) will only be encountered two times. In the 1,000 and 10,000 dialogue datasets, the number of interactions per user is set to be the same to avoid certain orders to appear more than others. For example, in PB1, user ID 5 is seen ten times in the 1,000 dialogue *training* set, and five times in the *validation* and

Table 6.4: The task sizes and the vocabulary sizes for Second Interaction, 1,000 and 10,000 dialogue Personalised Barista Datasets. Note that the task size increases in proportion to the number of dialogues, and differs according to the task.

	Dataset Size	PB0	PB1	PB2	PB3	PB4	PB5	PB6	PB7	PB8
Training Task Size	200 - 300	1,075	1,377	1,470	1,722	1,728	1,744	1,817	1,921	1,946
	1,000	3,482	3,480	3,787	4,765	5,113	5,082	5,348	5,934	6,337
	10,000	30,482	30,481	33,424	44,184	48,830	47,289	51,933	57,569	60,355
Validation Task Size	200 - 400	1,080	1,312	1,385	1,720	1,731	1,703	1,819	1,918	1,972
	1,000	3,474	3,466	3,748	4,772	5,189	5,017	5,479	5,850	6,176
	10,000	30,463	30,480	33,488	44,468	48,749	47,622	51,962	57,459	60,642
Test Task Size	200 - 400	1,064	1,298	1,386	1,580	1,715	1,752	1,803	1,860	2,018
	1,000	3,471	3,489	3,768	4,728	5,141	5,125	5,460	5,873	6,233
	10,000	30,481	30,484	33,404	44,762	49,364	47,515	52,205	57,420	60,689
OOV Task Size	200	1,064	1,066	1,106	1,210	1,244	1,286	1,324	1,325	1,426
	1,000	3,463	1,967	2,090	2,535	2,766	2,603	2,868	3,070	3,195
	10,000	30,469	15,467	16,969	22,540	24,572	23,912	26,184	28,999	30,309
Vocabulary Size	200 - 400	959	948	959	972	957	973	966	972	980
	1,000	959	959	971	975	965	983	977	981	989
	10,000	959	959	971	975	965	983	977	981	989

test sets, and user 105 (who is not seen in the *training* set) is also seen five times in the *validation* and *test* sets.

As can be seen from a comparison of Table 6.2 and Table 6.4, the number of customer-bot utterances has decreased. The reason is that personalisation decreases the necessary number of conversation turns to make an order (down to 3 turns), which improves the efficiency of an agent and decreases the time required to make an order, as we initially intended. On the other hand, the candidate set (i.e., the unique bot utterances in *training*, *validation* and *test* sets) is increased substantially: 7,715 for Second Interaction, 13,859 for 1,000 dialogues, and 53,304 for 10,000 dialogues, which is ten times the number in 10,000 dialogues for the Barista Dataset. The reason is the new users and the corresponding orders in the *validation* and *test* sets, in addition to the additional phrases for personalisation tasks as shown in Table 6.5, which also increases the vocabulary size. Similar to the Barista Dataset, one phrase is used per bot utterance in the datasets in Chapter 8, whereas, multiple phrases of the bot are used in the study in Chapter 7 to improve user engagement.

In order to understand how well the data-driven approaches perform for personalisation

Table 6.5: Number of additional phrases for customer and bot (multi-phrases) per each utterance in the Personalised Barista Dataset. Note that in the dataset used in Chapter 8, the bot has a single phrase per utterance. The utterances are colour-coded according to the interaction type corresponding to the dialogue examples presented.

Customer		Bot (Multi-Phrases)	
Greeting	4	Known customer - Suggest favourite order	8
Confirm preference	24	Note order location with name	3
Recognition error response	13	Apologise and take name	12
Give name	2	Look into records	2
Waiting for looking into records	5	Known customer - Suggest favourite order	3
Waiting for looking into records	5	New customer - Request drink	3
Incorrect order recall response	12	Request new drink	25
Changes to preference	9	Request size	3

of the interaction, the proportions of *personal(ised)* bot utterances (i.e., containing user name or preferences), *order details* (i.e., containing new order or preferences), the *other (remaining)* dialogues, and the phrases belonging to the B7 task are presented in Table 6.6 for the *test* set. Note that since both the *personal(ised)* and order phrases contain user preference in the Personalised Barista Datasets, the sum of percentages of personal, order and other phrases is higher than 100%. The percentage information for the *training*, *validation* and *OOV* sets are presented in Appendix D.2.

For each task, the orders of the customers are stored in a knowledge-base containing the interaction number, customer identity (i.e., ID number and name), and the final order in the dialogue. This knowledge-base was used in the experiments in Chapter 8 for

Table 6.6: The percentage of *personal(ised)* (i.e., containing user name or preference), *order details*, *other (remaining)* and Barista Task 7 (B7) phrase types in the bot utterances for the tasks of Second Interaction, 1,000 and 10,000 dialogue Personalised Barista *test* set.

Dataset Size	Phrases	PB0	PB1	PB2	PB3	PB4	PB5	PB6	PB7	PB8
200 - 400	Personal	28.2	39.6	41.7	35.63	29.85	38.18	32.22	30.7	33.15
	Order	24.81	27.27	27.71	29.62	32.13	29.79	32.61	31.72	32.11
	Other	56.39	49.61	47.98	48.1	50.38	46.23	48.64	49.03	47.13
	B7	90.6	83.51	77.85	83.35	87.64	76.77	82.7	85.43	79.88
1,000	Personal	54.74	54.46	55.94	45.79	36.96	46.89	38.39	36.69	37.98
	Order	30.86	31.21	31.05	32.42	36.12	32.64	36.03	35.28	35.54
	Other	40.33	40.13	39.49	40.82	44.43	40.1	43.77	43.35	42.44
	B7	74.07	74.2	67.94	75.36	82.49	70.54	78.21	80.33	76.54
10,000	Personal	65.29	65.28	65.45	51.18	40.31	52.17	42.1	39.79	41.13
	Order	33.07	33.08	33.09	34.25	38.39	34.25	38.03	37.05	36.81
	Other	34.12	34.12	34	36.69	41.35	36.4	40.8	40.4	40.06
	B7	67.52	67.52	61.56	71.16	79.94	66.88	75.07	77.62	73.67

evaluating the rule-based dialogue management system (RBDMS) (that will be introduced in Chapter 7) on the Personalised Barista Datasets, however, data-driven approaches do not have access to this information.

6.4 Personalised Barista with Preferences Information Dataset

The Personalised Barista Dataset evaluates if the data-driven approaches can learn to track the previous conversations to extract the most common order of a user, in addition to using that information to personalise the conversation. Thus, there are two problems of long-term interactions that it addresses, which is missing in the currently available datasets. However, the requirement for tracking previous orders and "calculating" the most common order may pose a high level of difficulty for a data-driven approach, which is a trivial problem to a rule-based approach that uses a knowledge-base. Hence, we created the Personalised Barista with Preferences Information (PBPI) Dataset to provide the information of the most common order of the user at the beginning of the conversation, to simulate extracting the information from a knowledge-base. This information is given alongside the user identity information at the beginning of the dialogue, similar to the Personalized bAbI dialog dataset. For example, for the customer in Figure 6.1, the information in this dataset will be in the format of [*True (Known), 8 (ID), Sarah Michelle Gellar (customer name), small (the most common size of the most common drink order), espresso (the most common drink order), pain au chocolat (the most common snack order)*]. The tasks, the phrases and the corresponding task, vocabulary and candidate set sizes of the Personalised Barista with Preferences Information Dataset are the same as that of the Personalised Barista Dataset.

6.5 Summary

This chapter introduced the text-based Barista and Personalised Barista Datasets, which contain simulated generic (i.e., non-personalised) and personalised dialogues for long-term interactions between a barista and a customer. The Personalised Barista Dataset contains previous and new user interactions, and the user recognition information is given at the beginning of the dialogue to identify the user. Additionally, the Personalised Barista with Preferences Information Dataset contains the order and user recognition information to simulate a knowledge-base query result. These datasets are used for structuring the dialogue of the non-personalised and personalised barista robots (Chapter 7) and evaluating the potential of data-driven approaches in generic and personalised long-term interactions (Chapter 8).

Chapter 7

Personalised Barista Robot: Real-World Study with Non-native English Speakers

Key points:

- Fully autonomous generic and personalised architectures are developed based on the Barista and Personalised Barista Datasets, using the Multi-modal Incremental Bayesian Network (MMIBN) for user recognition, online automatic speech recognition and a rule-based dialogue management system.
- The first real-world study that explores fully autonomous personalisation in dialogue for long-term human-robot interactions is conducted at an international student campus with non-native English speakers.
- Unreliable speech recognition and the inflexible structure of the rule-based dialogue manager negatively affected the user experience.
- The results indicate that personalisation mitigates the negative user experience.

Parts of the work presented in this chapter have been published in Irfan et al. (2020b)[1]. The final publication is available from ACM in the Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction[2], at DOI: 10.1145/3371382.3378278.

[1]This work was conducted in collaboration with Mehdi Hellou, who was an intern at SoftBank Robotics Europe between February-August 2019 under the main supervision of the author of this thesis and co-supervision of Alexandre Mazel. In addition to supervising Mehdi Hellou, the author designed the Barista Robot and the rule-based dialogue management system (RBDMS) architectures, created the corresponding datasets, designed the experimental procedure and the evaluation methods for the user study and conducted the analysis of the study. Mehdi Hellou implemented the rule-based dialogue management system (RBDMS) and combined the user recognition system developed by the author (as described in Sections 3 and 5) with automatic speech recognition systems on the robot, conducted the real-world HRI study outlined in this chapter, and contributed to the analysis of the results.

[2]Video presentation of the study is available online: https://www.youtube.com/watch?v=_g2H1Dk83wQ

7.1 Motivation

As we discussed in the previous chapter, our second objective (RO2) in this thesis is to design a personalised customer-oriented service robot. We hypothesise that the personalisation will improve the user experience and increase the task efficiency. We identified order taking at a coffee shop to be a task-oriented interaction that can be personalised with the use of a robot in a real-world application. Consequently, we created the Barista and Personalised Barista Datasets, as described in the previous chapter, to build the interaction upon. As we previously noted in Chapter 2, autonomy is an integral part of long-term interaction, especially in a real-world scenario. The previous studies (Kasap & Magnenat-Thalmann, 2012; Zheng et al., 2019; Churamani et al., 2017) explored fully autonomous personalisation in dialogue for long-term interactions and showed that personalisation improves the user experience, task efficiency and the perceived social intelligence of the robot. However, none of these studies was deployed in the real-world. Conducting real-world studies, as we previously highlighted in Chapter 2, brings additional challenges with it, such as the lack of robust technology, noisy data, high user expectations, the low number of users and non-native speakers. In this chapter, we touch upon these challenges and the impacts of personalisation with the first real-world study in exploring fully autonomous personalisation in dialogue for long-term Human-Robot Interaction (HRI).

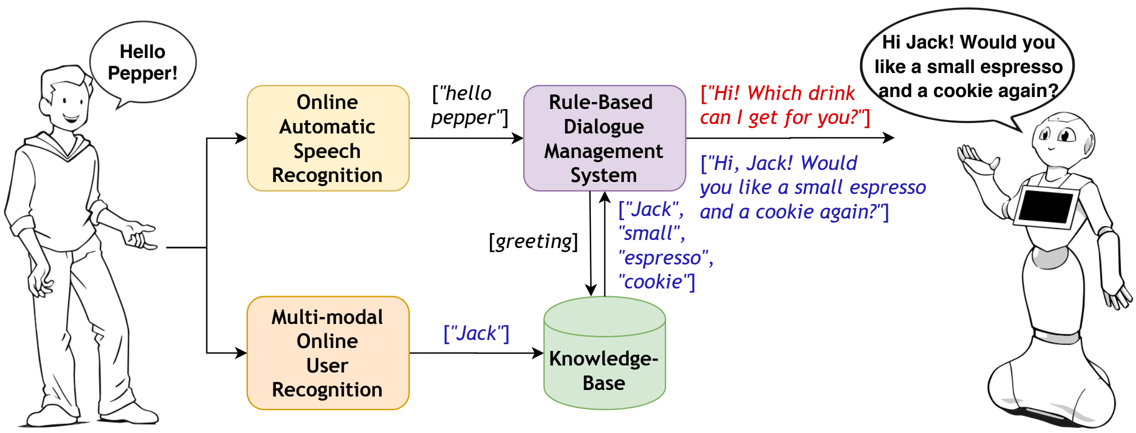


Figure 7.1: The general components of the Barista Robot architecture. The actions and inputs common to all conditions are coloured in black, and those specific to the *generic barista* are coloured in red and the *personalised barista* in blue. *The artwork of the Pepper robot and the man belong to SoftBank Robotics Europe. Permission was granted for use.*

7.2 Methodology

In the previous chapter, we created a set of rules to build the Barista and Personalised Barista Datasets to simulate a *generic* and *personalised barista* interaction. However, going from a text-based dataset to a fully autonomous personalised robot, especially for long-term interactions, is not a trivial task, and requires many robust components, such as user recognition, online automatic speech recognition (ASR), a dialogue manager and a knowledge-base (KB), as shown in Figure 7.1. The architecture of the Barista Robot[3] and the flow of the interaction is provided in Figure 7.2. This section further explains the architecture in detail.

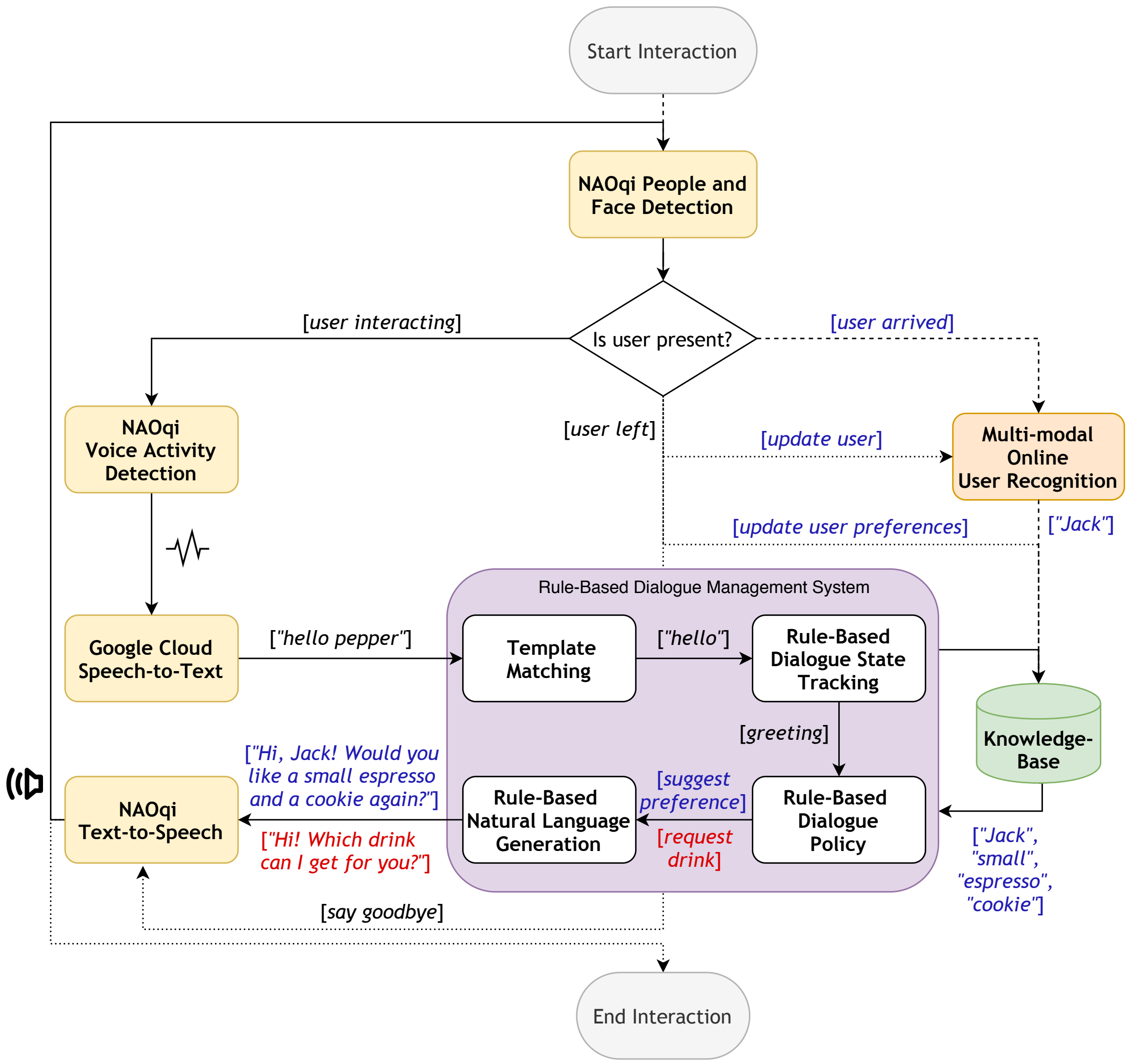


Figure 7.2: The flow chart of the Barista Robot architecture. The dashed line shows the actions taken when the user arrives, the solid line represents the action flow during the interaction of a user, and the dotted line represents the actions taken after the user left. The actions and inputs common to all conditions are coloured in black, those specific to the *generic barista* are coloured in red and the *personalised barista* in blue.

[3]Demonstration of the robot behaviours is available at: `https://youtu.be/eAOnHlDuHqw`

7.2.1 User Recognition

In Chapters 3 to 5, we described our proposed solution, Multi-modal Incremental Bayesian Network (MMIBN), which allows fully autonomous and incremental user recognition in long-term interactions that are fundamental in this real-world study. Our previous results showed that the non-adaptive MMIBN performed better than the approach with online learning (MMIBN:OL), hence, we used MMIBN for user recognition in this study.

Our previous results suggested that while our approach significantly improves face recognition, due to the noise in the identifiers, it cannot provide 100% recognition. Thus, the confirmation of the identity is necessary to ensure that the correct person has been identified. In our previous study in Chapter 4, we used explicit confirmation of the identity. In contrast, in this study, to reduce the efforts necessary by the users and provide naturalness in interaction, we decided to confirm the identity implicitly from the dialogue. We assume that if the estimated identity is incorrect, the user will notify the robot of its error. An example of indirect extraction of the confirmation, and the corresponding correction of the robot was given in Figure 6.3, and there are other examples of user utterances for stating the user recognition error in the Personalised Barista Datasets, such as *"That is not my name."*, *"I think you are mixing me with someone."*.

In order to fully autonomously detect users to start and end an interaction, in addition to keep eye contact with the user to improve the naturalness of the interaction, we use person and face detection modules of NAOqi[4]. The user model that refers to the identity information within MMIBN and the user preferences within the knowledge-base, is created at the first interaction of the user, which is the first time user's full name is obtained in the interaction. The user model is updated after each interaction with the corresponding user. All users are autonomously registered to the user recognition system after their first interaction, but the estimated user recognition information is only used by the *personalised barista*.

7.2.2 Speech Recognition

The interaction with a barista relies only on verbal communication in a coffee shop. Our aim is to provide an interaction similar to the real-world one, thus, we also rely mainly on

[4] `http://doc.aldebaran.com/2-5`

bi-directional verbal exchanges through automatic speech recognition (ASR) and text-to-speech (TTS). While we tried to build a representative database of phrases that may be used in a dialogue with a barista in the Barista Datasets, there may be other phrases that customers may use to express themselves. While users can adapt their interaction to the knowledge of the robot (Williams et al., 2018), this may take quite some time, whereas, our aim is to provide a natural interaction and increase the task efficiency by decreasing the time it takes to make an order. These objectives required allowing unconstrained speech (i.e., without a given a list of available sentences). In turn, we needed an online ASR that allows open-grammar (i.e., unconstrained) recognition, has a wide vocabulary and is reliable for real-world interactions with adults (Halpern et al., 2016; Ziman et al., 2018). Thus, we chose Google Cloud Speech-to-Text engine for ASR. We used the *Adapted Pepper* robot[5], which has an improved microphone system with lower noise compared to an off-the-shelf robot. In order to enable online speech recognition, we used NAOqi voice activity detection to determine the beginning and end of user speech, which is sent to Google ASR for processing, and the result is analysed by the dialogue manager, as described in the next section.

Speech recognition was optimised with a band-pass filter based on four monologues from the Personalised Barista Dataset with 12 non-native English speakers. The resulting speech recognition performance was 47% for exact match accuracy (i.e., percentage of words that match the ground truth values) and BLEU score[6] (Papineni et al., 2002) of 0.66.

While the robot is talking or processing speech, speech recognition is disabled to prevent loss of information. If the robot is listening green LEDs are lit in the robot's ears, otherwise, the light is blue. The participants in the study were requested not to speak when the robot is speaking, and have been notified of this feature.

Even in a real-world scenario, baristas can sometimes misunderstand the customer names due to the pronunciation of foreign names. This may pose a major problem for speech recognition as well, hence, we use the touchscreen interface on the robot to obtain customers' names for robustness. In addition, in order to ensure a natural level of interaction with mutual understanding, we combine non-verbal features, such as gaze (through face tracking) and body movements (i.e., animated speech feature of NAOqi).

[5]Created for MuMMER project: `http://mummer-project.eu`.

[6]BLEU score indicates how similar the candidate text is to the reference text, with values closer to 1 representing a higher quality machine translation.

7.2.3 Rule-Based Dialogue Management System

In Chapter 2, we identified that rule-based approaches that rely on a knowledge-base to extract and update information are commonly used in HRI for task-oriented dialogue (Gockley et al., 2005; Kanda et al., 2007, 2010; Giuliani et al., 2013; Kasap & Magnenat-Thalmann, 2012; Churamani et al., 2017; Williams et al., 2018; Zheng et al., 2019). Consequently, we built a rule-based dialogue management system (RBDMS). Similar to a typical task-oriented dialogue system (Tur, 2011; Young et al., 2013; Gao et al., 2019), a RBDMS is composed of four modules: natural language understanding (NLU) using template matching, (rule-based) dialogue state tracking (DST), (rule-based) dialogue policy (DP) and (rule-based) natural language generation (NLG), as shown in Figure 7.2.

As presented in the previous chapter, we allow mixed-initiative communication, in order to extract particular information from the user and allow the user to change their order and amend errors in recognition. We apply template matching and dialogue state tracking based on the phrases in the Barista and Personalised Barista Datasets, using Difflib[7] and Fuzzywuzzy[8] Python libraries. For instance, if a drink order is in the previous phrase, then template matching is applied to size order phrases because it is the next rule in line. We apply cutoff thresholds of 0.7-0.8 to phrases and 0.5 to order keywords, which were found using preliminary experiments, in other words, if the similarity score of the received phrase from the ASR to the expected phrase in the templates from the Barista Datasets is below 70% or 80% depending on the complexity of the expected phrase or below 50% for the order details, the robot asks the customer to repeat the phrase again.

Depending on whether a *generic barista* or a *personalised barista* robot will be used, the dialogue policy determines the next state (i.e., utterance type) in line. The NLG uses the phrase from the Barista Dataset for the *generic barista* or combines the phrase from the Personalised Barista Dataset with the user model information from the knowledge-base (KB) for the *personalised barista* robot.

The RBDMS was evaluated with the Barista and Personalised Barista Datasets, and the performance was 100% in each task, as expected.

NAOqi TTS is used to utter the chosen phrase back to the user, which reconnects to the voice activity detection if the user is still present. The interaction ends if the user said

[7]https://docs.python.org/3/library/difflib.html
[8]https://github.com/seatgeek/fuzzywuzzy

a farewell phrase or if a face is not detected for 30 seconds. Afterwards, the farewell phrase of the robot is triggered, the conversation is logged and the user model is updated (only for the personalisation condition). The reason we chose 30 seconds to wait for face detection is to prevent "losing" users during the interaction if they went out of the view of the robot.

7.3 Hypotheses

The primary purpose of this study is to answer our research question (RQ6), *what is the impact of personalisation in long-term human-robot interaction?* within a real-world application. In line with this research question and our main thesis, we designed a fully autonomous personalised customer-oriented service robot for a coffee shop that recognises customers and recalls their favourite orders, in order to improve user experience and increase the time efficiency of order taking. Based on the objective of our study and the findings from previous literature, we derived the following hypotheses for the study, as listed below:

H1 Personalisation will improve the user experience in comparison to a generic interaction for long-term interactions

H2 Personalisation will increase the time efficiency, by decreasing the number of turns in a dialogue

H3 Personalisation will improve the perceived level of social intelligence of the agent

Consequently, we designed three conditions for this study: (1) the *first interaction* with the robot, and the subsequent interactions (second and third interactions) with a (2) *generic barista* or a (3) *personalised barista*. The reason we separate the *first interaction* from the other conditions of the study is that the first interaction with the robot for all users will be non-personalised (i.e., generic) because the robot does not know the user. In addition, separating the initial interaction allows analysing the long-term perception in comparison to the short-term one, which may be affected by the *novelty effect*.

7.4 Experimental Procedure

We conducted a five day study in the coffee bar of an international student campus, Cité Internationale Universitaire de Paris, as shown in Figure 7.3.

(a) (b) (c)

Figure 7.3: (a) Experiment setup with *Adapted Pepper* robot, (b) image of the interaction from the external camera, (c) image of the user from the internal camera of the robot.

7.4.1 Participants

18 non-native English speakers (11 males, 7 females) within the age range of 22-47 ($M = 28.2$, $SD = 7.0$) participated in the study. The study was advertised through social media channels, posters around the campus and actively by the experimenter in the campus. In exchange for their time, the participants were given (by the experimenter) the drink and snack that they ordered from the robot. The experimenter was hidden behind a screen away from the view of the participant, while the participant interacted with the robot. The experimenter did not interfere with the experiment unless the robot was stuck at a phrase for a prolonged period of time or had an apparent connection failure, in which case, we asked if the participant would like to repeat the interaction.

In order to prevent users from being affected by the other users' interactions, and to avoid delays and queues which can cause negative perceptions, a schedule was created for the participants' attendance times.

Initially, we intended to assign 9 participants each for *generic* and *personalised barista* conditions, however, as this was a real-world long-term interaction, we could not force the users to come to the experiment in the next days. Hence, only 4 participants in the *generic robot* condition visited the coffee bar again (4 of them for the second time, and 2 of them for the third time), whereas, it was 5 participants in the *personalised barista* condition (5 of them for the second time, and 4 of them for the third time).

7.4.2 Measures

We measured the users' perception of the social intelligence of the robot through the Robotic Social Attributes Scale (RoSAS) (Carpinella et al., 2017) questionnaire, that has factors for measuring the robot's capability, knowledge, reliableness and competency. We additionally developed a questionnaire to evaluate the task performance and perceived personalisation of the robot, as presented in Appendix C. Moreover, we used open questions to allow users to freely express their perception of the robot.

In order to further analyse how the user interacts with the robot, we recorded videos of the interactions through the robot's camera and an additional hidden camera, as shown in Figure 7.3. The audio was recorded through the robot's microphones for sending the voice files to speech recognition. In addition, we used the recordings to evaluate whether speech recognition performed correctly. The participants were notified that the interactions would be recorded in the consent forms (Appendix C), with the option for the consent of sharing their videos and images for academic purposes.

7.5 Results

Due to the technical difficulties further outlined in this section, only 5 out of 18 *first interaction*, 4 out of 6 *generic barista*, 3 out of 9 *personalised barista* interactions were successful, that is, completed and the correct order delivered. Due to the low number of subsequent encounters, the resulting Bayes factors are between 0.3-3, suggesting inconclusive statistical significance between conditions (Jeffreys, 1939; Lee & Wagenmakers, 2013), thus, we interpret the implications of the trends in the results.

The results in Figure 7.4 support that a higher percentage of users received the correct order and had more complete interactions in the *generic barista* condition. However, a higher percentage of users enjoyed the interaction, looked forward to the next one, and preferred to interact with the robot as a barista in a coffee shop (in contrast to the findings by Churamani et al. (2017)) in the *personalised barista* condition in comparison to the other conditions, supporting **H1** (although not significantly). Also, the user experience has

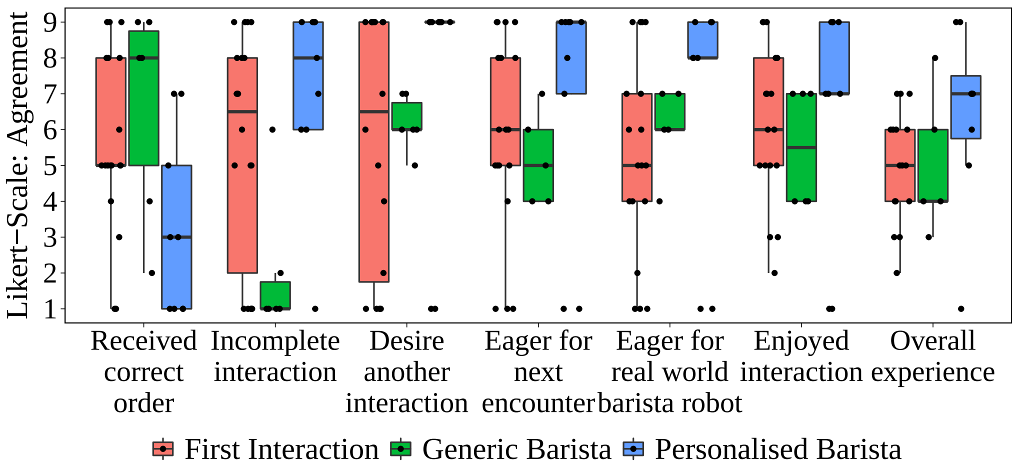


Figure 7.4: Perceived task performance and user experience from the task-specific questionnaire. The results show that while there was a lower percentage of successful interactions in the *personalised barista* condition, a higher percentage of users enjoyed the interaction and reported a more positive user experience, looked forward to the next interaction and preferred to interact with the robot as a barista in the real world.

improved in the *personalised robot* condition in comparison to the *first interaction*. The results for the additional questions for the *personalised robot* condition that evaluate the task performance of the robot were in correspondence with the actual performance of the robot, and the users were generally pleased that the robot was able to remember their previous orders (Likert Scale rating of 7-9), except for two users who experienced full speech and user recognition failures that resulted in the low scores from the participants. A participant additionally noted his delight about the personalisation of the robot in the open questions with *"I was very pleased to hear my name and my preferences"*. These findings suggest that personalisation can mitigate the negative user experience, which is a key result of conducting a real-world study.

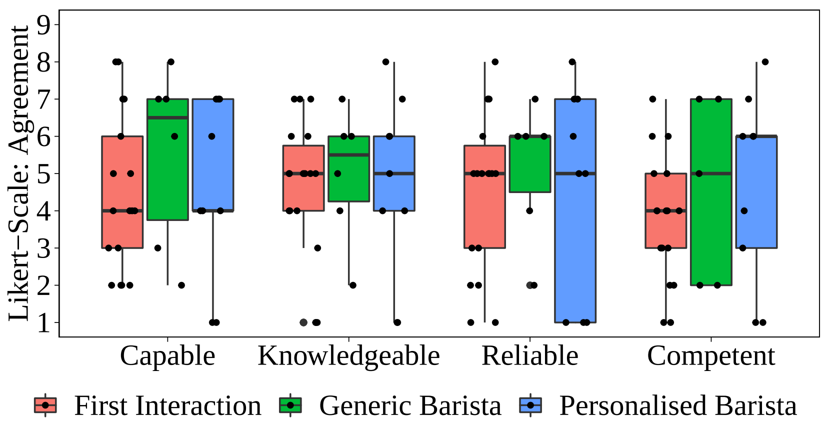


Figure 7.5: User responses for the factors of RoSAS that evaluate the user perception of the robot in terms of task performance. The *generic barista* was rated generally higher in capability and reliability in correspondence to the robot's task performance, whereas, the *personalised barista* was perceived to have a higher competency.

The RoSAS questionnaire results are in line with the perceived task performance of the robot. There were no significant differences between the conditions, as can be observed in Figure 7.5, however, the participants that interacted with the *generic barista* rated its capability and reliability generally higher, as expected, because of the higher percentage of success in the *generic barista* condition. Nonetheless, the competency ratings of the *generic barista* and the *personalised barista* are comparable in contrast to the poor performance of the *personalised barista* robot. However, due to the varying performance between the conditions, we cannot conclude support for **H3**.

The primary cause of failure was speech recognition, which often caused frustration (Figure 7.6), because users had to repeat phrases several times, as evident in Figure 7.7 and 7.8. Overall, only 30.2% of the utterances were processed (i.e., understood), and from those only 55.4% of the processed utterances matched correctly to the user utterances (i.e., per-response accuracy), 69.4% of the words were correctly recognised (i.e., exact match score) and the BLEU score was 0.49. The underlying reasons for failures in speech recognition are:

- Foreign accents of non-native speakers, which caused the ASR to match certain order words incorrectly (e.g., "MIT" for *"a mint tea"*, "black deer" for *"a black tea"*, "Arch" for *"large"*, "eliminated" for *"a lemonade"*)
- Latency due to connection problems, which caused the customers to repeat their phrases because the robot was not responding
- Quietly speaking users
- User's distance from the robot, which decreased the quality of the audio. However, users mentioned in the open questions that they were uncomfortable in staying at a close distance to the robot

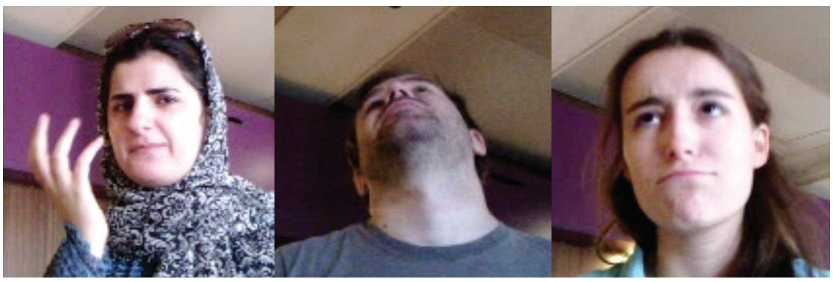


Figure 7.6: Examples of negative reactions from the users when the robot does not understand the order due to speech recognition or rule-based structure of the dialogue manager.

- Accuracy of the ASR on the audio obtained from the robot's microphones
- Delay between voice activity detection and recording of the voice, which caused *"no"* to not be recognised due to the corrupted phonemes in the voice recordings processed by the ASR

Due to the speech recognition errors, incorrect or incomplete phrases were delivered to the rule-based dialogue management system (RBDMS), which severely affected its performance. Repetitions of users also worsened the performance of the RBDMS because

Profile:	True *(Known)*, 3 *(ID)*, X *(Name)*
Preference:	Black tea, medium, chocolate cookie
Customer:	Hello
Speech recognition:	*Hello*
Personalised barista:	Hello, X! Would you like to have a medium black tea and a chocolate cookie again?
Customer:	No, thanks! *[speech not detected]* No, thank you!
Speech recognition:	*[Unable to process speech]*
Personalised barista:	Sorry, I am having trouble understanding you. Could you speak a little louder?
Customer:	No!
Speech recognition:	*No*
Personalised barista:	Ah, I must be mixing orders! Which drink I get for you, X?
Customer:	A coffee.
Speech recognition:	*Purses*
Personalised barista:	Which size for the cup?
Customer:	Medium. *[speech not detected]* Medium!
Speech recognition:	*[Unable to process speech]*
Personalised barista:	I am having trouble hearing you. Could you repeat please?
Customer:	Medium size. *[speech not detected]* Medium.
Speech recognition:	*[Unable to process speech]*
Personalised barista:	Could you repeat please?
Customer:	Medium.
Speech recognition:	*Medium*
Personalised barista:	Can I offer you anything to go with your drink?
Customer:	A sponge cake.
Speech recognition:	*A sponge cage*
Personalised barista:	So that would be a medium coffee and a sponge cake. Did I get that right?
Customer:	Yes! *[speech not detected]* Yes!
Speech recognition:	*[Delayed] Yes*
Personalised barista:	Could you repeat please?
Personalised barista:	We will get that ready for you at the next counter, X.
Customer:	Thank you!
Speech recognition:	*Thank you*
Personalised barista:	Have a great day!

Figure 7.7: Example of a **successful** dialogue from the study, where the *personalised barista* recognised the customer and recalled their preference correctly, but the customer wanted to order something else, and the robot was correctly able to update the order, corresponding to Task 4 in the Personalised Barista Dataset (Figure 6.5). The speech detection and recognition errors made the interaction tedious and longer than desired (2.5 minutes). Customer's name replaced with X for anonymity.

Profile:	False *(New)*, 0 *(ID)*
Customer:	Hi! *[speech not detected]* Hi!
Speech recognition:	*Hi*
Generic barista:	Hello! Which drink can I get for you?
Customer:	Erm, please lemonade. *[speech not detected]* A lemonade? *[incorrect pronunciation of lemonade]*
Speech recognition:	*[Delayed] Please*
Generic barista:	Which size for the cup?
Customer:	A medium *[pause]* please. *[wait for robot]* Medium? *[speech not detected]* Please medium. Medium. *[pause]* Medium. *[incorrect pronunciation of medium]*
Speech recognition:	*Please medium medium medium*
Generic barista:	Can I offer you anything to go with your drink?
Customer:	Erm, yes. Please a chocolate cookie. *[pause]* Chocolate cookie. *[pause]* Chocolate cookie. *[incorrect pronunciation of chocolate]*
Speech recognition:	*[Unable to process speech]*
Generic barista:	I am sorry. Could you repeat please?
Customer:	Erm, a pound cake. *[incorrect pronunciation of pound]*
Speech recognition:	*[Unable to process speech]*
Generic barista:	I am sorry. Could you repeat please?
Customer:	Erm, a chocolate biscuit. *[pause]* A chocolate biscuit. *[long pause]* A chocolate biscuit? *[incorrect pronunciation of chocolate]*
Speech recognition:	*[Unable to process speech]*
Generic barista:	Sorry, I did not get that!
Customer:	*[inaudible voice]*
Speech recognition:	*[Unable to process speech]*
Generic barista:	Could you speak a little louder?
Customer:	Yes, sure. A chocolate cookie! *[speech not detected]* A chocolate cookie. *[speech not detected]* A chocolate cookie.
Speech recognition:	*[Delayed] Is chocolate cookie*
Generic barista:	Could you speak a little louder?
Generic barista:	Your order is a medium coffee and a chocolate cookie. Is that right?
Customer:	Yes yes yes yes yes. Thank you.
Speech recognition:	*[Unable to process speech]*
Generic barista:	Sorry, I am having trouble understanding you.
Customer:	What I should do?
Speech recognition:	*[Unable to process speech]*
Generic barista:	Could you repeat please?
Customer:	Yes
Speech recognition:	*Yes*
Generic barista:	May I take your name, please? Please use the tablet for that.
Customer:	Okay! *[enters name on tablet]*
Generic barista:	You can pick up the order at the next counter, Y.
Customer:	Okay, thank you!
Speech recognition:	*Ok, thank you*
Generic barista:	I wish you an amazing day!

Figure 7.8: Example of an **unsuccessful** dialogue from the study, where the *generic barista* delivered the incorrect order (coffee instead of lemonade) due to errors in speech recognition. The interaction corresponds to Task 6 in the Barista Dataset (Figure 6.1). The speech detection and recognition errors made the interaction frustrating and longer than desired (3.07 minutes), which caused the customer to accept the wrong order. Customer's name is replaced with Y for anonymity.

the delivered phrase from the ASR included these repetitions (e.g., *"Please medium medium medium"* in Figure 7.8), which reduced the similarity score to the template phrases. Incomplete phrases also caused the RBDMS to misinterpret the user's response. For instance, when the *personalised barista* robot offered the user their previous preference, the ASR returned *"Thank you"* instead of *"May I have a coffee, please? Thank you!"*, hence, the RBDMS processed it as a confirmation instead of a change in the order. In addition, the fixed order structure of RBDMS, prevented users to change their order after confirmation. The system failed to understand when the user ordered the items in a combined sentence (e.g., *"I would like a small coffee and chocolate cookie please."*) or switched the order of items. Moreover, the RBDMS was unable to understand *"What is my order?"* and *"What snack do you have?"*, because these phrases were not in the Barista Datasets. The reason is that the user was presented with a menu behind the robot (as shown in Figure 7.3), hence, we assumed that the user would not ask the robot for the information.

Blurry images due to user movement, as well as bad lighting conditions in the cafeteria, reduced the performance of the user recognition system. In addition, the users did not realise when the robot incorrectly identified them, thus, MMIBN and RBDMS updated the wrong user, thereby, causing the *personalised barista* to have a worse success rate. Consequently, 11 users were enrolled in the user recognition system and the knowledge-base, instead of the 18 recruited participants. Only in 4 of the *personalised barista* interactions, the participant was identified correctly, and the previous orders of the users were remembered, however, 2 of these interactions were incomplete. In these interactions, because the speech recognition and RBDMS worked sufficiently correctly, the number of dialogue turns has decreased to 3 turns instead of 7 in support of **H2**, however, due to the problems encountered during the study and the several repetitions necessary to make an order, we cannot conclude that the personalisation increased the time efficiency, as presented in Figure 7.9.

The video analysis showed that most of the users who learn to give short and specific orders and speak loud and clear got their order right, and this interaction style was retained in the subsequent interactions. In addition, the users changed their phrasing during repetitions, in order for the robot to understand them, which is similar to the results in (Williams et al., 2018).

Profile:	True *(Known)*, 1 *(ID)*, Z *(Name)*
Preference:	Coffee, small, chocolate cookie
Customer:	Hi!
Speech recognition:	*Hi*
Personalised barista:	Hello, Z! Would you like to have a small coffee and a chocolate cookie again?
Customer:	Yes, please! *[speech not detected]* Yes, please!
Speech recognition:	*Yes please*
Personalised barista:	You can pick up your order at the next counter, Z.
Customer:	Thanks!
Speech recognition:	*[Unable to process speech]*
Personalised barista:	Sorry, I did not get that!
Customer:	*[walks out, then walks back in front of the robot]* Thanks!
Speech recognition:	*Thank you*
Personalised barista:	Enjoy your day!

Figure 7.9: Example of a **successful** dialogue from the study, where the *personalised barista* recognised the customer and recalled their preference correctly, and the customer confirmed the suggestion, corresponding to Task 1 in the Personalised Barista Dataset (Figure 6.2). The interaction took 1.17 minutes. Customer's name is replaced with Z for anonymity.

7.6 Discussion

The technical difficulties in this study caused the participants to repeat their phrases several times, change their wording, and even accept wrong orders, but these are unlikely to happen in the real world when the customers are in a hurry. For deploying robots to the real world, we need solutions that are reliable and can recover from failures.

Our results showed that ASR is not accurate enough for real-world applications, hence, a touchscreen interface for text or image-based interaction can be used. However, such methods decrease the naturalness of the interaction. Thus, it is preferable to improve the accuracy of ASR, by constraining grammar of ASR (Kennedy et al., 2017), ensuring a reliable internet connection or using an onboard ASR, and using high-quality microphones. Low ASR accuracy in foreign accents can be overcome by personalising the interaction with the user's native language, which was even requested by some of our participants in the open questions.

We should also account for user errors by designing systems that are flexible and robust. For example, explicit confirmation of the identity before the order can overcome errors in MMIBN. Moreover, a neural network with a long-term memory could be more suitable than a rule-based dialogue management system for reverting changes in the state of the dialogue (Bordes et al., 2016) and dealing with the combination of order items. We will evaluate the potential of such systems based on the Barista Datasets in the next chapter.

Nevertheless, these failures enabled us to observe the high positive impact of personalisation on the negative user experience, which showed the importance of evaluating technologies outside of controlled environments and studying how people respond to failures.

7.7 Summary

This chapter described the fully autonomous generic and personalised Barista Robot architectures built upon the Barista and Personalised Barista Datasets. We conducted the first real-world study that explores fully autonomous personalisation in dialogue for long-term human-robot interactions in an international student campus for five days with non-native English speakers. We evaluated how personalisation affects the user experience, task efficiency and user perception of social intelligence. We experienced several challenges due to speech recognition failures, arising from the foreign accent of non-native speakers, latency due to connection problems, quietly speaking users, user's distance from the robot, low accuracy due to the robot's microphones, and the delay between voice activity detection and recording. Nonetheless, these failures showed that personalisation can overcome a negative user experience.

Chapter 8

Towards Using Data-Driven Approaches in Personalised Long-Term Interactions

Key points:

- State-of-the-art retrieval-based and generative data-driven dialogue models are evaluated on the Barista Datasets to evaluate their potential in generic and personalised long-term interactions.
- The (generative) vanilla Sequence-to-Sequence model achieves the best and near-perfect per-response accuracy in generic task-oriented dialogue.
- The (retrieval-based) vanilla End-to-End Memory Network achieves best accuracy in personalised task-oriented dialogue, but it does not perform sufficiently well to be deployed in personalised long-term real-world interactions.
- Most models cannot learn new customer names or new order items, which decreases the performance in continual (or lifelong) learning (i.e., incremental learning with adaptation) for personalised task-oriented dialogue. Thus, user preferences information that simulates a knowledge-base extraction only slightly improves the accuracy, and the separate profile memory architectures do not markedly improve the performance.
- Time order within the conversation context is important for dialogue state tracking and detecting changes in the user orders.

- Generative models learn sentence grammar and structure well and perform best in dialogue state tracking and few-shot learning.
- High sample size improves model accuracy for generic task-oriented dialogue, however, it has a varying effect on personalisation depending on the model and task.
- Memory Network and Split Memory Network have the lowest time complexity, but all models are suitable for real-time interaction.

Parts of the work are under review at the Frontiers in Robotics and AI journal. The Barista Datasets and the adapted data-driven dialogue models described in this chapter are available online[1] for academic use based on the license terms.

[1] `https://github.com/birfan/BaristaDatasets`

8.1 Motivation

While previous research in Human-Robot Interaction (HRI) pre-dominantly uses rule-based architectures for communication, the previous chapter showed that these approaches are not robust or flexible enough for long-term interactions in the real world. In contrast, data-driven approaches map the user input to the agent output directly, hence, provide more flexibility with the variations in user utterances without requiring any set of rules. However, previous research in data-driven approaches focused on the current dialogue exchange with a single user and does not build up a memory over long-term conversation with different users (Dodge et al., 2015), whereas this is essential for long-term interactions and personalisation in HRI. The models need to learn users and their preferences incrementally and recall previous interactions with users to adapt and personalise the interactions, which is a continual (or lifelong) learning problem. Continual learning in machine learning refers to learning information continuously, incrementally and adaptively, on top of previously learned information. In other words, it refers to incremental learning with adaptation, but it covers both batch and online (sequential) learning. In addition, it is desirable for the model to learn users preferences from a few samples of interactions (i.e., *few-shot learning*). These are challenging problems in machine learning (Parisi et al., 2019; Triantafillou et al., 2017), while they are trivial for rule-based approaches. Correspondingly, in this chapter, we explore the state-of-the-art data-driven dialogue models, namely the variants of Memory Networks (Bordes et al., 2016; Joshi et al., 2017; Zhang et al., 2018), Supervised Embeddings (Bordes et al., 2016; Joshi et al., 2017), and Sequence-to-Sequence (Sutskever et al., 2014) to evaluate their potential in generic and personalised long-term interactions, based on the Barista Datasets. These approaches are strong baselines in other domains of personalisation based on single interactions, such as adapting to general user attributes in task-oriented dialogue (Joshi et al., 2017) or "person"alising an open-domain dialogue by maintaining a given personality (Zhang et al., 2018).

8.2 State-of-the-Art Data-Driven Approaches

In Section 2.3, we outlined the state-of-the-art data-driven approaches for task-oriented interactions and personalisation of the interaction. These approaches can be categorised based on the response generation: retrieval-based and generative models. Retrieval-based models choose a response from a list of phrases (called the *candidate set*), hence, the response can be syntactically correct, but these models may fail to respond appropriately to novel questions. Generative models form a response iteratively (i.e., word-by-word), thus, novel responses can be generated. However, it is challenging to learn the grammar and the structure of the sentence, in addition to learning the correct responses, hence, they are prone to grammatical errors.

As previously described in Section 6, while there are no publicly available datasets for user-specific personalisation in dialogue, there are two publicly available datasets for other applications of personalisation in dialogue: Personalized bAbI dialog (Joshi et al., 2017) and Persona-Chat (Zhang et al., 2018).

The Personalized bAbI dialog is a simulated text-based dataset for personalising conversation style and recommendations in task-oriented dialogue according to general user attributes, such as gender, age, favourite food item and dietary preference. Retrieval-based models, namely Supervised Embeddings (Dodge et al., 2015; Bordes et al., 2016), End-to-End Memory Networks (MemN2N) (Sukhbaatar et al., 2015) and Split Memory architecture (Joshi et al., 2017) were evaluated on this dataset. The best performing model for the complete dialogue task was found to be Split Memory.

Persona-Chat is a crowd-sourced text-based dataset that contains conversation based on assigned personalities, which was created to increase user engagement and improve consistency of the agent in open-domain dialogue. Several retrieval-based models, such as Nearest Neighbor Information Retrieval (Sordoni et al., 2015), Ranking Profile Memory Network, Key-Value Profile Memory Network based on (Miller et al., 2016) and a supervised embedding model, StarSpace (Wu et al., 2017), along with generative models, such as Generative Profile Memory Network and Sequence-to-Sequence (Seq2Seq) (Sutskever et al., 2014), were evaluated on the Persona-Chat dataset. The best performing model was found to be the Key-Value Profile Memory Network, in terms of automated metrics and human evaluation (for fluency, engagingness and consistency).

We evaluate the baselines from the Personalized bAbI dialog and the best performing baseline from the Persona-Chat dataset, in addition to the generative baselines, to get an insight on their potential for user-specific personalisation in long-term interactions. Instead of the user (customer) attributes (e.g., gender, age, favourite food) in the Personalized bAbI dialog or the personality determining sentences in the Persona-Chat dataset, we use user identity information (i.e., whether user is enrolled (True or False), user's ID number and name) in the Personalised Barista Dataset, and the user identity information along with the user preferences (i.e., most preferred drink, size and snack) in the Personalised Barista with Preferences Information Dataset, which we call *user profile*, similar to earlier work. In this section, we briefly describe these approaches and their performance in the previous literature.

8.2.1 Supervised Embeddings

Word embedding models are generally used for learning unsupervised embeddings over unlabeled datasets such as in Word2Vec (Mikolov et al., 2013). However, they are strong baselines for predicting the response given the previous conversation in both open-domain and (bAbI) task-oriented dialogue (Dodge et al., 2015; Bordes et al., 2016).

One common approach in the previous literature (Bai et al., 2009; Dodge et al., 2015; Bordes et al., 2016; Joshi et al., 2017) sums the bags-of-embeddings of the input and the target, and then scores the inner product of the candidate responses against the input (e.g., user response). The embeddings are trained with stochastic gradient descent (SGD) using a margin ranking loss to ensure that the correct targets are ranked higher than any other targets (i.e., *negative candidates*). This approach corresponds to a Memory Network with no attention over memory (Dodge et al., 2015) and a classical information retrieval model where the matching function is learnt (Bordes et al., 2016).

We use the implementation[2] of Supervised Embeddings used in (Joshi et al., 2017). Similar to that work, we do not handcraft any special embeddings for the user profile, and treat it as a turn in the dialogue. The Adam optimiser (Kingma & Ba, 2014), which is an extension of SGD, is used for minimising the model loss.

Inputs and outputs are represented with binary bag-of-words vectors that have a size equal to that of the vocabulary. Each item (i.e., words and punctuations) in the input text is represented with 1 corresponding to its order in the vocabulary and the remaining vector is filled with 0's. For example, for the user utterance "Hi there!", the vector representation will be [1,0,0,0,1,...,0,1,0], because *Hi* is the 5th item, *there* is the 1st item and *!* is the 399th item for a vocabulary size of 400. This vector representation is then converted to

[2] `https://github.com/chaitjo/personalized-dialog`

an embedding of a specified size (e.g., 32). Note that due to the structure of this method, the order of the words within the input (e.g., user utterance, bot response or conversation context) is not preserved, that is, the embedding only contains the list of words instead of a sentence. In addition, repeating words would also be lost in the embedding. Hence, we believe that this model is not suitable for dialogue with the implementation used in (Joshi et al., 2017). Moreover, it performed poorly in Personalized bAbI dialog dataset. However, we selected it as a baseline to determine its strong and weak points for user-specific personalisation in comparison to generic dialogue.

8.2.2 Sequence-to-Sequence

Sutskever et al. (2014) introduced the Sequence-to-Sequence (Seq2Seq) model, which is a generative data-driven model for language translation. The model uses a long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997; Graves, 2013) to read the input sequence (i.e., as the encoder) to obtain a fixed-dimensional vector representation, and another LSTM to extract the output sequence from that vector (i.e., as the decoder). The order of the words in the sentence is reversed for the input sequence, which improves the performance due to the closer proximity of the input to the output while using SGD for optimising the loss. The original paper showed that the model performs well for language translation (English to French). The model was shown to differentiate between the different ordering of the words, e.g., "John admires Mary" and "Mary admires John" were in separate clusters based on their meaning. In addition, the clusters of the active voice formulation of a sentence (e.g., "She gave me a card in the garden.") and the passive voice formulation (e.g., "I was given a card by her in the garden.") were close, meaning that the model is fairly insensitive to the voice. These properties make the model suitable for dialogue. Correspondingly, it was shown to be a strong baseline in task-oriented and open-domain dialogue (Vinyals & Le, 2015; Sordoni et al., 2015; Li et al., 2016a,b; Zhang et al., 2018).

We use the implementation from ParlAI[3] that was used at ConvAI2[4] challenge (Dinan et al., 2019) with the Persona-Chat dataset. The final hidden state is fed into the decoder as the initial state. For each time step, the decoder produces the probability of a word occurring in that place via a softmax of the weighted hidden state. An example of the

[3]`https://github.com/facebookresearch/ParlAI/tree/master/projects/convai2/baselines/seq2seq`

[4]`http://convai.io`

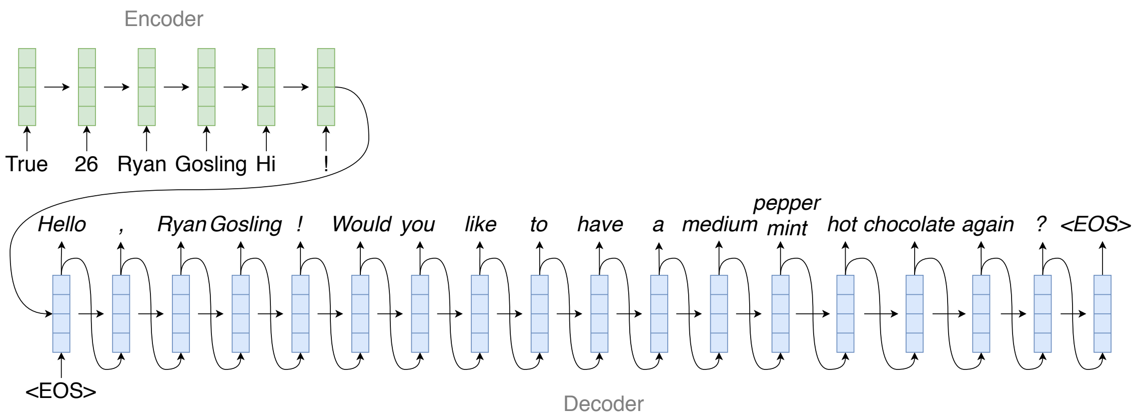


Figure 8.1: An illustrative example for encoding (of the user response) and decoding (of the bot response) within the Sequence-to-Sequence (Seq2Seq) model. <EOS> token is used to determine the end of the sentence within the fixed embedding structure. The user profile information (i.e., user identity and preferences) is concatenated to the beginning of the input. The example corresponds to the beginning of the interaction in Figure 6.5.

encoder-decoder structure of the Seq2Seq is given in Figure 8.1. In (Zhang et al., 2018), GloVe (Pennington et al., 2014) word embeddings were used. However, we found that using randomly initialised embeddings that are trained with the data performed better than pre-trained GloVe embeddings in Personalised Barista Datasets (e.g., achieving 60.58% accuracy in comparison to 41.75% in task 8 for 1,000 dialogues dataset), whereas, GloVe performed slightly better in the Barista Dataset (99.85% with random embeddings and 99.94% with GloVe). Since the difference is considerably higher in the Personalised Barista Datasets, we chose random embeddings. In addition, in contrast to (Zhang et al., 2018), using previous sentences in the dialogue (i.e., conversation *context*) performs better in Personalised Barista Datasets. This is expected because the order is made in several dialogue turns, hence, the model should keep the previous history to gather all of the items in the order, whereas in open-domain dialogue the dependency to the context is less. Similar to (Zhang et al., 2018), we train the model with negative log likelihood and we prepend the user profile to the input sequence (i.e., concatenated to the beginning of the input).

8.2.3 Memory Networks

Humans focus on salient parts of information to efficiently accomplish a task at hand or for recalling the key aspects of an event. Similarly, *attention* in deep learning focuses on particular elements of a task to respond to queries, through non-uniformly weighting parts of the input to optimise the learning and recall processes. Attention mechanisms

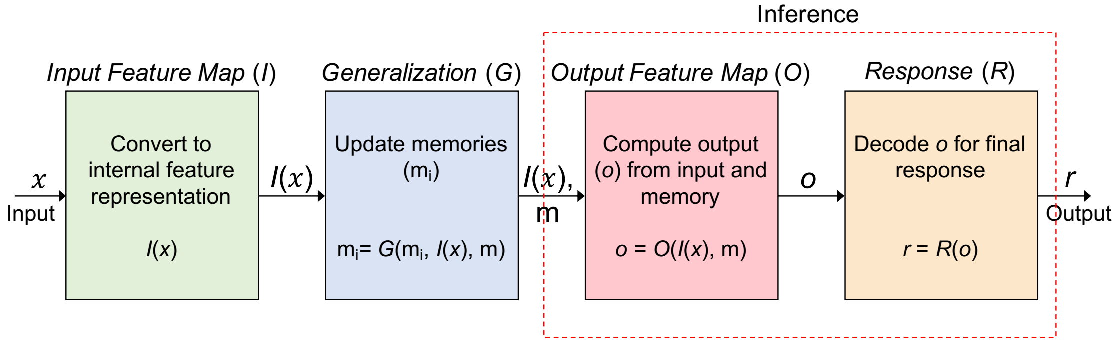


Figure 8.2: Components of the Memory Network. The mapping and scoring functions within the components vary depending on the implementation.

are especially important for personalisation of the dialogue in long-term HRI, such that the responses of the robot can be selected more efficiently and more correctly given the expanding volume of data over time.

A Memory Network (Weston et al., 2014) is an attention-based model with a long-term memory, which was initially designed for question answering (QA). The initial model required full supervision, that is, during training, the correct answer to the question had to be labelled, along with (two) *supporting memories* which are sentences that carry additional information to answer the question. The system is composed of *memory*, *input feature map*, *generalization*, *output feature map*, and *response*, as shown in Figure 8.2. The *memory* stores the representations of information (e.g., character, word, sentence, image, or audio signal) converted by the *input feature map*, e.g., parsing. The *generalization* component writes into and updates memory, which can be organised into categories or topics by hashing or clustering. The *output feature map* scores the input and the memory to find the highest scoring candidate(s) (e.g., through hard max of all scores), and the *response* module decodes the output to produce the response.

The relative time of the events (e.g., sentences) is encoded into the memory, which allows keeping the memory up-to-date and maintaining to-and-fro relations between events. Memory Networks can discover simple linguistic patterns based on verbal forms, such as (X, took, Y) for “Alice took the teacup.”, hence, it can generalise the meaning of their instantiations for previously unseen vocabulary.

In general, the model is flexible and allows using different machine learning architectures for its components, such as decision trees, Support Vector Machine (SVM), recurrent neural networks (RNN), long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997), and it can be used as a generative or a retrieval-based approach.

End-to-End Memory Networks (MemN2N) (Sukhbaatar et al., 2015) removed the need for supporting memories. The input (i.e., *query*) is converted to an internal state with an embedding matrix and the previous input sentences (called *context* or *conversation history*) is converted to a memory representation using another embedding matrix. The match between the input embedding state and each memory is computed via an inner product followed by a softmax (instead of the hard max in the original Memory Network). The output is a weighted sum of the resulting probability vector over the inputs multiplied by another embedding matrix representation of the *context*. Due to the smooth function between input and output, the gradients can be computed that allows training the system end-to-end with backpropagation. The final prediction of the bot response is found through the weighted sum of the output and the input embedding state with a softmax applied over it. They also introduced *hops*, which processes the sum of the input query and the output of the network in multiple layers, enforcing the network to increase its attention on the correct supporting sentences. For dialogue, the *query* corresponds to the last user utterance, and the *context* consists of the previous utterances of the user and the responses of the bot. The dialogue examples from the *recognition error* task (2) of the Personalised Barista (PB2, Table E.2 in Appendix E.2) and Personalised Barista with Preferences Information (PBPI2) Datasets (Table 8.1) show the attention weights on the conversation context in varying hops.

MemN2N outperformed Supervised Embeddings and information retrieval approaches, such as the Term Frequency-Inverse Document Frequency (TF-IDF) and the Nearest Neighbor in task-oriented dialogue (Bordes et al., 2016). In addition, the vanilla model outperformed Supervised Embeddings in the Personalized bAbI dialog.

Similar to the Supervised Embeddings, we use the implementation[2] used in (Joshi et al., 2017), which is a retrieval-based model for vanilla End-to-End Memory Networks. Similar to that work, we do not handcraft any special embeddings for the user profile, and treat it as a turn in the dialogue. The embedding vectors for the inputs contain indices of the words in the vocabulary (i.e., the numbers denoting the order of the words in the vocabulary) and have 0 if the word is not in the vocabulary. The embeddings and memory have a fixed size, hence, the beginning of the conversation context may be cut off. The answer array is returned as a one-hot encoding. The Adam optimiser is used for minimising the cross entropy loss.

Table 8.1: A dialogue example from the *recognition error* task (2) of the Personalised Barista with Preferences Information Dataset (PBPI2) shows the attention weights in the End-to-End Memory Networks (MemN2N) model for varying hops. In contrast to the Personalised Barista Dataset (Table E.2 in Appendix E.2), the model focuses on the customer preferences information to predict the correct response, especially with increasing hops. Zero attention weight signifies a very small value ($< 10^{-5}$).

Time	Speaker	Conversation Context	Hop1	Hop2	Hop3
1		True , 182 , Tom Welling , large , latte	0.029	0	0
2	Customer	Hey there !	0.115	0.00004	0
3	Barista	Hello , Tom Welling ! Would you like to have a large latte again ?	0.047	0	0
4	Customer	That is not my name .	0.067	0.0002	0
5	Barista	I am terribly sorry ! May I take your name ? api_call getCustomerName	0.041	0.0002	0
6	Customer	It is Anne Hathaway .	0.334	0.04	0
7	Barista	Let me see if I have any previous records of you , Anne .	0.258	0.023	0
8		True , 23 , Anne Hathaway , small , mocha , blueberry muffin	0.108	0.937	1
Customer Input		Okay .			
Correct Response		I thought you looked familiar , **Anne** ! Would you like a **small mocha** and a **blueberry muffin** again ?			
Predicted Response		I thought you looked familiar , **Anne** ! Would you like a **small mocha** and a **blueberry muffin** again ?			

8.2.4 Split Memory Network

The Split Memory (Joshi et al., 2017) architecture combines two MemN2Ns at each layer: one for conversation context (i.e., dialogue history) and the other for the user profile attributes to ensure that attention is paid to the user's profile. The outputs from both MemN2Ns are summed element-wise to get the final response of the bot for each conversation turn. For multiple hops, each MemN2N separately processes the output in multiple layers, and then the resulting outputs are summed.

Split Memory outperformed Supervised Embeddings in the Personalized bAbI dialog dataset in all tasks, and outperformed MemN2N in recommending the correct restaurant and conducting a full dialogue, however, it performed worse for responding to user queries and when the user requested changes (e.g., requesting a different type of cuisine than the previously requested one), suggesting that the simpler MemN2N model is more suitable for tasks which do not require compositional reasoning over various entries in the memory. In contrast to MemN2N, their results showed that performing multiple hops might perform worse when there is more than one aspect (e.g., favourite food and dietary preference) or memory event to focus on, as evident in Table 8.2.

Table 8.2: A dialogue example from the *recognition error* task (2) of the Personalised Barista with Preferences Information Dataset (PBPI2) shows the attention weights in the Split Memory model for varying hops. Split Memory allows focusing attention separately on the user profile (i.e., the customer's identity and most preferred order), in addition to the last bot response (containing the customer's name to be used in the response), which reinforces dialogue state tracking and predicting the correct response. Preferences information helps choose the correct items in the suggestion, which decreases the risk of mixing customers (and preferences), in contrast to the Personalised Barista Dataset (Table E.3 in Appendix E.2). Hops facilitate focusing attention to relevant inputs, however, it can decrease the performance when there are multiple target items (for preference suggestion or order confirmation), as evident in Hop 3. Zero attention weight signifies a very small value ($< 10^{-5}$).

		Profile	Hop1	Hop2	Hop3
		True	0.077	0.012	0.00002
		23	0.087	0.033	0.031
		Anne Hathaway	0.24	0.412	0.777
		small	0.087	0.004	0.00002
		mocha	0.244	0.297	0.18
		blueberry muffin	0.265	0.241	0.012
Time	**Speaker**	**Conversation Context**			
1	Customer	Hey there !	0.145	0	0
2	Barista	Hello , Tom Welling ! Would you like to have a large latte again ?	0.06	0	0
3	Customer	That is not my name .	0.151	0.0002	0
4	Barista	I am terribly sorry ! May I take your name ? api_call getCustomerName	0.055	0.006	0
5	Customer	It is Anne Hathaway .	0.251	0.00009	0
6	Barista	Let me see if I have any previous records of you , Anne .	0.339	0.994	1
Customer Input		Okay .			
Correct Response		I thought you looked familiar , **Anne** ! Would you like a **small mocha** and a **blueberry muffin** again ?			
Predicted Response		I thought you looked familiar , **Anne** ! Would you like a **small mocha** and a **blueberry muffin** again ?			

We use the implementation[2] used in (Joshi et al., 2017), which is a retrieval-based model. The embedding and memory structures are the same as MemN2N. The Adam optimiser is used for minimising the cross entropy loss. The profile attributes are added as separate entries in the memory before the start of the dialogue. In contrast to (Joshi et al., 2017), in our datasets, the profile can be updated during the conversation, such as for registering a new user or due to a recognition error, hence, in those cases, we overwrite the profile memory with the new profile information. Moreover, we found that the implementation of the Split Memory[2] contained an error, which causes the context memory to have only the most recent utterance (i.e., the last bot utterance), instead of the full dialogue history, in contrast to the reported results in (Joshi et al., 2017). This decreases the performance accuracy of the model in the majority of the tasks. For instance, for the task 8 of the Personalised Barista Dataset (PB8) with 1,000 dialogues, the model has 66.95% accuracy with the context, whereas 64.78% without the context; for the task 8 of the Personalised Barista Dataset with Preferences Information (PBPI8) with 1,000 dialogues, the model has 71.65% with the context, whereas 68.71% without the context. Hence, we modified the code to include all the conversation context, and we report the corresponding results. Split Memory is equivalent to the MemN2N without the profile information, thus, this method is not evaluated with the Barista Dataset.

8.2.5 Key-Value Profile Memory Network

Key-Value Memory Network (Miller et al., 2016) is an extension of retrieval-based MemN2N, which stores facts in key-value structured memory slots before reasoning on them to predict an answer. The keys are used to address (lookup) relevant memories concerning the input, and the corresponding values are returned. This structure allows end-to-end training with standard backpropagation via SGD. The input query can be used to pre-select a small subset of the memory, where the key shares at least one word with the input. The subset is assigned a relevance probability by comparing the question to each key. Subsequently, the output is found by reading the values of the memories by taking their weighted sum using the assigned probabilities. This key addressing and value reading can be done repeatedly with hops to focus on and retrieve more pertinent information in subsequent accesses. If the key and value are set to be the same for all memories, the model becomes equivalent to the MemN2N. However, Key-Value Memory Network outperformed MemN2N in QA in various datasets (Miller et al., 2016).

Zhang et al. (2018) applied Key-Value Memory Network to dialogue by using dialogue history from the *training* set as the keys, and the values as the next dialogue utterances, such as the user response, which allows the model to have a memory of past dialogues that can be used to predict for the current conversation. Profile attributes are separately used to perform attention to find the relevant lines from the profile to combine with the input, which is used to predict the next utterance, hence, it is called Key-Value Profile Memory Network, but we will here on refer to it as Key-Value. The next utterance is found by computing the cosine similarity of the input to the profile attributes and applying softmax, taking the weighted sum with profile sentences and summing with the input query. The candidate set responses are ranked according to their similarity for this value to determine the bot response. For multiple hops, this value is used to attend over the keys and output a weighted sum of values as before, which is again used with the candidate set to predict the next utterance. Key-Value outperformed Seq2Seq and Generative Profile Memory Network in both automated metrics and human evaluation (in terms of fluency, engagingness and consistency).

We use the implementation[5] that was used at the ConvAI2 challenge. Similar to the Split Memory, we overwrite the user profile attributes if it is updated during the conversation. We only evaluate the 1-hop model, similar to (Zhang et al., 2018). Similar to (Zhang et al., 2018), but in contrast to the other methods, we only keep the last bot utterance and the corresponding last user response in the dialogue context. This was reported to perform better in the implementation, and our results in the majority of the tasks also support this: using all dialogue context provides 55.68%, 19.2%, 23.15%, whereas, using only last bot-user utterance pair provides 70.98%, 40.09%, 42.95% in the task 7 of the Barista Dataset (B7), task 8 of the Personalised Barista Dataset (PB8), and task 8 of the Personalised Barista with Preferences Information Dataset (PBPI8), respectively. We have not evaluated other methods in this way, either because the structure was not implemented or a difference was not reported in the original work.

Because the set of (key-value) pairs is large in Key-Value, the training of the model is very slow. In (Zhang et al., 2018), they trained the Ranking Profile Memory Network (equivalent to the first layer of Key-Value without keys) and used the set of weights from that model and applied Key-Value architecture at test time instead. However, the authors noted that training the model directly would give better results, hence, we trained the model directly on the datasets.

8.2.6 Generative Profile Memory Network

Generative Profile Memory Network (Zhang et al., 2018), here on referred to as Profile Memory, extends the Seq2Seq model by encoding each of the profile entries as individual memory representations in a Memory Network. Words are weighted with their inverse

[5] https://github.com/facebookresearch/ParlAI/tree/master/projects/convai2/baselines/kvmemnn

term frequency[6]. The decoder attends over both the encoded profile entries and the context as illustrated in Figure 8.3.

Profile Memory was shown to outperform Seq2Seq for automated metrics in the Persona-Chat dataset, however, it performed considerably worse than Key-Value (e.g., next-utterance classification (Lowe et al., 2015) score of 0.125 in comparison to 0.511 by Key-Value).

We use the implementation from ParlAI[7] that was used in (Zhang et al., 2018). Similar to the work in (Zhang et al., 2018), we use GloVe embeddings (Pennington et al., 2014) and train with the Adam optimiser, however, we also use the previous context (i.e., history of the user and bot utterances) in the dialogue, because we found that this improves the accuracy as in the Seq2Seq model in the majority of the tasks (e.g., for 1,000 dialogues datasets, 57.82% accuracy with context, 55.22% without context for the PB8; 58.98% with

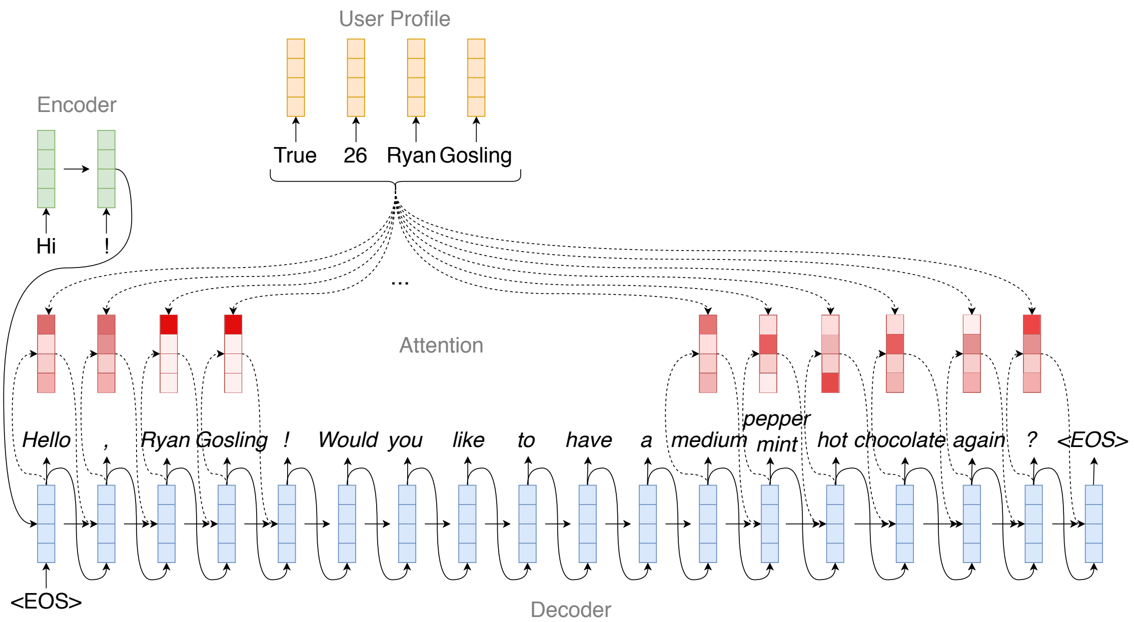


Figure 8.3: A diagram of the Generative Profile Memory Network that shows the architecture consisting of an encoder (for the user response), decoder (for generating the bot response), user profile embeddings and the attention mechanism. The decoder attends over both the encoded profile entries and the context. The attention returns higher probabilities (represented with darker red colours) when the network is more certain of the response (e.g., when using profile information). <EOS> token is used to determine the end of the sentence within the fixed embedding structure. The example corresponds to that of Figure 8.1.

[6]Term frequency (tf) represents how frequently a word occurs in a document, and is computed in (Zhang et al., 2018) from the GloVe index using Zipf's law, tf$= 1e6/(\mathtt{index}^{1.07})$. The weights of words are then calculated by $\alpha_i = 1/(1 + \log(1 + \mathtt{tf}))$.

[7]Note that while this baseline was used in (Zhang et al., 2018) with the Persona-Chat dataset, it was deprecated from the ParlAI library on March 2019. Hence, we use the last available version before the deprecation: `https://github.com/facebookresearch/ParlAI/tree/6a76a555ea84b06e2914cdea4c56a46a5f495821/projects/personachat`

context, 55.13% without context for the PBPI8). While Profile Memory implementation in (Zhang et al., 2018) uses the user response-correct bot responses pairs in *training* for the conversation context, it uses user response-model's prediction pairs in the *validation*, *test* and *out-of-vocabulary (OOV)* sets. This surprisingly performs better than using the correct response (e.g., 51.93% in PB8 and 42.39% in PBPI8). Hence, we kept this method and it also allowed us to have a fair comparison of this baseline to its performance in the Persona-Chat dataset. Similar to the Split Memory, we overwrite the profile attributes if it is updated during the conversation. Profile Memory is equivalent to the Seq2Seq without the profile information, hence, this method is not evaluated with the Barista Dataset.

8.3 Research Questions

Which dialogue architectures are appropriate for long-term interactions in the real world? This research question (RQ3) led us to explore rule-based and data-driven approaches. Consequently, we generated the text-based Barista Datasets based on sets of rules that simulate real-world interactions with a barista, to form the base of the rule-based dialogue management system (RBDMS) and as training and evaluation data for the data-driven architectures. While our RBDMS achieved 100% accuracy in each of the text-based tasks, the results of the study described in Chapter 7 suggested that it is not suitable for real-world interactions. Data-driven approaches may offer more flexibility in dealing with various user responses in the real-world interactions, however, they have not been previously evaluated for long-term interactions or user-specific personalisation. Hence, we cannot hypothesise which model would be more appropriate, or if any would be suitable at all for real-world interactions. Thus, we need to explore the performance of the data-driven architectures on the simulated text-based Barista Datasets to determine which model can perform sufficiently well (i.e., at least 90% accuracy). Consequently, we formulated the following sub-questions to our original research question to explore in this chapter:

- RQ3.1: *Which architecture is the most suitable for generic (non-personalised) task-oriented dialogue?* We will explore the answer to this question using the Barista Dataset.
- RQ3.2: *Which architecture is the most suitable for personalised interactions in task-oriented dialogue?* We will explore the answer to this question using the Personalised Barista Dataset.

- RQ3.3: *How much improvement do user preferences information provide?* The Personalised Barista with Preferences Information Dataset simulates extracting user preferences information from a knowledge-base and providing this alongside the user identity information at the beginning of a conversation. We will evaluate the results of the two Personalised Barista Datasets comparatively to identify the effects of the external information.

- RQ3.4: *What causes inaccuracies in a model?* In Chapter 6, we categorised the bot utterances in the Barista Datasets in terms of phrase types, such as *personal(ised)* (i.e., containing user name or preference), *order details* and *other (remaining)* phrases, in addition to the phrases specific to task 7 of the Barista Dataset (B7). Here we will investigate the errors of the models in the corresponding categories to determine the underlying reasons for inaccuracies in the models. Moreover, we will evaluate the performance in dialogue state tracking (DST), i.e., how well the models choose the correct template corresponding to the dialogue turn.

- RQ3.5: *What is the effect of the out-of-vocabulary (OOV) words, such as new menu items, on the performance?* We will explore the answer to this question using the *OOV* sets.

- RQ3.6: *What is the effect of the dataset size?* We will examine the performance on few-shot learning with the Second Interaction sets within the Personalised Barista Datasets to see whether the models can learn from only a few samples of data. Moreover, we will compare the results of the 1,000 and 10,000 dialogue datasets to observe how much performance improvement does the increase in training data bring.

- RQ3.7: *What is the applicability of the architectures to real-time interaction?* We will evaluate the training and computation time for response generation in the models to understand whether they can be utilised for real-time interactions.

8.4 Experimental Procedure

The experiments relied on the Ghent University IDLab cloud servers and took 6 months to run (February to August 2020). We trained and evaluated the described baselines on the Barista, Personalised Barista and Personalised Barista with Preferences Information

Datasets with varying dataset sizes (i.e., 1,000, 10,000 dialogues and Second Interaction) based on the *test* and *OOV* sets and separately for each task.

The hyperparameters for each method are given in Appendix E.1. The hyperparameters used in the experiments have by no means been extensively explored, and correspond to the hyperparameters from the original implementations (Joshi et al., 2017; Zhang et al., 2018), unless otherwise noted in the text for creating a more fair comparison without decreasing the performance of the original models, based on the limitations of our systems. For instance, in contrast to the original work, we used 100 epochs for training each baseline to ensure equal comparison, except for Key-Value and Supervised Embeddings, which were only trained for 25/15 epochs due to the vast amount of time required to train them (Section 8.5.7 shows the training and test times in more detail). However, the corresponding number of epochs or training time given in the original work were less than or equal to ours. Key-Value was trained for 20 hours (in equivalence to our computational power) on the Persona-Chat dataset, whereas, in our case, the training lasted between 17 to 40 **days per task** in the 10,000 dialogues datasets, even though the tasks in our datasets have 1/5th number of utterances. In addition, we increased the batch size of the Supervised Embeddings to 128 (was 32 in the original implementation) to decrease the training time, and used a batch size of 1 for Seq2Seq and Generative Profile Memory Network on the *test* and *OOV* sets due to out-of-memory errors.

Key-Value, Seq2Seq and Generative Profile Memory Network were trained and evaluated using the ParlAI[8] (Miller et al., 2017) framework with PyTorch (1.1.0) on Python 3.6, while the MemN2N, Split Memory and Supervised Embeddings use Tensorflow (1.13.1) on Python 3.6, without an external framework. A Docker[9] container was created with the code for the modified baselines[10] and the datasets, and the experiments were run in parallel on (a limited number of) cloud servers for each baseline. The training is separate for each task, that is, we do not use the trained model on Task 1 for training on Task 2.

Each *training*, *validation* and *test* set are randomly divided into batches of dialogue examples, in which the *conversation context* (i.e., the conversation history), the *user query* (i.e., the last user response), and the *correct response* (i.e., the correct bot response) are given. All methods have access to the *candidate set* (i.e., set of all bot responses) from all sets during

[8] `https://parl.ai/`
[9] `www.docker.com`
[10] Available online at: `https://github.com/birfan/BaristaDatasets`

training and test. The task is to predict the *correct response*. The *correct response* is compared with the *predicted response* for text (i.e., string) or embedding equality comparison, depending on the model. Correspondingly, the model performance is measured by the *per-response accuracy metric* (Bordes et al., 2016; Joshi et al., 2017), which is the percentage of correct matches within the number of total examples. For the retrieval-based tasks, this metric is also known as *Next-Utterance-Classification* (Lowe et al., 2016). We evaluate both retrieval-based and generative models using the per-response accuracy metric, because each response must be completely correct in a task-oriented dialogue for the interaction to be successful. However, as previously mentioned, Supervised Embeddings, the (predicted and correct response) bag-of-embeddings do not preserve the order of the words in the phrase, and contains the unique words in the utterance, as used in (Joshi et al., 2017). For other models, the word order and the exact words in the phrase are preserved. The *validation* set is used for finding the best performing model during training after each epoch.

Beyond the intrinsic difficulty of each task, *OOV* sets evaluate whether the models could generalise to new entities (i.e., drinks, size types, and snacks) unseen in any training dialogue, which embedding methods are not capable of doing. Thus, Persona-Chat, and (bAbI and) Personalized bAbI dialog papers have a different approach to this evaluation. The former builds a vocabulary from the *training*, *validation* and *test* sets leaving out the *OOV* set, and replaces unknown words with a special token. The latter papers add them to the vocabulary during training due to the fixed size vectors used in MemN2N, Split Memory and Supervised Embeddings. Since we wanted to remain faithful to reproducing these approaches within a different context, we did not change this structure. In addition, removing the *OOV* words from the vocabulary causes erroneous performance measurement in the latter methods. Thus, we will cautiously examine the *OOV* results.

8.5 Results

In this section, we present the results from our experiments that we will examine under the research questions described in Section 8.3 to explore the potential of data-driven approaches in personalised long-term interactions. It is important to recall that the RBDMS, presented in Chapter 7, performs 100% on these datasets because the datasets were created from a set of rules with deterministic bot utterances.

Similar to the evaluations in (Bordes et al., 2016), the best performing methods or the methods that perform within 0.1% margin of the best performing method are highlighted in bold for the per-response accuracy metric[11]. We will report the performance of each method in the order of its average rank in performance.

8.5.1 RQ3.1: Generic Task-Oriented Dialogue

The Barista Dataset consists of non-personalised (i.e., generic) barista interactions. As described in detail in Section 6.2, we defined seven tasks to evaluate the performance of data-driven approaches on various barista interactions that can occur in a coffee shop:

- **Task 1 (B1):** *Greetings*. This task consists of greeting the customer, requesting their drink order, taking the customer's name, noting the order pick up location and saying goodbye to the customer. No order is made in this task.
- **Task 2 (B2):** *Order drink (without greetings)*. In this task, the customers order only a drink.
- **Task 3 (B3):** *Order drink with changes*. The customers order a drink, but they can change the order (up to two times) during the interaction.
- **Task 4 (B4):** *Order drink and snack*. The customers order a drink and (probably) a snack without changes. The probability of ordering a snack is 0.5 (i.e., 50% chance), sampled from a uniform distribution.
- **Task 5 (B5):** *Order drink and snack with changes*. The customers order a drink and a snack (50% chance), but they can change the order (up to two times). The probability of a change is 0.5, sampled from a uniform distribution.
- **Task 6 (B6):** *Order drink and snack with greetings*. This task is the combination of tasks 1 and 4.
- **Task 7 (B7):** *Order drink and snack with changes and greetings*. This task is the combination of tasks 1 and 5, and contains interaction types from all tasks.

The *training*, *validation* and *test* sets have the same customers and drink, size, and snack types, but the customer orders vary randomly. The corresponding performances of

[11] No statistical analysis can be performed on the results, because each method could only be trained and tested once due to the training times and the available resources. Hence, we apply the method of Bordes et al. (2016), which is a common practice in comparing performances of dialogue models in machine learning.

the state-of-the-art data-driven approaches on Barista Dataset with the *test* set on 1,000 dialogues are presented in Table 8.3 based on the per-response accuracy metric.

Sequence-to-Sequence: *Best model* Seq2Seq model performs best between all models for generic task-oriented dialogue, except in the first task where it is a close second. This is a remarkable result, given that Seq2Seq is a generative model that forms sentences word-by-word, meaning that it learned both the grammar and the correct responses, in contrast to the retrieval-based models which only need to learn the correct responses. It also provides a near-perfect performance, showing that it is suitable for generic task-oriented dialogue.

Memory Network While End-to-End Memory Networks (MemN2N) achieves 100% accuracy in the greetings (B1) and ordering a drink without greetings (B2) tasks, the introduction of changes (B3) and an additional order item (B4) decreased its performance. Even though using 2 hops provides the best of MemN2N in B7 (that contains interaction types from all tasks), this model performed poorly in B5. On average, using 3 hops performs the best, which suggests the importance of focusing the attention.

Supervised Embeddings This model was seen as a strong baseline in other works, hence, we used it as a baseline in our evaluations. The results show that it performs well in greetings (B1) task, as reported in (Bordes et al., 2016), but not as favourable as the other models. However, as previously mentioned, Supervised Embeddings model does not preserve the order of the words within the sentence or the time order of the conversation

Table 8.3: The *test* set results of the Barista Dataset with 1,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average and for task 7 (containing all tasks), Seq2Seq is the best performing model, providing near-perfect accuracy.

Task	**MemN2N**			**Key-Value**	**Seq2Seq**	**Supervised**
	Hop1	Hop2	Hop3			
1	**100**	**99.98**	**100**	98.8	99.85	98.72
2	**100**	**99.98**	**99.95**	75.3	**99.92**	76.33
3	97.96	97.73	97.67	65.9	**99.98**	63.59
4	93.33	95.85	98.45	65.7	**99.95**	85.25
5	91.97	79.92	94.85	60.46	**98.97**	69.54
6	96.66	98.7	99.69	78.27	**100**	89.64
7	94.78	96.29	95.99	70.98	**99.85**	80.25

context (i.e., the user utterance or the conversation context is embedded according to their order in the vocabulary) and embeddings contain only unique words, both reasons resulting in a poor performance in changes in the order (B3) task. Therefore, we believe that it is not appropriate for dialogue[12].

Key-Value Profile Memory Network While Key-Value was the best model in the Persona-Chat dataset, our results show that, in general, it is the worst in performance for generic task-oriented dialogue. Its initial good performance in the greetings (B1) task may be attributed to its *chit-chat* capabilities in open-domain dialogue.

8.5.2 RQ3.2: Personalised Task-Oriented Dialogue

The Personalised Barista Dataset, as described in Section 6.3, contains personalised subsequent barista interactions in which the barista recognises customers and recalls their preferences, in terms of the most common (or most recent) drink, size, and snack, and suggests to the customer whether they would like to have it again. These interactions are built on top of the Barista task 7 (B7), that is, the customers' initial interactions corresponds to a dialogue from B7, and the subsequent interactions may contain relevant phrases. The personalised phrases depend on the interaction, which we defined as tasks:

- **Personalised Task 0 (PB0):** *Confirmed personalised order suggestion for new customers*. This task consists of recalling and suggesting the preferences of new customers (i.e., customers that are not in the *training* set). The customers always confirm the suggestion.
- **Personalised Task 1 (PB1):** *Confirmed personalised order suggestion for previous and new customers*. This task has the same type of dialogue interactions as PB0, but contains (previous) customers from the *training* set, as well as the new customers to evaluate continual learning. The following tasks are built upon this task.
- **Personalised Task 2 (PB2):** *Recognition error*. This task evaluates the bot to correct itself after incorrect recognitions in open world recognition, where 90% of the known customers are recognised correctly and 10% of the new customers are confused with another customer.

[12]This conclusion applies to the Supervised Embeddings implementation of Joshi et al. (2017). An implementation that uses an embedding to preserve the word order in the sentence would be more appropriate and may provide different results.

- **Personalised Task 3 (PB3):** *Incorrect recall*. This task evaluates the bot to correct itself after an incorrect recall of the preferences of the customer, with 30% of the dialogues containing incorrect recalls.

- **Personalised Task 4 (PB4):** *Changes to preference*. The task evaluates whether the barista can learn the most common preferences of the customer, as the customer may ask for a different drink/size/snack at each interaction. A change in preference has a probability of 0.5, sampled from a uniform distribution.

- **Personalised Task 5 (PB5):** *Recognition error and incorrect recall*. This task is the combination of tasks 2 and 3.

- **Personalised Task 6 (PB6):** *Recognition error and changes to preference*. This task is the combination of tasks 2 and 4.

- **Personalised Task 7 (PB7):** *Incorrect recall and changes to preference*. This task is the combination of tasks 3 and 4.

- **Personalised Task 8 (PB8):** *All tasks*. This task is the combination of tasks 2, 3 and 4. This task evaluates all the scenarios that can occur in a personalised barista interaction.

The user identity information is presented at the beginning of a dialogue and after the customer gives their name (e.g., for new customers or recognition errors), in the format: *True* (for known customer/ *False* for new customer), *8* (customer ID/ *0* for new customer), *Sarah Michelle Gellar* (customer name/ empty for new customer). The *training*, *validation* and *test* sets have the drink, size, and snack types, but the customer orders vary randomly.

The performance of the data-driven approaches on this dataset, based on the per-response accuracy metric, are presented in Table 8.4. The results show that in all the models, the performance drops considerably below the level in the generic task-oriented dialogue, showing that personalisation in long-term interactions is a challenging problem for the state-of-the-art data-driven approaches. This is, of course, anticipated due to the *catastrophic forgetting* problem (i.e., drastic loss of performance on previously learned classes upon learning a new class) in continual learning. Especially when the *test* set contains completely different users than those in the *training* set (PB0), the accuracy drops substantially in all the models, except for Key-Value, which surprisingly performs almost equally

Table 8.4: The *test* set results of the Personalised Barista Dataset with 1,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average and for task 8 (containing all tasks), MemN2N is the best performing model.

Task	**MemN2N**			**Split Memory**			**Key-**	**Profile**	**Seq2-**	**Super-**
	Hop1	Hop2	Hop3	Hop1	Hop2	Hop3	**Value**		**Seq**	**vised**
0	42.52	43.01	43.27	44.97	44.89	44.63	**61.91**	40.91	40.28	55.85
1	70.42	70.97	70.82	**71.83**	71.34	71.05	61.97	69.73	50.21	67.03
2	69.35	69.32	69.77	**70.25**	70.06	69.93	55.28	69.32	41.45	63.65
3	**68**	67.22	67.51	66.24	67.53	66.69	47.65	61.42	46.24	62.06
4	72.71	**76.58**	75.67	72.44	74.42	74.81	47.31	63.86	64.33	61.16
5	62.83	63.82	**65.37**	63.45	62.83	63.8	43.2	59.73	45.13	57.61
6	70.48	**74.71**	73.85	69.98	72.99	72.88	44.38	62.22	54.78	57.12
7	70.42	**72.77**	72.04	68.26	70.2	63.9	43.74	58.18	68.43	60.26
8	68.8	**71.81**	70.01	64.93	66.95	66.95	40.09	57.79	60.58	56.17

well in tasks 0 and 1. While the personalised barista can decrease the number of turns, the presence of user recognition errors (PB2) and incorrect recalls (PB3) may increase the number of turns. Hence, the additional turns that contain misinformation about the customer or their order decrease the performance of the models in those tasks. Overall, the results show that MemN2N model shows potential for user-specific personalisation in task-oriented dialogue, however, none of the models perform adequately well to be used in real-world interactions. Below we analyse the performance of each of the models in detail.

Memory Network: ***Best model*** On average (in 6/9 tasks) and in task 8, which contains all non-generic and personalised tasks, MemN2N performs best of all models, especially using 2 hops. However, in task 0, the performance is poor and below that of the Split Memory and Key-Value, indicating that this model is not appropriate for applying it to only new users. Moreover, it performs slightly worse than Split Memory for PB1 and recognition errors (PB2), however, it is more competent in handling incorrect recalls (PB3) and changes to preference (PB4) than all other models, showing that it is more capable of changing an incorrect order and also tally orders to find and suggest the most preferred order.

Split Memory Network While it is the second best model based on the results, we expected Split Memory to perform better overall than MemN2N because it pays separate attention to the user identity information (as shown in Table 8.2), which is why it performs slightly better than MemN2N in tasks 0, 1 and 2 (recognition error). However, for overcoming incorrect recalls (PB3) and making changes to preference and tallying (PB4), the model is not as good as MemN2N. We believe this is due to its inferior performance in issuing application program interface (API) calls (i.e., choosing the correct response based on the user utterance) and updating (the response according to changes in the user requests), as reported in (Joshi et al., 2017), suggesting that the simpler MemN2N model is more suitable for tasks which do not require compositional reasoning over various entries in the memory.

Supervised Embeddings Within a close competition with the Generative Profile Memory Network, the Supervised Embeddings is the third best model on average. Its performance in dealing with changes to the preference (PB4) or the order (in the Barista tasks 3 and 5) is in contrast with the findings of Joshi et al. (2017), in which it performed very poorly (12%) in updating API calls.

Generative Profile Memory Network Profile Memory allows focusing on the user identity information, thus, performs better than Seq2Seq in tasks focusing on such information. However, it performs worse in tasks that involve changes to the preference. The most prominent reason is that it uses the model predictions, in the *validation* and *test* sets, instead of the correct labels, which may have decreased its performance in tracking the dialogue state or the order items in the dialogue. On the contrary, using correct labels surprisingly decreases the performance of the model, as we previously mentioned.

Sequence-to-Sequence Despite achieving near-perfect accuracy in generic task-oriented dialogue, the Seq2Seq model does not perform well in personalising the dialogue.

Key-Value Profile Memory Network While for open-domain dialogue, Key-Value considerably outperformed (30-50% accuracy) both Profile Memory and Seq2Seq (8-10% accuracy) as well as the other retrieval-based models (Zhang et al., 2018), the results show the contrary in personalised long-term task-oriented dialogue. Nonetheless, the model is indifferent to the customer database (i.e., performing almost equally in PB0 and PB1), which is an important aspect, for instance, for deploying the model in different locations of the same coffee shops. However, this model is not able to handle the inaccuracies of real-world dialogue (PB2 and PB3), as well as changes in customer preferences.

8.5.3 RQ3.3: User Preferences Information

Rule-based approaches relying on a knowledge-base have the advantage of knowing the preferences of the user prior to the conversation, whereas, this information needs to be obtained from dialogue within the Personalised Barista Dataset, which makes it a very challenging dataset. Thus, we created the Personalised Barista with Preferences Information Dataset, as described in Section 6.4, by providing the user preference information (e.g., *small* (the most common size of the most common drink order), *espresso* (the most common drink order), *pain au chocolat* (the most common snack order)), alongside the user identity information, at the beginning of the conversation, similar to the Personalized bAbI dialog dataset (Joshi et al., 2017). The resulting performance of the data-driven approaches on this dataset, based on the per-response accuracy metric, are presented in Table 8.5.

We expected that adding user preference information, alongside the user recognition information, would improve the accuracy of all the models, especially for tasks focusing on learning and recalling the user preference (0, 1, 4). However, it seems to have a varying effect depending on the task and model. Nonetheless, it improved the accuracy in all models for task 8 up to an increase of 15.58% (for Seq2Seq). Similar to the Personalised Barista Dataset, on average MemN2N performs the best, however, in contrast, Seq2Seq performs best in task 8. However, the overall accuracy still remains considerably below that of the RBDMS (100%), consequently, no model is adequate for personalised long-term interactions in the real world.

Table 8.5: The *test* set results of the Personalised Barista with Preferences Information Dataset with 1,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average MemN2N is the best performing model, however, Seq2Seq performs best for the task 8 (containing all tasks).

Task	**MemN2N**			**Split Memory**			**Key-Value**	**Profile**	**Seq2-Seq**	**Super-vised**
	Hop1	Hop2	Hop3	Hop1	Hop2	Hop3				
0	42.9	41.75	43.01	43.39	44.31	43.04	**62.37**	41.37	40.33	53.05
1	70.71	70.42	70.19	70.59	70.39	**70.91**	62.28	69.13	42.62	69.56
2	68.84	68.92	68.82	**69.32**	69.16	**69.29**	54.54	67.65	40.79	65.75
3	**70.3**	**70.26**	70.05	68.44	68.06	69.01	49.39	62.67	66.73	65.27
4	73.6	76.15	75.92	74.17	75.96	**76.31**	50.9	63.68	68.92	67.15
5	65.93	**66.2**	**66.13**	64.47	65.23	64.76	45.35	60.82	62.48	61.86
6	71.94	74.82	**75.15**	71.14	74.3	74.21	45.86	62.31	52.51	61.96
7	73.66	74.34	73.88	70.65	72.19	72.96	45.72	59.12	**75.21**	63.47
8	73.19	72.98	73.3	68.75	71.65	69.73	42.95	58.96	**76.16**	60.66

8.5.4 RQ3.4: Reasons for Inaccuracies

In the previous sections, we compared the performance of the methods task-by-task. However, the overall accuracy in a task does not provide much information about the underlying reasons for the inaccuracies in the models. Hence, we recorded logs during the *test* sets to categorise each error according to the phrase types we presented in Tables 6.3 and 6.6 in Chapter 6, namely in terms of *personal(ised)* (i.e., containing customer name or preferences), *order details* (i.e., containing order item), *other* (remaining) phrases and Barista Task 7 (B7) phrase types. In addition, we evaluated the dialogue state tracking (DST) errors, that is, whether the model responded with the correct template for the conversation turn, regardless of the specific details (i.e., order or user name) in the response. Table 8.6 presents the results of the tasks that contain all the tasks within the Barista (B7), Personalised Barista (PB8) and Personalised Barista with Preferences Information (PBPI8) datasets, in addition to the personalisation tasks where the customer preferences are recalled and suggested, and the customers confirm the suggestion (PB0 and PB1). These latter tasks show whether the models can learn and use customer names and preferences for new customers (PB0) and additionally for previous customers (PB1).

The percentage of errors corresponds to the number of errors divided by the task size (number of user-bot utterance pairs), that is, it corresponds to the percentage of the error in the overall task performance, which can help identify the most common errors in the overall performance and facilitate equal comparison between the models. The percentage of errors within the parentheses correspond to the percentage of the error within the respective phrase types, calculated by the number of errors within the phrase type divided by the total number of user-bot utterance pairs in the phrase type. The sum of errors in *personal(ised)*, *order details* and *other* phrases (i.e., phrases without customer name or order item) equal the total error in the per-response accuracy for the Barista Dataset. In contrast, in the Personalised Barista Datasets, *personal(ised)* phrases also include order items (i.e., for suggesting customers their most preferred order). Similarly, B7 phrases also include order details (i.e., for the confirmation of the order) and *personal(ised)* phrases (i.e., for noting the order location referring to the customer name). Hence, the total error of these categories would be higher than the overall error. However, this allows evaluating the errors within each perspective. Note that while the error within the total utterances may seem low, it may correspond to a high error within the phrase type (given in parentheses).

Table 8.6: Percentage of errors in dialogue state tracking (DST), *personal(ised)*, *order details*, other and Barista Task 7 (B7) phrase types for 1,000 dialogue *test* sets. The best performing methods (or methods within 0.1%) are given in bold for the error in per-response accuracy, and the error percentage within the phrase type is given in parentheses.

Task	**Error Type**	**MemN2N** Hop1	**MemN2N** Hop2	**MemN2N** Hop3	**Split Memory** Hop1	**Split Memory** Hop2	**Split Memory** Hop3	**Key-Value**	**Profile**	**Seq2-Seq**	**Super-vised**
B7	DST	0.54	0.35	0.18	-	-	-	1.89	-	**0**	0.78
	Personal	0.36 (2.80)	0.15 (1.20)	0.14 (1.10)	-	-	-	**0.00** (0.00)	-	**0.00** (0.00)	3.87 (30.09)
	Order	4.84 (21.13)	3.48 (15.19)	3.87 (16.87)	-	-	-	20.97 (91.48)	-	**0.15** (0.67)	15.03 (65.58)
	Other	**0.01** (0.02)	**0.08** (0.12)	**0.00** (0.00)	-	-	-	8.05 (12.54)	-	**0.00** (0.00)	0.85 (1.32)
PB0	DST	21	2.13	0.49	22.62	21.78	3.03	28.44	**0**	0.29	1.73
	Personal	53.36 (97.47)	52.58 (96.05)	52.52 (95.95)	50.79 (92.79)	50.16 (91.63)	50.99 (93.16)	**29.13** (53.21)	54.22 (99.05)	54.74 (100.00)	52.95 (96.74)
	Order	**29.85** (96.72)	30.11 (97.56)	**29.90** (96.91)	30.14 (97.65)	30.74 (99.61)	30.14 (97.65)	30.74 (99.61)	30.80 (99.80)	30.86 (99.99)	30.25 (98.03)
	Other	0.20 (0.50)	0.23 (0.57)	0.23 (0.57)	**0.03** (0.07)	0.14 (0.36)	0.17 (0.43)	4.06 (10.07)	**0.00** (0.00)	**0.06** (0.14)	0.86 (2.14)
	B7	**4.12** (5.56)	4.41 (5.95)	**4.21** (5.68)	4.24 (5.72)	4.96 (6.69)	4.38 (5.91)	8.96 (12.10)	4.87 (6.57)	4.98 (6.73)	5.19 (7.00)
PB1	DST	8.97	4.36	0.66	11.84	8.46	0.97	33.08	**0**	0.29	1.63
	Personal	24.79 (45.52)	24.85 (45.63)	24.33 (44.68)	**23.65** (43.42)	24.08 (44.21)	24.62 (45.21)	28.52 (52.37)	24.85 (45.63)	44.22 (81.21)	33.99 (62.42)
	Order	16.19 (51.89)	**15.59** (49.96)	16.05 (51.43)	15.96 (51.15)	15.71 (50.33)	**15.65** (50.14)	30.84 (98.81)	17.02 (54.55)	29.75 (95.32)	24.13 (77.32)
	Other	**0.06** (0.14)	0.14 (0.36)	0.26 (0.64)	**0.03** (0.07)	0.34 (0.86)	0.14 (0.36)	4.21 (10.50)	**0.00** (0.00)	0.14 (0.36)	4.53 (11.28)
	B7	4.79 (6.45)	**4.18** (5.64)	4.84 (6.53)	4.53 (6.10)	4.59 (6.18)	4.33 (5.83)	9.52 (12.82)	5.42 (7.30)	5.56 (7.49)	9.37 (12.63)
PB8	DST	2.81	2.87	4.89	2.57	3.45	1.96	49.67	2.09	**0.29**	9.88
	Personal	**21.24** (55.93)	21.92 (57.70)	22.20 (58.46)	22.11 (58.21)	23.31 (61.38)	23.46 (61.76)	22.33 (58.80)	23.63 (62.22)	25.43 (66.95)	31.53 (83.01)
	Order	20.39 (57.38)	**16.89** (47.54)	18.47 (51.96)	23.68 (66.63)	20.79 (58.50)	20.76 (58.41)	34.11 (95.97)	30.07 (84.60)	29.78 (83.78)	33.72 (94.89)
	Other	0.71 (1.66)	0.56 (1.32)	0.66 (1.55)	0.66 (1.55)	0.58 (1.36)	0.51 (1.21)	18.02 (42.45)	0.71 (1.66)	**0.03** (0.08)	2.68 (6.31)
	B7	9.67 (12.64)	**6.27** (8.20)	7.60 (9.94)	12.87 (16.81)	9.74 (12.72)	9.47 (12.37)	36.05 (47.10)	18.56 (24.25)	14.04 (18.34)	20.47 (26.75)
PBPI8	DST	4.17	2.68	2.89	3.1	2.79	2.39	55.54	**0.34**	0.75	4.2
	Personal	27.56 (72.57)	21.34 (56.18)	21.76 (57.28)	20.55 (54.11)	21.15 (55.68)	21.08 (55.51)	22.00 (57.91)	21.72 (57.20)	**18.76** (49.38)	27.24 (71.73)
	Order	18.02 (50.70)	13.91 (39.14)	**13.27** (37.33)	19.17 (53.95)	16.24 (45.68)	17.92 (50.42)	21.05 (59.23)	29.36 (82.61)	13.52 (38.06)	27.77 (78.14)
	Other	0.85 (2.00)	0.47 (1.10)	0.48 (1.13)	0.48 (1.13)	0.24 (0.57)	0.48 (1.13)	15.48 (36.48)	**0.05** (0.11)	**0.05** (0.11)	2.89 (6.80)
	B7	6.32 (8.26)	5.58 (7.29)	**4.76** (6.23)	10.62 (13.88)	7.19 (9.39)	9.11 (11.91)	33.53 (43.81)	19.30 (25.22)	5.12 (6.69)	19.43 (25.38)

For instance, MemN2N with 1-hop has 29.85% error due to order phrases in the PB0 task, but this makes up 96.42% of all the *order details* phrases, which means almost all the phrases containing an order is wrong. The key points derived from the detailed analysis of the logs based on these categories will be presented in this section.

New customer names cannot be learned All models except Key-Value can only use names that occur in the *training* set, showing that they are not suitable for incremental learning of new names. Which is why the performance is very low for *personal(ised)* and *order details* utterances in PB0 in comparison to PB1 or PB8. Note that each model has in its vocabulary the customer names from *training*, *validation* and *test* sets, as well as candidates from all sets. This finding is similar to the findings in (Bordes et al., 2016; Joshi et al., 2017)[13], which showed that the performance decreased substantially when entities (e.g., restaurant names, cuisines) are not seen in *training*, which they defined as out-of-vocabulary (OOV). However, those works have not examined the details of these errors. Because of this reason, having a separate memory for user profile information or using preferences information (e.g., in PBPI8) does not markedly improve the performance, as evident in Table 8.7.

Table 8.7: Incorrectly predicted customer name in the PBPI0 task due to the lack of training on the name in the Split Memory model for varying hops. While the model focuses on the name and the preferences of the customer, especially in the first two hops, the predicted response contains a customer name from the *training* set, because the model is unable to learn or use new names for any of the hops. Similar to Table 8.2, using multiple hops resulted in low attention weights in the relevant items and a wrong item suggestion, in contrast to the higher overall accuracy of hop 3 in the datasets. Note that there is no conversation context, because this is the beginning of the dialogue.

	Profile	**Hop1**	**Hop2**	**Hop3**
	True	0.02	0.001	0.00001
	119	0.234	0.201	0.027
	Jared Padalecki	0.236	0.236	0.032
	small	0.166	0.112	0.015
	espresso con panna	0.179	0.059	0.0001
	chocolate chip cookie	0.165	0.39	0.926
Customer Input	Hi !			
Correct Response	Hello , **Jared Padalecki** ! Would you like to have a **small espresso con panna** and a **chocolate chip cookie** again ?			
Predicted Response	Hello , **Angie Harmon** ! Would you like to have a **small breakfast tea** and a **chocolate chip cookie** again ?			

[13] The *OOV* results of (Joshi et al., 2017) are presented in: `https://github.com/chaitjo/personalized-dialog/tree/master/MemN2N-split-memory`

Customer: Hello!
Barista: Hello! Which drink can I get for you?
Customer: A **hot chocolate** please.
Barista: Which size?
Customer: **Small** cup.
Barista: Would you like anything else with that?
Customer: I would like a **doughnut** please.
Barista: So that would be a **small hot chocolate** and a **doughnut**. Is that right?
Customer: Erm, I would like a **slice of banana bread** actually.
Barista: Sure! So that would be a **small hot chocolate** and a **slice of banana bread**. Is that right?
Customer: Let's change slice of banana bread to a **croissant** instead.

Correct Response: *Sure! So that would be a* ***small hot chocolate*** *and a* ***croissant****. Is that right?*
Predicted Response: *Sure! So that would be a* ***small hot chocolate*** *and a* ***slice of banana bread****. Is that right?*

Figure 8.4: Incorrectly recalling the previous order item during a change of order in task 7 of the Barista Dataset within the Sequence-to-Sequence (Seq2Seq) model.

Size of the conversation context affects the performance Models are more prone to errors in longer conversation context, causing confusions of order items (Figure 8.4 and Figure 8.5) or in dialogue state tracking, thus, increasing the percentage of B7 errors, especially in the presence of incorrect recalls or user recognition errors. Also, the entities (e.g., order item or drink size) that did not occur in the conversation context can be used incorrectly in responses within all of the models, as shown in Figure 8.6, especially during changes or longer conversation context. Memory Network (MemN2N) performs best in recalling *order details* in the conversation for personalised task-oriented dialogue, whereas, Seq2Seq is the best in generic task-oriented dialogue.

Customer: Hey there!
Barista: Hello! Which drink can I get for you?
Customer: Can I get a **cordusio**?
Barista: Which size?
Customer: **Large**, thank you!
Barista: Would you like anything else with that?
Customer: May I get a **classic oatmeal**?
Barista: So that would be a **large cordusio** and a **classic oatmeal**. Is that right?
Customer: Oh I am sorry, but I would like an **espresso** instead.

Correct Response: *Sure! So that would be a* ***large espresso*** *and a* ***classic oatmeal****. Is that right?*
Predicted Response: *Sure! So that would be a* ***large espresso****. Is that right?*

Figure 8.5: Missing parts of the order during change of order in task 7 of the Barista Dataset in the Sequence-to-Sequence (Seq2Seq) model.

Customer: Hey there!
Barista: Hello! Which drink can I get for you?
Customer: **Green tea**, please?
Barista: Which size?
Customer: **Large**, thank you!
Barista: Would you like anything else with that?
Customer: I would like a **chocolate caramel shortbread**.

Correct Response: *So that would be a **large green tea** and a **chocolate caramel shortbread**. Is that right?*
Predicted Response: *So that would be a **large green tea** and a **chocolate chip cookie**. Is that right?*

Figure 8.6: Dialogue example for incorrect item that does not occur in the conversation context in task 7 of the Barista Dataset in the End-to-End Memory Networks (MemN2N) with 3 hops.

User recognition errors and incorrect recalls are frequent in all personalised tasks Even in models with a separate memory for user profile information, customers are confused with new or other known customers and their preferences are incorrectly recalled, as shown in Table E.3 (Appendix E.2). These show the importance of training the models to appropriately react to the corresponding user responses in the presence of these errors in the real world.

Generative models learn sentence grammar and structure well Despite generating sentences word-by-word, Seq2Seq and Profile Memory performs very well in learning the correct grammar and template in responding to the users, with rare errors with missing words in the response, as shown in Figure 8.7 or mixing words from different phrases. One peculiar error in the Profile Memory was responding to a variety of user utterances with *"Your order will be ready at the next counter, Rachel."*.

Profile: True, 96, Ariana Grande

Customer: Hi!

Correct Response: *Hello, Ariana Grande! Would you like to have a medium caramel macchiato again?*
Predicted Response: *Hello, Ariana?*

Figure 8.7: Grammatically incorrect bot response with missing words in task 8 of the Personalised Barista Dataset in the Profile Memory.

Generative models are better in dialogue state tracking Within all tasks in all datasets, Seq2Seq or Profile Memory perform best in dialogue state tracking.

Time order within the conversation context is important Supervised Embeddings does not maintain the time order of the dialogue turns in the conversation context due to its bag-of-words embedding structure, as described in Section 8.2.1. Due to this, it cannot track and correctly respond to the changes in the preference, user recognition errors and the incorrect recalls, because the corresponding words may appear in a different order in the bag-of-words embeddings than the order of the sentence and the conversation context.

Key-Value is not suitable for task-oriented dialogue Key-Value performs poorly in dialogue state tracking, and the most prominent reason is that it repeats the previous bot utterance. It can use the correct customer name, however, it does so frequently within the wrong context. For instance, instead of suggesting the user preference at the first turn in the dialogue, it responds with the phrase for the incorrect recall, *"Alright! Which drink can I get for you, Lena?"*. In addition, since the method uses only last bot-user response pair in the context (which performs better than using the full context, as stated in Section 8.2.5), the performance in the order confirmation is poor because the drink and size are missing from the conversation context. In contrast, having the last information should improve the performance for changes in the user preferences, since the previous order confirmation and the user change is both available in the context (e.g., as evident in Table E.4 in comparison to Table E.5 in Appendix E.2), however, this also does not appear to be the case. Overall, the performance of the model is very low in all datasets, hence, we can conclude that this model is not suitable for task-oriented dialogue.

8.5.5 RQ3.5: Out-of-Vocabulary

The previous section showed that customer names that do not appear in the *training* set cannot be used by most of the models in the *test* sets, hence, the performance of the models dropped drastically. In this section, we evaluate the performance of the models on new customers that do not appear in the *training* set, in addition to new order items (i.e., drinks, sizes and snacks). In these evaluations, the previous customers from the *training* set do not appear, similar to the PB0 and PBPI0. However, as we previously noted, the definition of out-of-vocabulary (OOV) for Bordes et al. (2016) and Joshi et al. (2017) differs from that

Table 8.8: The *out-of-vocabulary (OOV)* set results of the Barista Dataset with 1,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average and for task 7 (containing all tasks), Seq2Seq is the best performing model, similar to the *test* set.

Task	**MemN2N** Hop1	Hop2	Hop3	**Key-Value**	**Seq2Seq**	**Supervised**
1	**79.9**	78.62	76	75.15	76.85	77.45
2	74.05	73.8	70	18.07	**75**	74.28
3	55.07	54.53	58.34	9.01	**62.87**	60.96
4	67.8	65.58	65.7	11.07	**75**	74.58
5	60.07	50.32	60.26	6.7	**63.59**	57.72
6	62.31	59.63	60.03	34.39	71.93	**72.64**
7	55.42	61.54	60.01	32.83	**64.96**	61.67

of Zhang et al. (2018): the former work adds the new entities to the vocabulary, whereas the latter does not. In addition, removing the *OOV* words from the vocabulary for the methods implemented by Joshi et al. (2017) resulted in erroneous accuracy metrics. We followed each work according to their definition to remain faithful to the work. Table 8.8 presents the *OOV* set results of the Barista Dataset, and the remaining results for the 1,000, 10,000 dialogues and Second Interaction datasets are presented in Appendix E.3 for reasons of perspicuity. Table 8.9 presents the percentage errors for the *OOV* sets in *personal(ised)*, *order details*, other and B7 phrase types and dialogue state tracking performance within *OOV* set of 1,000 dialogues. Note that because the recognition errors (PB2) task may confuse customers with those from the *training* set and offer their preferences, there are a few *training* set order items within the Personalised Barista tasks within the *OOV* set. Thus, the *order details* errors in the PB8 and PBPI8 tasks is less than that of B7, PB0 and PB1.

Out-of-vocabulary entities decrease the accuracy drastically The results show that regardless of whether the OOV words are included in the vocabulary or not, all methods have a drastic drop in performance in the *OOV* sets. When we compare the performance of the methods with the performance in the Barista Datasets *test* sets, most methods lost 20-40% of accuracy. Seq2Seq model still performs best in all the models for the Barista Dataset containing OOV entities, however, the Supervised Embeddings model performs best overall in the Personalised Barista Datasets.

Table 8.9: Percentage of errors in DST, *personal(ised)*, *order details*, other and Barista Task 7 (B7) phrase types for 1,000 dialogue *out-of-vocabulary (OOV)* sets. The best performing methods (or methods within 0.1%) are given in bold for the error in per-response accuracy metric, and the error percentages within the phrase types are given in parentheses.

Task	**Error Type**	**MemN2N** Hop1	Hop2	Hop3	**Split Memory** Hop1	Hop2	Hop3	**Key-Value**	**Profile**	**Seq2-Seq**	**Super-vised**
B7	DST	19.49	11.5	13.48	-	-	-	26.57	-	**0.8**	10.08
	Person-	**10.04**	10.41	11.19	-	-	-	11.55	-	11.95	12.43
	al	(77.80)	(80.60)	(86.70)				(89.50)		(92.60)	(96.30)
	Order	22.51	22.30	22.54	-	-	-	**21.38**	-	22.54	21.56
		(99.89)	(98.91)	(100.00)				(94.85)		(100.00)	(95.65)
	Other	12.02	5.76	6.26	-	-	-	34.30	-	**0.54**	4.34
		(18.62)	(8.92)	(9.70)				(53.14)		(0.84)	(6.72)
PB0	DST	23.19	4.07	2.45	24.23	22.49	4.71	31.33	2.02	2.97	**1.76**
	Person-	52.96	52.15	53.05	**50.27**	51.05	52.18	54.35	53.57	54.87	53.51
	al	(96.52)	(95.05)	(96.68)	(91.62)	(93.05)	(95.10)	(99.04)	(97.62)	(99.99)	(97.52)
	Order	30.67	30.70	30.70	30.70	30.70	30.70	**30.41**	30.70	30.70	30.64
		(99.89)	(99.99)	(99.99)	(99.99)	(99.99)	(99.99)	(99.05)	(99.99)	(99.99)	(99.80)
	Other	0.64	**0.26**	1.24	0.90	0.61	1.16	20.33	1.04	0.49	0.92
		(1.57)	(0.64)	(3.07)	(2.21)	(1.50)	(2.86)	(50.28)	(2.57)	(1.21)	(2.29)
	B7	5.31	**4.97**	5.95	5.60	5.31	5.86	25.01	5.75	5.20	5.57
		(7.18)	(6.71)	(8.04)	(7.57)	(7.18)	(7.92)	(33.79)	(7.76)	(7.02)	(7.53)
PB1	DST	17.9	8.9	7.22	26.33	18.96	6.76	34.37	**0.36**	1.27	6.3
	Person-	43.67	43.92	44.23	**42.20**	42.81	43.72	45.14	44.08	44.94	44.59
	al	(95.45)	(96.01)	(96.68)	(92.23)	(93.57)	(95.57)	(98.68)	(96.34)	(98.23)	(97.46)
	Order	28.83	28.83	28.83	28.77	28.83	28.83	**28.27**	28.83	28.83	28.83
		(99.98)	(99.98)	(99.98)	(99.81)	(99.98)	(99.98)	(98.05)	(99.98)	(99.98)	(99.98)
	Other	0.31	1.07	4.27	2.54	1.37	2.90	25.72	**0.20**	0.56	8.08
		(0.67)	(2.33)	(9.33)	(5.56)	(3.00)	(6.33)	(56.23)	(0.44)	(1.22)	(17.67)
	B7	8.80	9.56	12.76	10.98	9.86	11.39	33.91	**8.69**	9.05	16.57
		(11.04)	(12.00)	(16.02)	(13.79)	(12.38)	(14.30)	(42.57)	(10.91)	(11.36)	(20.81)
PB8	DST	10.74	16.4	18.31	11.55	17.59	12.46	59.84	9.48	**1.78**	17.59
	Person-	**32.61**	33.18	33.30	32.83	33.46	33.30	34.52	33.68	32.90	34.68
	al	(92.63)	(94.23)	(94.58)	(93.25)	(95.03)	(94.58)	(98.05)	(95.65)	(93.43)	(98.49)
	Order	**33.05**	**32.96**	33.08	**33.05**	**33.05**	33.11	33.18	33.33	33.55	33.15
		(98.31)	(98.03)	(98.40)	(98.31)	(98.31)	(98.50)	(98.68)	(99.15)	(99.80)	(98.59)
	Other	5.35	6.79	7.86	4.85	7.92	3.26	31.96	6.85	**1.28**	1.66
		(11.83)	(15.02)	(17.37)	(10.73)	(17.51)	(7.20)	(70.65)	(15.15)	(2.84)	(3.67)
	B7	24.76	26.23	27.07	24.29	27.42	22.69	49.92	26.35	20.85	**20.72**
		(31.38)	(33.24)	(34.31)	(30.78)	(34.75)	(28.76)	(63.27)	(33.40)	(26.42)	(26.26)
PBPI8	DST	15.24	16.96	15.71	16.24	16.71	9.7	74.71	**7.67**	8.48	8.36
	Person-	33.08	33.68	33.24	33.24	33.15	33.62	**31.46**	33.99	33.21	34.05
	al	(93.96)	(95.65)	(94.40)	(94.40)	(94.14)	(95.47)	(89.34)	(96.54)	(94.31)	(96.71)
	Order	32.71	32.14	32.93	32.55	32.93	32.83	**29.86**	33.05	33.02	32.68
		(97.29)	(95.61)	(97.94)	(96.82)	(97.94)	(97.66)	(88.81)	(98.31)	(98.22)	(97.19)
	Other	5.98	7.48	6.60	7.23	10.30	3.76	39.03	**3.07**	6.82	**3.16**
		(13.22)	(16.54)	(14.60)	(15.99)	(22.77)	(8.30)	(86.29)	(6.78)	(15.09)	(6.99)
	B7	25.16	26.10	25.76	26.23	29.70	23.07	57.06	22.44	26.38	**22.07**
		(31.89)	(33.08)	(32.65)	(33.24)	(37.65)	(29.24)	(72.32)	(28.44)	(33.44)	(27.97)

Customer: Hi!
Barista: Hello! Which drink can I get for you?
Customer: May I have iced **pumpkin spice chai latte**?
Barista: Which size?
Customer: Grande, please!

Correct Response: *Would you like anything else with that?*
Predicted Response: *Your order will be ready at the next counter, Ariana.*

Figure 8.8: Confusion of Ariana Grande with grande drink size in task 8 of the Personalised Barista Dataset *OOV* set in the MemN2N with 3 hops.

All models perform poorly for new customer names or new orders Comparatively analysing the percentage of errors within the phrase types (i.e., the error percentages are given in parentheses) within the *test* (Table 8.6) and *OOV* (Table 8.9) sets show that all models perform extremely poorly in *personal(ised)* and *order details* phrases (with 0 to 11.19% per-response accuracy) in the Personalised Barista Datasets, because of the new entities. The correct *personal(ised)* phrases in most models (i.e., all models except Key-Value) correspond to using the phrase "Your order will be ready at the next counter" with customers that have their first name same with a customer in the *training* set. On a related note, an amusing and peculiar error specific to MemN2N is confusing *Ariana Grande* (from the *training* set) with the *grande* drink size, as shown in Figure 8.8.

Out-of-vocabulary entities increase dialogue state tracking errors The dialogue state tracking errors increased in all models due to confusing known customers with new customers because the models do not have training on these customers.

Key-Value can learn new entities, but performs worst due to dialogue state tracking errors Despite being able to use new entities (i.e., names or orders), even without having them in the vocabulary, Key-Value performs extremely poorly in *OOV* sets, worse than the other models. The most prominent underlying reason is the increased number of dialogue state tracking errors, especially due to its tendency to repeat the previous bot utterance. The other reason is because of the increased number of words in the vocabulary, it confuses more the orders, names and preferences, as can be seen in Figure 8.9. "Brand", "short" and "raspberry lemonade" are not part of the vocabulary, whereas, Russell is a first name in the *training* set. The __UNK__ is the special token used to represent words that are not in the vocabulary in ParlAI framework. Despite the special token in the profile or

Profile: True, 157, Russell __UNK__

Customer: Hey there!

Correct Response: *Hello, Russell Brand! Would you like to have a* ***short peach iced tea*** *and a* ***chicken and pesto panini*** *again?*
Predicted Response: *Hello, Russell Brand! Would you like to have a* ***short raspberry lemonade*** *again?*

Figure 8.9: Incorrectly recalling the previous order item during a change of order in task 8 of the Personalised Barista Dataset *OOV* set in Key-Value.

the conversation context, Key-Value is able to learn and use those new words.

Memory Network can learn new order items, but fails to use them While MemN2N was not able to use any new customer names, in rare occasions, it was able to use the new order items. This suggests that the model *can* learn new entities, in contrast to our initial conclusion, however, it *does not*, in general.

8.5.6 RQ3.6: Dataset Size

The accuracy of machine learning approaches tends to improve with more data, as the models have more data to better learn the correlations between correct labels and the queries. On the other hand, few-shot learning is a challenging problem (Triantafillou et al., 2017), because the model only has a few samples to learn the patterns in the responses. Especially in combination with continual learning and out-of-vocabulary words, it becomes very difficult for the models to learn new entities, thus, the accuracy may drop. Thus, in this section, we evaluate the effects of the dataset size on the per-response accuracy of the models using Second Interaction dataset that has only 2-3 dialogues per customer in the *training* set, and the 10,000 dialogues datasets which has more data (100 dialogues per customer). Note that directly comparing the performances between Second Interaction, 1,000 dialogues and 10,000 dialogues may result in incorrect conclusions, because the percentage of personalised and order phrases differ between datasets. Hence, we instead compare the percentage of errors within the phrase types across datasets (based on Table 8.6 and tables in Appendix E.4) and compare the performance of the models within each dataset based on the per-response accuracy.

8.5.6.1 Second Interaction

Table 8.10 shows the few-shot learning performance of the models on the Personalised Barista Dataset using the Second Interaction set. The results for the Personalised Barista with Preferences Information Dataset are presented in Appendix E.4.1.

Sequence-to-Sequence is the best model for few-shot learning While there is a varying effect of using few samples on models depending on tasks, Seq2Seq performs best overall in all datasets. For task 0, the performance of most models (i.e., all models except Key-Value and Supervised Embeddings) seems to have remarkably improved, whereas, for the task 8, the performance dropped in all models except Seq2Seq.

Low sample size causes high dialogue state tracking errors The analysis of the types of errors in the models (presented in Table E.14 in Appendix E.4.1) provides the underlying reason for the inaccuracies. The errors in dialogue state tracking increased in most of the models, which caused an increase in the error for B7 phrases. This is anticipated because models have less training on responding correctly to the user utterances, thus, they make more mistakes. In addition, most models perform worse in *personal(ised)* and *order details* phrases in task 8, however, there is no clear pattern for PB0 and PB1.

Table 8.10: The *test* set results of the Personalised Barista Dataset with Second Interaction set (few-shot learning). The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average and for task 8 (containing all tasks), Seq2Seq is the best performing model.

Task	**MemN2N**			**Split Memory**			**Key-Value**	**Profile**	**Seq2-Seq**	**Super-vised**
	Hop1	Hop2	Hop3	Hop1	Hop2	Hop3				
0	59.3	58.55	58.83	59.02	**59.49**	57.99	55.17	56.02	56.95	43.99
1	74.19	74.27	73.34	73.27	**75.27**	73.81	54.62	71.26	62.56	61.52
2	70.06	68.76	68.11	68.4	**70.56**	70.2	46.97	70.13	60.46	55.99
3	62.41	62.85	62.66	63.48	62.85	62.28	45.51	**64.81**	62.91	60.03
4	66.36	67.17	63.97	64.66	64.9	64.02	41.34	66.36	**80.29**	56.3
5	56.56	57.02	55.02	57.53	57.65	56.16	37.33	**62.27**	62.04	50.35
6	59.84	59.29	56.52	60.07	60.73	58.57	37.88	62.73	**78.31**	54.85
7	60.16	61.83	58.87	60.32	59.57	60.81	41.67	62.63	**68.92**	58.3
8	55.55	51.93	55.65	54.26	54.96	55.45	36.17	45.54	**63.43**	51.6

8.5.6.2 10,000 Dialogues

Table 8.11 and 8.12 show the performance on the Barista and Personalised Barista Datasets with 10,000 dialogues, respectively. The results for the Personalised Barista with Preferences Information Dataset and the error analysis based on phrase types are presented in Appendix E.4.2.

Sequence-to-Sequence is the best model for generic task-oriented dialogue Similar to the results in Table 8.3, Seq2Seq performs best in all models, performing perfect or near-perfect in all tasks. MemN2N is also able to respond fully accurately in four out of seven tasks in generic task-oriented dialogue, which shows that the high number of samples improved its accuracy efficiently. The consistent results in both small and large sets confirm that these two models are suitable for generic task-oriented dialogue.

None of the models is suitable for personalised task-oriented dialogue in real-world interactions Split Memory performs more accurately in both Personalised Datasets with a higher number of samples, outperforming the previously best performing model, MemN2N, in most of the tasks except tasks 5, 6 and 8. However, none of the models performed sufficiently well (i.e., above 90% accuracy) to be deployed in personalised long-term real-world interactions. The reasons behind this, as previously discussed, is the lack of ability to learn or use new customer names, as evidenced by the poor performance in task 0 in both Personalised Barista Datasets and the high percentage of error (i.e., mostly above 90%) in *personal(ised)* and *order details* phrases.

Table 8.11: The *test* set results of the Barista Dataset with 10,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average and for task 8 (containing all tasks), Seq2Seq is the best performing model.

Task	**MemN2N** Hop1	Hop2	Hop3	**Key-Value**	**Seq2Seq**	**Supervised**
1	**100**	**100**	**100**	**100**	**99.9**	**100**
2	**100**	**100**	**99.99**	75.19	**100**	66.33
3	99.02	99.49	99.31	67.24	**100**	63.08
4	**100**	**99.99**	**99.99**	75.45	**100**	72.73
5	98.87	99.13	98.74	63.83	**99.93**	70.34
6	**100**	**99.99**	**100**	85.39	**99.98**	97.23
7	99.26	99.38	99.12	74.84	**99.98**	87.37

Table 8.12: The *test* set results of the Personalised Barista Dataset with 10,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average, Split Memory is the best performing model, however, End-to-End Memory Networks (MemN2N) is the best model for task 8 (containing all tasks).

Task	**MemN2N** Hop1	Hop2	Hop3	**Split Memory** Hop1	Hop2	Hop3	**Key-Value**	**Profile**	**Seq2-Seq**	**Super-vised**
0	35.19	35.79	37.01	39.1	40.8	40.82	34.3	34.53	34.89	**51**
1	68.17	67.44	68.19	69.73	70.57	69.54	**72.87**	63.54	64.74	68.17
2	69.39	69.37	69.65	**70.33**	70.21	69.63	68.83	64.15	65.07	63.44
3	69.78	71.22	70.8	69.5	71.56	**71.79**	47	58.09	65.27	61.59
4	80.05	79.91	79.42	**80.49**	79.88	79.23	47.77	62.48	75.29	61.1
5	69.44	**72.35**	72.05	69.28	70.88	71.6	42.72	57.1	58.47	58.09
6	**80.61**	79.73	79.12	79.42	79.43	79.01	44.62	32.92	75.03	58.55
7	78.05	77.7	76.9	78.05	**78.34**	78.19	41.79	73.03	75.78	61.5
8	**78.13**	77.79	**78.18**	77.32	77.41	77.32	39.88	55.35	71.34	56.42

High sample size improves model accuracy for generic task-oriented dialogue As expected, the performance of the models improved with a higher number of *training* samples in the Barista Dataset and in the B7 and *other* phrases for the Personalised Barista Datasets. However, the results suggest that the sample size has a varying effect on recalling or using customers' names or preferences within the dialogue, as evidenced by the inconsistent model performance in tasks based on the percentage of error in *personal(ised)* and *order details* phrases. Nonetheless, most models perform better on task 8 with more training data.

8.5.7 RQ3.7: Training and Execution Times

The quality of real-world interaction depends on the real-time capabilities of the system. Hence, we analyse the training and test times on task 8 of the Personalised Barista Dataset[14], as presented in Table 8.13. The test time per example is calculated by dividing the execution time for the task (in the *test* set) by the number of utterances in each dataset (as given in Table 6.4). Thus, the test time corresponds to the average amount of time the model takes to respond after the user utterance. Hence, we can determine if the model can be used in real-time for human-agent (or human-robot) interaction.

All models are suitable for real-time interaction Based on the training and test times, MemN2N and Split Memory have the lowest time complexity, thus, they are the most

[14]The results are similar in other tasks.

Table 8.13: Training and test times of the models for the task 8 of the Personalised Barista Dataset. The test time per example is calculated by dividing the executing time for the task by the number of utterances in each dataset. The MemN2N and Split Memory models have the lowest time complexity.

Dataset	Dataset Size	MemN2N	Split Memory	Key-Value	Profile	Seq2Seq	Supervised
Training time (hours)	300	**0.02**	0.03	2.70	0.86	0.21	3.98
	1,000	**0.08**	0.09	13.79	2.13	0.87	30.58
	10,000	**1.46**	1.65	1049.87	29.55	6.18	805.28
Test time per example (seconds)	400	**0.0005**	0.001	0.25	0.06	0.15	0.29
	1,000	**0.0006**	**0.0006**	0.39	0.22	0.27	0.41
	10,000	**0.0003**	0.0004	0.68	0.52	0.15	1.22

suitable for real-time interaction. Nonetheless, even with training on 10,000 dialogue dataset, most models (i.e., all models except Supervised Embeddings) can respond under 1 second, which is sufficient for a dialogue model. However, in verbal interaction (e.g., human-robot interaction), the time to process the audio (i.e., voice detection and automatic speech recognition) can increase the time to respond, hence, the lower the response time, the better would be the interaction quality.

Memory Networks and Split Memory are suitable to learn progressively Given the short training and test times required for these models in all dataset sizes, new customer names and preferences can be learned progressively from sequential interactions by re-training the models, which can improve their task performance for personalised task-oriented dialogue.

8.6 Discussion

For the past decade, deep learning approaches have been in high demand and they have achieved great performance in some areas, such as closed-set face recognition, and even "near-human" open-domain conversation capabilities (Adiwardana et al., 2020). Our results show that they are indeed suitable for generic task-oriented dialogue, especially Sequence-to-Sequence and Memory Networks models. However, for long-term human-robot interaction, as we have repeatedly stated in the previous chapters, personalisation is necessary to achieve a good quality interaction to meet user expectations. We have, thus, evaluated the state-of-the-art data-driven approaches in personalised long-term conversation in the task-oriented domain. Our results show that they are not suitable for continual learning. Hence, they are currently not in the stage to be applied to real-world long-term interactions, thus, we cannot compare their real-world performance with that of the rule-based dialogue management system. This is not to say that these models do not have potential. For instance, End-to-End Memory Networks (MemN2N) performs best in the Personalised Barista Datasets, however, it fails to use new customer names or entities. Key-Value Memory Networks, on the other hand, can use new entities, however, performs poorly overall due to dialogue state tracking errors. This suggests that there is potential in using Memory Networks, however, not solely in either form. Since MemN2N can also train and respond to users in a short amount of time, new customer names and preferences can be learned progressively by re-training the model, such that the model can recall these names in the subsequent interactions.

Another important note is that contrary to the usual belief, our results show that generative models are very good in learning grammar and perform best in dialogue state tracking in all models. Moreover, Seq2Seq model also performs best in few-shot learning. While it does not perform as well as Memory Networks in personalised task-oriented dialogue, we believe that it is also a fundamental approach and variants of it (e.g., using transformers (Vaswani et al., 2017) as used by Roller et al. (2020)) may also be the right approach in solving personalisation with data-driven approaches.

8.7 Summary

In this chapter, the state-of-the-art data-driven approaches in dialogue, namely Supervised Embeddings, Sequence-to-Sequence (Seq2Seq) (Sutskever et al., 2014), End-to-End Memory Networks (MemN2N) (Sukhbaatar et al., 2015; Bordes et al., 2016), Split Memory Networks (Joshi et al., 2017), Key-Value Profile Memory Networks (Miller et al., 2016; Zhang et al., 2018) and Generative Profile Memory Networks (Zhang et al., 2018), were evaluated in the Barista and Personalised Barista Datasets developed in Chapter 6. The results showed that Seq2Seq model achieved best and near-perfect per-response accuracy in generic task-oriented dialogue, and MemN2N achieved the best accuracy in personalised task-oriented dialogue, however, no model performed sufficiently well to be deployed in personalised long-term real-world interactions. The reasons behind the inaccuracies of the models in the personalised task-oriented dialogue were identified to be the lack of capability to use new customer names or order items, the poor performance in recalling the user preferences, and user recognition errors. When the user preference information was provided within the dialogue, similar to a knowledge-base extraction, the performance improved only slightly due to the poor performance in (or lack of ability to) using the new customer names. Our results also showed that generative models learn grammar well and are the best methods in dialogue state tracking. In addition, Seq2Seq model performs best when the training data only contains a few examples (few-shot learning), however, the performance drops in all models for few-shot learning due to increased dialogue state tracking errors. Increasing the dataset size, as expected, improves the performance for generic task-oriented dialogue, however, the effect on suggesting the preferences of the customers varies from model to model, depending on the task. The time order within the conversation context was also noted to be an important factor in the accuracy for dialogue state tracking and detecting changes in the user orders. Finally, our evaluations showed that while all models are suitable for real-time interactions, MemN2N and Split Memory Networks have the lowest time complexity in training and test, which makes them suitable for learning new customer names and preferences progressively through re-training.

Chapter 9

Personalisation in Socially Assistive Robotics: A Long-Term Real-World Study

Key points:

- The components of a personalised socially assistive robot architecture for cardiac rehabilitation are described: (a) the sensor interface provides continuous monitoring based on physiological (i.e., heart rate, exertion level) and spatiotemporal (i.e., gait, step length, cadence) parameters and treadmill inclination, (b) the social robot gives immediate feedback and motivation based on the sensor values and cervical posture, and (c) the therapy progress of the patient is tracked and the user is recognised to personalise the feedback with the aim to improve user motivation and adherence.
- The real-world long-term study ran for 2.5 years in the Fundación Cardioinfantil-Instituto de Cardiología clinic and is presented with the corresponding conditions for each of the components. A therapy session ran for 18 weeks or 36 sessions per patient.
- The majority of the patients in all conditions had an improvement in the recovery heart rate, which supports the benefits of cardiac rehabilitation programme.
- The long-term perceptions of the *personalised robot* were highly positive and maintained throughout the programme.

- Both robot conditions improved the expectations about the robot and the system and improved motivation to attend the cardiac rehabilitation sessions.
- The *personalised robot* was perceived slightly more positively than the *social robot* for perceived sociability, ease of use, safety and social presence, but the user recognition errors arising from the face recognition failures resulted in a drop for perceived usefulness, utility and trust.
- Multi-modal Incremental Bayesian Network with Online Learning performed better than the non-adaptive model (MMIBN) and the base face recognition, in the presence of malfunctioning in face recognition.
- The patients in the *personalised robot* condition maintained gaze and social interaction with the robot throughout the programme, and fully complied to the robot's posture correction requests, suggesting that the user engagement was maintained throughout the programme.
- The low exertion levels self-reported by the patients need to be combined with continuous monitoring for reliability and robustness and to facilitate immediate intervention by the medical team in critical situations.
- Continuous monitoring allowed high-intensity training in *personalised robot* condition.

Parts of the work presented in this chapter[1] have been published in Lara et al. (2017a,b); Casas et al. (2018a,b,c, 2019, 2020); Irfan et al. (2020a). Additionally, two articles are under review at the Frontiers in Neurorobotics and the User Modeling and User-Adapted Interaction journals.

[1]This chapter presents the methodology and findings from 3.5-year collaboration with the research group in Colombian School of Engineering Julio Garavito (Nathalia Céspedes Gomez, Jonathan Casas, Juan S. Lara and Andres Aguirre, under the supervision of Marcela Munera and Carlos A. Cifuentes), the medical specialists from Fundación Cardioinfantil-Instituto de Cardiología, Bogotá, Colombia (Monica Rincon-Roncancio and Luisa F. Gutiérrez) and two PhD students from the University of Plymouth (the author of this thesis and Emmanuel Senft, under the supervision of Tony Belpaeme). The contributions of each group are as follows: the research group in Colombian School of Engineering Julio Garavito designed the sensor interface for obtaining continuous and online measurements of the patients, integrated the sensor interface with the robot, conducted the experiments and questionnaires, and contributed to the analysis of the results; the medical specialists from Fundación Cardioinfantil-Instituto de Cardiología contributed a clinical point of view on the design and execution of the overall system, provided feedback on the clinical implications of the results, and applied the designed systems in real-world cardiac rehabilitation programme at the hospital; Emmanuel Senft designed and implemented a robot with generic feedback based on the sensor values, designed the gaze estimation system for posture correction and contributed to the analysis of the results; the author of this thesis designed and implemented the personalised feedback of the robot, integrated the multi-modal user recognition system, and contributed to the design process of the overall system and the analysis of the findings. This work was supported in part by the Royal Academy of Engineering IAPP project Human-Robot Interaction Strategies for Rehabilitation based on Socially Assistive Robotics (grant IAPP/1516/137), Colciencias (grant 813-2017), the EU H2020 Marie Skłodowska-Curie Actions ITN project APRIL (grant 674868), the EU FP7 project DREAM (grant 611391) and the Flemish Government (AI Research Program).

9.1 Motivation

Cardiovascular diseases, which are the disorders of the heart and blood vessels, are considered to be the most critical causes of death, costing 17.7 million lives a year that represent 31% of the global deaths[2]. Cardiac rehabilitation (CR) is the Class I recommendation of the European Society of Cardiology, the American Heart Association, and the American College of Cardiology following a cardiovascular event (Thomas et al., 2007; Piepoli et al., 2010). CR aims to provide therapy to those who have suffered a cardiovascular event to accelerate recovery and reduce the risk of suffering recurrent events through structured exercise prescription, education, and risk factor modification (Giuliano et al., 2017; Kraus & Keteyian, 2007). CR is a long-term programme often lasting 13-14 months, where adherence is vital for the complete recovery of a patient and reduce the risk of suffering recurrent events (Jolly et al., 2007; Suaya et al., 2009; Hammill et al., 2010). Nonetheless, in addition to the low participation in the programme (Altenhoener et al., 2005; McKee et al., 2014), a high percentage (24-50%) of patients who enrol in cardiac rehabilitation programs drop out (Carlson et al., 2000; Scane et al., 2012). The reasons behind dropout or non-attendance vary, such as motivation factors (e.g., lack of interest or faith in the programme, increasing lack of motivation throughout the duration of the programme, anxiety about the exercise component, group cohesion), presence of comorbidities (e.g., depression, obesity, diabetes), health coverage, location and accessibility, and scheduling or work commitments (McKee et al., 2014; Maclean & Pound, 2000; Siegert & Taylor, 2004; Beswick et al., 2005; Cooper et al., 2007; Bethell et al., 2009; Shahsavari et al., 2012; Turk-Adawi et al., 2013; Ruano-Ravina et al., 2016). While other factors are beyond the control of the clinicians, motivational issues can be addressed by providing individual support within the sessions, such as through rigorous supervising during the patient's exercise and quick support in emergent situations (Shahsavari et al., 2012). Moreover, while physical training during the CR is required to enable increased fitness and is safe, it is important to closely monitor the patient's physiological parameters to avoid any complications that may arise during the therapy (Bethell et al., 2009). However, CR programme is generally conducted with large groups, and it is challenging for clinicians to provide continuous and individual support during the session.

[2]World Health Organization statistics: `https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`

As the previous research presented in Chapter 2 shows, a socially assistive robot can provide monitoring, feedback and assistance, increase user motivation, adherence and improve task performance and progress. This, in turn, can facilitate the clinicians to focus on the individual needs of the patients, immediately detect any complications during the session, analyse the patient's progress within the therapy in more detail and provide a more tailored plan. Nevertheless, as previously highlighted, while short term interactions benefit from the novelty effect for high user engagement, long-term studies require robust and complex systems, because the limitations of the robot often come to the fore with repeated interactions, which may result in a decrease of user interest and engagement (Leite et al., 2013). Moreover, the behaviour of the robot might not be attractive enough to keep up with the patient's expectations, and interest by the patient and medical staff might wane over time (Süssenbach et al., 2014; Kidd & Breazeal, 2008), resulting in a declining frequency of use and interaction with the robot (Fernaeus et al., 2010). Thus, designing a physically embodied socially assistive agent that is pleasant and valued to interact with in long-term interactions, and demonstrates a marked improvement in training or recovery of the user in a therapy remains a grand challenge of socially assistive robotics (SAR) (Tapus et al., 2007). Moreover, because the robot is deployed in a real-world therapy with non-expert users (e.g., doctors, nurses, patients), it should be autonomous and require minimal effort from users and medical staff (Feil-Seifer & Matarić, 2005). Personalisation (e.g., addressing the patient with their name, and referring to previous sessions) has been shown to have added benefits in improving user motivation and engagement, helping the clinicians monitor the progress of the patient, and facilitating rapport and trust over long-term interactions (Richardson et al., 2018; Scassellati et al., 2018; Winkle et al., 2018; Clabaugh et al., 2019).

Motivated by improving user's motivation, engagement and adherence to the programme, we established our third and last research objective (RO3) as designing and deploying a fully autonomous personalised socially assistive robot in a real-world cardiac rehabilitation programme at the Fundación Cardioinfantil-Instituto de Cardiología (FCI-IC) hospital (Bogotá, Colombia), as shown in Figure 9.1. In this chapter, we describe the components of the designed system, analyse the impacts of personalisation of the robot in long-term interactions (RQ6) through the physiological evolution of patients throughout the programme, the long-term perceptions of the patients and their interactions with the robot, and an overall comparison of the conditions within the study.

(a) (b)

Figure 9.1: Setup of our system for cardiac rehabilitation programme at the Fundación Cardioinfantil-Instituto de Cardiología (FCI-IC) (Bogotá, Colombia): patient interacting with the (a) tablet interface, (b) personalised socially assistive robot.

9.2 Methodology

CR is conducted in FCI-IC through three phases: (I) *inpatient*, (II) *outpatient* and (III) *maintenance* phase. The patient begins the *inpatient* phase after being hemodynamically stable, and its duration depends on the severity of the cardiac event (e.g., 48 hours). In this phase, the patient is asked to perform passive movements for maintaining muscular tone and reducing risks or any complication. The *outpatient* phase begins immediately after the patient leaves the hospital, lasts an average of 18 weeks with sessions twice a week, in which the patient performs various physical exercises at the hospital, and receives an education program about the risk factors and learning healthy habits (e.g., controlling blood pressure, cholesterol, weight, and stress management). The *maintenance* phase lasts

Figure 9.2: A conventional *training* stage of the *outpatient* phase of cardiac rehabilitation programme at Fundación Cardioinfantil-Instituto de Cardiología (Bogotá, Colombia).

on average about nine months with one or two sessions per week, with the aim to reinforce the information and habits gained during the previous phase.

A conventional outpatient CR session lasts 20-30 minutes and consists of three main sub-stages: (1) *warm-up* via stretching exercises, (2) *training* through physical exercises on a treadmill (Figure 9.2), and (3) *cooldown*, in which low intensity exercises are carried out. During the *warm-up* and *cooldown* stages, the medical staff measures the initial and final heart rate (HR), as well as the initial and final blood pressure (BP). During *training*, the medical staff regularly asks for the exertion level of the patient using the Borg scale (BS) (Borg, 1998; Aamot et al., 2014). The *training* performance of the patient depends highly on the intensity of the session, which is determined by the treadmill speed and inclination. The intensity of the exercise sessions increases through these parameters to improve the physical fitness of the patient (Simms et al., 2007). The overall progress of the patient guides the physiatrists in determining these parameters. In the face of any alerts during the session (e.g., high Borg scale or heart rate), the intensity should be promptly adjusted by the physiatrist. However, because of the high number of patients in the programme and the lack of a telemetry in the CR unit, it is very difficult for the medical staff to monitor the patient continuously during the sessions. Hence, our work addresses the *training* stage to provide continuous and personalised monitoring and feedback to support the medical staff in providing immediate assistance in emergent situations and help them focus more directly on the patients.

It is important to structure the social interaction such that the therapy is not negatively affected, especially considering the medical context, the vulnerability of the patient, the potential of unanticipated events and the typical noise of a real-world environment (Goodrich & Schultz, 2007). Thus, in collaboration with medical specialists, we designed a rule-based system for providing and adapting feedback by the robot to the patient. The architecture of the system, as shown in Figure 9.3, is composed of three main parts: (a) sensor interface for collecting physiological and session parameters, (b) socially assistive robot for providing immediate feedback and motivation, and (c) personalising the feedback according to the patient's progress and adherence.

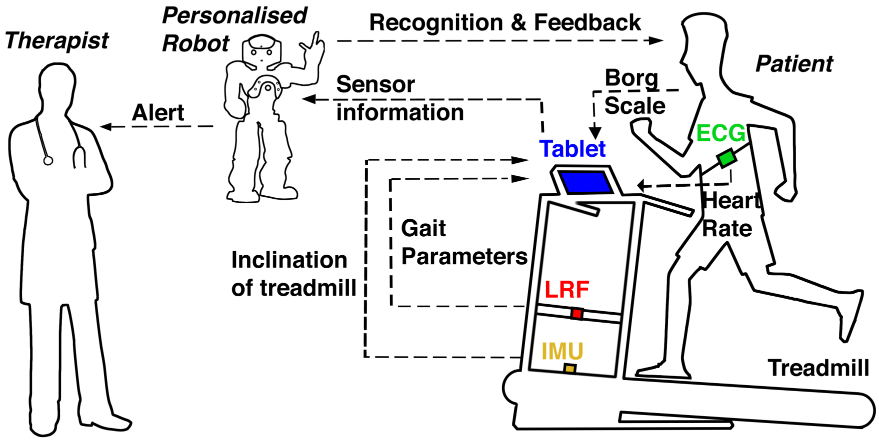


Figure 9.3: The architecture of our system with a personalised socially assistive robot for cardiac rehabilitation programme.

9.2.1 Sensor Interface for Continuous Monitoring

A sensor interface is designed to continuously measure the physiological parameters of the patient and the intensity of the exercise sessions:

- heart rate (HR) and recovery heart rate (R-HR) are estimated with an electrocardiogram (ECG: Zephyr HxM, Medtronic, USA) on the patient during both *training* and *cooldown*,
- Borg scale (BS) is regularly requested through a tablet interface on the treadmill console, as shown in Figure 9.4,
- the spatiotemporal parameters (i.e., patient's gait, step length, cadence) are estimated by a laser range finder (LRF: URG-04LX-UG01, Hokuyo, Japan) that uses infrared electromagnetic wave (785nm) to measure the distance based on the light phase difference,
- the session intensity parameters are estimated by an inertial measurement unit (IMU: MPU-9150, Invensense, USA) for treadmill inclination up to 5 degrees, and the LRF for treadmill speed,
- the cervical posture of the patient is determined by the gaze direction obtained from the tablet camera,
- the systolic blood pressure (BP) is taken by the physiatrist at the beginning and end of a session and entered through the tablet interface.

The tablet interface is also used to provide online (bio)feedback (Figure 9.4) to help the physicians assess the patient more closely and reduce the risks inherent to the CR programme.

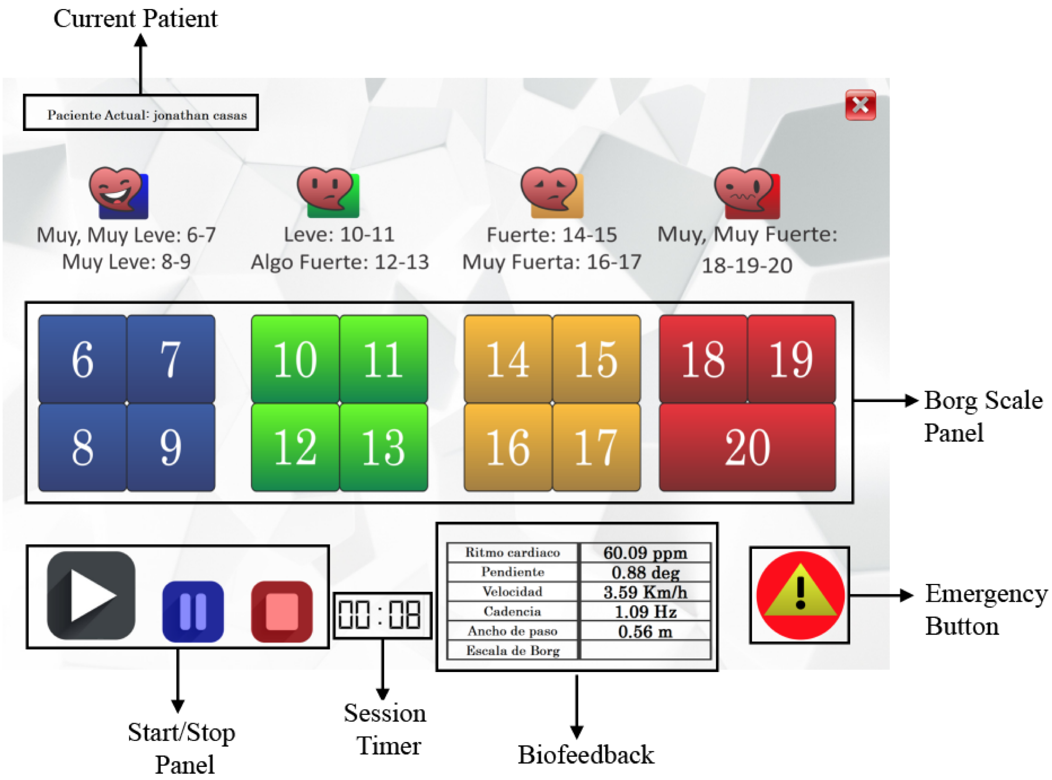


Figure 9.4: Graphical User Interface (GUI) on the tablet for obtaining Borg scale and visualising the sensory information.

9.2.2 Socially Assistive Robot for Immediate Feedback and Motivation

A large body of work in Human-Robot Interaction (HRI) supports the importance of physical embodiment (Fasola & Matarić, 2013b; Deng et al., 2019), including its role in increasing compliance (Bainbridge et al., 2008), likeability (Fasola & Matarić, 2013b; Li, 2015), social engagement (Lee et al., 2006; Wainer et al., 2006; Vasco et al., 2019), adherence (Bickmore & Picard, 2005; Kidd & Breazeal, 2007) and task performance (Vasco et al., 2019), which are essential in, especially long-term, therapy. The main factors that affect user expectation, engagement and motivation, as we discussed in Chapter 2, are the appearance and the behaviour of the robot. A child-like appearance or anthropomorphic but less realistic appearance is more suitable for assistive tasks (Tapus et al., 2007), hence, we use a NAO robot (Softbank Robotics, France) for the therapy, as shown in Figure 9.1. As SAR focuses on providing assistance through social interaction instead of physical interaction (Feil-Seifer & Matarić, 2005), the robot's behaviour relies on the communication and interaction skills that allow it to properly act in a human environment (Shin & Choo, 2011), which leads us back to our research question (RQ2): *How should the robot communicate with users to acquire and convey information?* Previous research shows that socially assistive robots require a set of features, especially for long-term deployment, such as a high level of autonomy, automated perception of human behaviour, quantitative diagnosis and assessment, sensor-based automated health data acquisition and context-appropriate assistance through user interfaces (Okamura et al., 2010; Prescott & Caleb-Solly, 2017). In addition, verbal and non-verbal communication play a crucial role in SAR to increase the ease of interaction and make the robot appear more intuitive and natural (Tapus et al., 2007). However, as Chapter 7 highlighted, current automatic speech recognition (ASR) approaches are not suitable for real-world interactions. Especially since the environment for therapy in FCI-IC contains a high level of noise arising from exercise machines, motivational music and the conversations of the physiatrists with the patients. A tablet interface enables more effective communication by giving visual feedback and providing a means for the user to easily provide more precise input (Liappas et al., 2019). Hence, we use the tablet and sensor interface described in the previous section, and a fully autonomous socially assistive robot that uses verbal and non-verbal communication (i.e., body movements and gaze).

We observed typical physiatrist-patient interactions and collaborated with medical specialists to develop suitable behaviours for giving feedback and motivation. The resulting various types of robot feedback and requests are as follows:

- announcing the session parameters (speed and inclination of the treadmill) at the beginning of a session,
- motivating the patient throughout the session (e.g., *"Let's go! You can do it!"*),
- requesting for entry of the Borg scale (BS) on the tablet interface at certain periods,
- requesting correction of the cervical posture (i.e., ask the patient to look straight ahead instead of down to the treadmill) to reduce the risk of dizziness, falls and nausea,
- warning high heart rate (HR) and requesting confirmation from the patient of the health status, which leads to an alert to the medical staff if the patient is not feeling well (e.g., "You seem like you are starting to get tired, is everything all right?"),

- requesting confirmation of the Borg scale (BS) on the tablet interface if the value exceeds the critical threshold, but the heart rate value is in a healthy range,
- alerting the medical staff when the heart rate or Borg scale exceed critical values ("Your heart rate is too high, I am calling for help. Doctor, could you please come here?" with a waving gesture),
- ending the session with a farewell.

The corresponding finite state machine (FSM) is shown in Figure 9.5. The warning and critical alert thresholds of heart rate are determined by the physiatrists based on the progress of the patient throughout the programme. The *critical heart rate threshold* corresponds to the maximum heart rate allowed for the patient ($\mathtt{HR}_{\mathtt{max}}$), calculated using the Karvonen formula (She et al., 2014) by the clinicians as presented in Equation 9.1, where $\mathtt{HR}_{\mathtt{optimal}}$ represents the optimal heart rate during the exercise, $\mathtt{IHR}$ refers to the initial resting HR (taken by the clinicians when the patient arrives to the clinic), and $\%\mathtt{Effort}$ represents the percentage of desired exercise intensity. Exceeding this level may result in complications, hence, it is extremely important to immediately alert the medical team when this value is reached. Correspondingly, the robot directly alerts the medical team without confirmation from the patient. While the *high heart rate* warning does not represent a critical situation, it can give an initial warning to the medical staff about the patient's condition. The critical Borg scale value is considered to be above 12 in FCI-IC. In

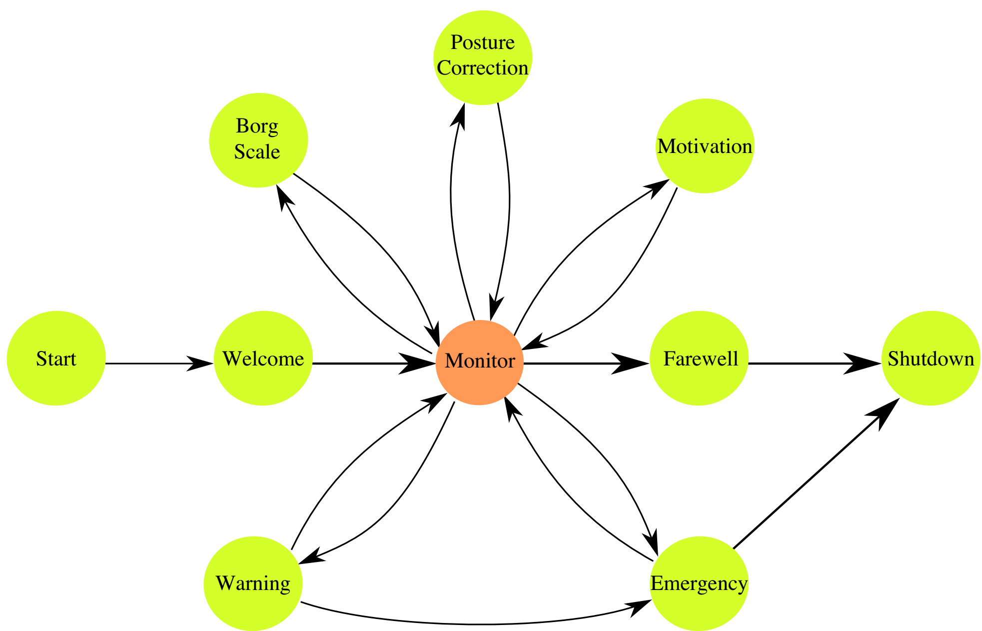


Figure 9.5: Finite state machine (FSM) of the robot behaviours.

order to ensure that the medical staff has heard the robot's alert, the alert (for critical heart rate or Borg scale) is repeated until the medical staff comes and touches the robot's head. Similarly, Borg scale requests and high heart rate confirmation are compulsory, i.e., the robot repeatedly requests these parameters if the patient does not respond within a given time. We use body movements (i.e., animated speech) and face tracking (for gaze) using NAOqi to improve the naturalness of the communication of the robot.

$$\mathsf{HR}_{optimal} = [(\mathsf{HR}_{max} - \mathsf{IHR}) * \%\mathsf{Effort}] + \mathsf{IHR} \tag{9.1}$$

9.2.3 Personalised Socially Assistive Robot

In order to personalise the interaction, we need a system that can autonomously and incrementally recognise users in a real-world therapy, with minimum efforts from the patients and the doctors. Hence, we used Multi-modal Incremental Bayesian Network with Online Learning (MMIBN:OL) with explicit confirmation of identity in order to avoid any errors in the therapy. The reason we chose online learning on the contrary to our findings in Chapter 5 is to track the attendance of the user and evaluate MMIBN:OL in a real-world environment.

As previously pointed out, personalisation has been identified as a critical tool to improve user compliance, engagement and motivation, and in turn, the adherence. *The type of information that should be recalled for personalisation* (RQ4) varies depending on the application of rehabilitation, but a common technique includes user profiling (e.g., name, performance, preferences, schedule) to adapt the feedback within short-term and long-term therapy (Tapus et al., 2007; Ahmad et al., 2017). Therapists also personalise their feedback for improving motivation, and use reminders and prompts for holding the patient accountable (Winkle et al., 2018). In addition, the medical team in FCI-IC identify personalisation, sociability and social presence of the robot as the key elements for motivation, engagement and compliance in the CR programme (Casas et al., 2019). Based on the input of medical specialists in FCI-IC, we decided to use the patient's name, therapy progress and attendance for personalising the feedback, as described in detail below.

In order to increase the sociability of the robot, we personalise the robot's feedback by referring to the patient with their name periodically throughout the session. In addition, the progress of the patient (based on the number of alerts experienced) and the relative intensity of the sessions are tracked, which is used to motivate the patient for the current and upcoming sessions. (1) At the beginning of a session, the current session parameters are announced along with the *relative intensity* of the session and the previous session progress, such as *"In the previous session, you experienced an alert for the heart rate. I am sure it will be all fine this time!"*. The *relative intensity* is defined (by the physiatrists in FCI-IC) as **higher** if either the speed or inclination of the treadmill is higher than the previous session, and **lower** if both of these parameters are lower. (2) At the end of the session, session progress is compared based on the relative intensity, e.g., *"We had a lower number of alerts in this session than the previous one, even though the session intensity was higher. Let's keep up the good work, [PATIENT_NAME]!"*, *"We had more alerts... Next time will be better, [PATIENT_NAME]!"* or *"Wonderful, we had no alerts this session... I am glad to have been here for you, [PATIENT_NAME]!"*. In order to provide only positive feedback (as decided by the physiatrists), we removed the comparison of the relative intensity if it was less intense.

We enforce social presence by tracking the attendance of the patient to improve the adherence to the CR programme. If a patient does not attend the therapy sessions twice a week except for national holidays, the robot comments on the situation with *"You didn't come to the session last (X) session(s). I hope everything is all right!"*. We also aim to increase positive sociability by commenting on the weekend/national holiday, e.g., *"I hope you had a nice weekend/holiday!"* based on the date of the session.

9.3 Experimental Procedure

The primary purpose of this collaborative project is to evaluate the applicability and effects of a socially assistive robot and personalisation in a cardiac rehabilitation programme. Consequently, we designed three conditions within the project to evaluate the effects of using a robot and personalisation:

- *Control*: using only the sensor and tablet interface, as described in Section 9.2.1. No feedback is given by the tablet in order to closely resemble the conventional CR programme conditions.
- *Social robot*: combining the sensor and tablet interface with a socially assistive robot, as described in Section 9.2.2.

- *Personalised robot*: recognising patients and tracking their progress and attendance to adapt the feedback and motivation of the robot, as described in Section 9.2.3.

Our work focuses on the *personalised robot* condition, which is aimed to improve user engagement and motivation in order to increase adherence to the programme, in addition to helping the medical team and patients to track the therapy progress.

The study took place directly at the Fundación Cardioinfantil-Instituto de Cardiología (FCI-IC) clinic (Bogotá, Colombia) for treating patients in the outpatient phase of the CR programme lasting 18 weeks with two sessions per week. There was an experimenter present in the room during the therapy sessions for safety purposes, but the experimenter did not interfere with the therapy, except in the case of system failures. We validated the reliability, robustness and suitability of both the sensor interface and the social robot for the CR programme under laboratory and clinical settings (Lara et al., 2017a,b; Casas et al., 2018a,b,c) before starting the clinical trials.

9.3.1 Participants

The study started in August 2017 for the *control* and *social robot* conditions, and in October 2019 for the *personalised robot* condition. The patients' schedules were arranged such that during a therapy session with 20 patients, only one subject at a time participated in the session. We recruited 15 patients each (with various starting times) for the *control* and *social robot* conditions and 13 patients for the *personalised robot* condition. Initially, 18 weeks were considered as the programme duration, in which patients would attend twice per week. However, some patients missed therapy sessions, correspondingly, this initial policy resulted in a shorter programme for the patients (23-33 sessions). Hence, we revised this policy in 2018 to improve the CR programme offered to the patients for instead lasting 36 sessions. In addition, we introduced the "dropout criteria", such that if the patient does not attend three sessions in a row without a justification, the patient is dropped from the study (i.e., they would continue the CR programme without the robot). In addition, there were patients that could not complete the programme due to a critical health condition, funding, or because of the outbreak of COVID-19, which halted all the cardiac rehabilitation programme in FCI-IC in March 2020. We considered this situation as an "incomplete therapy", since these reasons were beyond the control of the patients, and were not related to motivation. Correspondingly, we had 26 patients that actively

Table 9.1: Demographic data of the patients who have finished the outpatient phase of the cardiac rehabilitation programme within the study.

	Control	*Social Robot*	*Personalised Robot*
Participants	9	11	6
Gender	9 males	10 males, 1 female	6 males
Age (years), mean (SD)	56.6 (7.8)	55.7 (11.2)	60.3 (6.5)
Body Mass Index, mean (SD)	26.2 (2.6)	29.2 (3.9)	25.0 (2.1)
- Obese		54.5%	
- Overweight	66.7%	36.4%	50%
- Healthy weight	33.3%	9.1%	50%
Level of education			
- Elementary school degree	22.2%	18.2%	16.7%
- High school degree	22.2%	27.3%	
- Technologist		18.2%	
- Bachelor's studies/ degree	55.6%	18.2%	50%
- Postgraduate studies/ degree		18.2%	33.3%

participated in the rehabilitation and completed the outpatient phase as established by the medical team, as presented in Table 9.1.

The patients who attended the study did not have any visual, auditive, cognitive or physical impairments. Only patients with acute myocardial infarction, percutaneous coronary intervention, coronary artery bypass graft, valve replacement, ischemic heart disease, hypertension and ejection fraction greater than 40% were recruited for the study to allow comparability between the patients. The patients signed consent forms (presented in Appendix F) to take part in the study.

9.3.2 Measures

A variety of evaluation methods exist for HRI depending on the task at hand (see Steinfeld et al. (2006) and Bartneck et al. (2019) for extensive lists of approaches and methodologies). Combining different types of methods, based on objective and subjective measurements increase the quality and the reliability of the data (Ganster et al., 2010), hence we decided to use both *quantitative* (i.e., physiological analysis) and *qualitative* (i.e., questionnaires and video material) methods. All of the questionnaires applied in the study can be found in Appendix F.

According to the medical team in FCI-IC, the most important parameters that determine the therapy progress of a patient during CR are the heart rate (HR) during the *training*, and the recovery heart rate corresponding to the difference between the heart rate at *training* and the heart rate acquired one minute after *cooldown* begins. Additionally, the Borg scale (BS) and the blood pressure (BP) are also important to observe the exertion level and the effects of the exercise throughout the programme.

For socially assistive robotics, adapted Unified Theory of Acceptance and the Use of Technology (UTAUT) (Heerink et al., 2010; Venkatesh et al., 2003) questionnaire (also known as the Almere model) and the Working Alliance Inventory (WAI) (Horvath & Greenberg, 1989) are commonly used to evaluate the task performance of a robot. UTAUT evaluates the key aspects of a socially assistive therapy through several concepts. An adapted version of UTAUT questionnaire, in addition to open questions, were applied to the patients who completed the *social robot* condition, in the constructs of perceived utility (PU), usefulness (U), sociability (SP), trust (T), social presence (SP) and safety (S), as presented in Table F.1 in Appendix F (more details of the study and the constructs can be found in Casas et al. (2019)). In order to evaluate the perception of the robot with patients without any prior experience with robots or our system, a *focus group* was formed of 20 patients in their early outpatient or maintenance phase, without any prior experience with robots or our system. This *focus group* did not include the patients in the *control* condition, because we did not want to affect the expectations and perceptions of the patients in that condition. A debriefing was made about the systems, their benefits and the parameters measured in the system, followed by a video presentation of the *social robot* condition. The UTAUT questionnaire was applied after this presentation. Similarly, another focus group was formed of 15 clinicians (e.g., nurses, occupational therapists, physiatrists) in FCI-IC who did not have prior experience with the robot. The clinicians filled the questionnaire before being described the benefits of the socially assistive robot by the research group in Colombia. Afterwards, a discussion was conducted with the clinicians to determine the needs, challenges, modifications, and improvements that can be developed in the interface.

For the *personalised robot* condition, additional questions were developed, in the constructs of perceived usefulness (U), utility (PU), enjoyment (PE), sociability (PS) and adaptivity (PA), social presence (SP) and attitude (A), for the UTAUT questionnaire to measure the perceived personalisation features, as shown in Table 9.2. The questionnaire was applied after the final session of the patient in both robot conditions. In addition, we adapted the WAI questionnaire (Table F.2 in Appendix F) and applied it in the personalisation

Table 9.2: Additional questions developed for the Unified Theory of Acceptance and the Use of Technology (UTAUT) questionnaire to evaluate the perceptions of the patients specific to the *personalised robot* condition, in terms of perceived usefulness (U), perceived utility (PU), perceived enjoyment (PE), perceived adaptivity (PA), perceived sociability (PS), social presence (SP) and attitude (A).

Construct	No.	Question
U	1	I feel encouraged to come to the sessions.
	2	I feel engaged in the therapy.
	3	I feel that the robot helped me progress in my therapy.
	4	I feel encouraged about my therapy when the robot comments on my session performance.
PU	1	The robot recognises me correctly.
	2	The robot remembers my previous sessions correctly.
	3	The robot tracks my session performance correctly.
PE	1	I am pleased that the robot recognises me.
	2	I am pleased that the robot uses my name.
	3	I am pleased to hear about my therapy progress.
	4	I am pleased that the robot remembers me.
	5	I am pleased to work with the robot.
PA	1	I feel that the robot personalises its interaction.
PS	1	I feel that the robot knows me well.
SP	1	I feel that the robot has a personality.
	2	I feel compelled to come to the sessions because the robot comments on my absence.
A	1	I feel attached to the robot.

condition at the middle and end of the study to analyse the long-term perception of the robot by the patient. The WAI questionnaire measures the performance of the robot in terms of *Task*, *Bond*, and *Goal*. It uses negative (e.g., *I feel uncomfortable with the help of the robot.*) and positive (e.g., *The robot perceives my objectives of the rehabilitation properly.*) formulations to limit the bias in the results. Additionally, we recorded the sessions in the *personalised robot* condition to evaluate the interaction with the robot more closely. The analysis of the video was performed by an independent coder, who labelled various types of interactions (i.e., medical staff interaction, response to robot requests, posture correction, gaze to the robot, and social interaction) based on a previously-established protocol. In contrast, we do not have video recordings of the *control* or *social robot* conditions due to the varying ethical concerns at the beginning of the study and the lack of available resources. WAI was not applied to the other conditions.

We use Bayesian analysis, as previously used in Chapter 7, for comparing patients' progress and perceptions over the duration of the programme in the *personalised robot* condition and for analysing the differences between the conditions. Bayesian analysis is a powerful tool to inspect the main effects in modelling the data, especially when the sample size is low (Biel & Friedrich, 2018). A null hypothesis is formed and the alternative hypothesis is compared against the null using the Bayes factor (Jeffreys, 1939) to determine whether the results suggest a tendency towards one or the other. We use the R package BayesFactor[3] (Morey & Rouder, 2018) and, unless otherwise stated, default parameter values are used. When reporting the Bayes factors (BF), we apply the methodology of Navarro (2018) in which the level of evidence[4] for or against a main effect of a variable in modelling the data is presented. In other words, the presence of a main effect suggests that the data varies due to the variable, whereas, evidence against an effect suggests that the data does not depend on the variable. Our study is a long-term study with the same measurements taken at each session with three conditions. Correspondingly, we use a Bayesian one-way repeated measures analysis of variance (ANOVA) for evaluating the progress of the patients throughout the CR programme within the *personalised robot* condition and a Bayesian two-way mixed-factor ANOVA (i.e., the combination of repeated measures and between-subjects) for comparing between the conditions. The random factor is the patient and the independent variables are the stage of the programme (e.g., the first stage corresponds to sessions 1-4, the second stage for sessions 5-8) as the within-subjects factor and the condition (i.e., *control*, *social robot*, *personalised robot*) as the between-subjects factor. The assumptions of the ANOVA[5] are validated. We compared questionnaires using Wilcoxon signed rank test and Mann-Whitney U-test for WAI and UTAUT, respectively, as Likert scales are ordinal, and applying ANOVA may produce incorrect results, in addition to violations in the assumptions of the ANOVA test. For comparing all conditions, we applied a robust two-way mixed ANOVA using the trimmed means, using WRS2 package in R (Wilcox, 2017) (*bwtrim* function), because the ANOVA assumptions were violated due to the varying variances between conditions and the presence of non-normality in

[3] `https://richarddmorey.github.io/BayesFactor/#mixed`

[4] Jeffreys (1939)'s levels of evidence for an effect: BF > 100 is *decisive*; 30-100 is *very strong*; 10-30 is *strong*; 3-10 is *moderate*; 1-3 is *anecdotal*. BF $=1$ is no conclusive evidence for or against an effect. Levels of evidence against an effect: 0.33-1 is *anecdotal*; 0.1-0.33 is *moderate*; 0.033-0.1 is *strong*; 0.01-0.033 is *very strong*; BF <0.01 is *decisive*.

[5] Assumptions of ANOVA: (1) Continuous dependent variable, (2) Normally distributed dependent variable and residual for all combinations of factors, (3) Homogeneity of variance between conditions/stages, (4) No outliers. Note that the central limit theorem in statistics states that, given a sufficiently large sample size (generally greater than 30), the sampling distribution of the mean for a variable will approximate a normal distribution and ANOVA is robust to violations of this assumption. In the presence of outliers, the results were compared to the analysis without outliers, but no considerable differences were found in terms of evidence levels or ANOVA. Hence, ANOVA still holds valid and we report the values for analysis with the outliers. Outliers can be observed from the boxplots.

the data resulting from low and unequal group sizes, which may have resulted in lower significance values in these comparisons.

9.4 Results

As previously mentioned, not all patients who enrolled in the CR programme completed it. Figure 9.23 shows the proportion of therapy status of patients in each condition. While the *control* condition has finished, one patient in the *social robot* condition and six patients in the *personalised robot* condition could not complete the programme due to the outbreak of COVID-19. The patient in the *social robot* condition had three sessions left, and the patients in the *personalised robot* condition had less than 18 sessions (half the programme). At the time of writing of this thesis, CR programme (conventional or with our system) is ceased in FCI-IC, since patients with cardiovascular diseases carry a high risk from the virus.

As this work focuses on personalisation, we will do an in-depth analysis of a patient in the *personalised robot* condition, which will allow us to interpret the factors that affect the perception of a patient. We will then compare the case study to the other patients in the *personalised robot* condition to evaluate the generalisability of the findings, and finally compare the *personalised robot* condition to the other conditions.

9.4.1 Personalised Robot: A Case Study

While a quantitative analysis enables examining the overall effects of a study, it prevents detecting the reasons behind the perception of users. Since our work focuses on the individuality and the personalisation of a long-term therapy, a case study of a patient can enable an in-depth look into each of the sessions and see how the interaction between the robot and the patient occurred and varied throughout time. That is why, this section describes the case study of a male patient (60 years old, body mass index: 25.7 - overweight, high school degree - studying towards a Bachelor's degree) in the *personalised robot* condition. The patient was diagnosed with myocardial infarction and underwent an angioplasty procedure. After being discharged from the hospital, the patient started the outpatient phase of the CR programme.

Within the first session of the outpatient phase, the patient was informed about the purpose of this study and the role of the robot. Upon this information, the patient agreed to take part in the study, hence, throughout the remaining 35 sessions of the outpatient phase, the personalised robot was present and took an active role in supporting and monitoring the session's progress. The 13th session is of note. During that session the patient experienced fatigue, and his heart rate was very high over a critical threshold. This was detected by the robot and the medical staff were called over for an intervention. This patient was referred to the Emergency Room and a percutaneous transluminal coronary angioplasty plus stent was performed.

This section will analyse the overall physiological progress of the patient along with an in-depth look at this "critical" session. In addition, the results of the WAI and UTAUT questionnaires and the video analysis are presented in order to observe the long-term perception of the patient.

9.4.1.1 Physiological Progress

While the *relative intensity* of the sessions has progressively increased (visible in Figure 9.6), the *very mild* exertion level perceived by the patient and the physiological progress of the patient (Figure 9.7) show a positive outcome regarding the patient's cardiovascular health and the success of the CR programme. As we previously indicated, the most important parameters showing the physiological improvement of a patient in CR is the training heart rate (HR, Figure 9.7a) and the recovery heart rate (R-HR, 9.7b). The average heart rate of the patient mostly stayed below the critical level (120 bpm) that corresponds to the robot alerting the medical staff. Moreover, the threshold for the critical value was increased by the physiatrists, showing that the physical fitness of the patient improved. Recovery heart rate was higher at the end of the programme than the first session, which the physiatrists identified as an important improvement on the patient's health. In addition, the systolic blood pressure (BP) (Figure 9.7c) was maintained in a safe range (110-130 mmHg as determined by the physiatrists) in most of the sessions.

However, the number of alerts in a session (Figure 9.8) lights a different perspective than the average heart rate and Borg scale. While the average heart rate hardly increased over the critical limit throughout the programme, in 9 out of 35 sessions[6], the medical staff were alerted to help the patient, which was critical to the therapy.

[6]The number of alerts in the 24th session corresponds correctly to the detected critical rates, but in session 27, there was an excessive number of alerts due to a problem with the sensors of the robot.

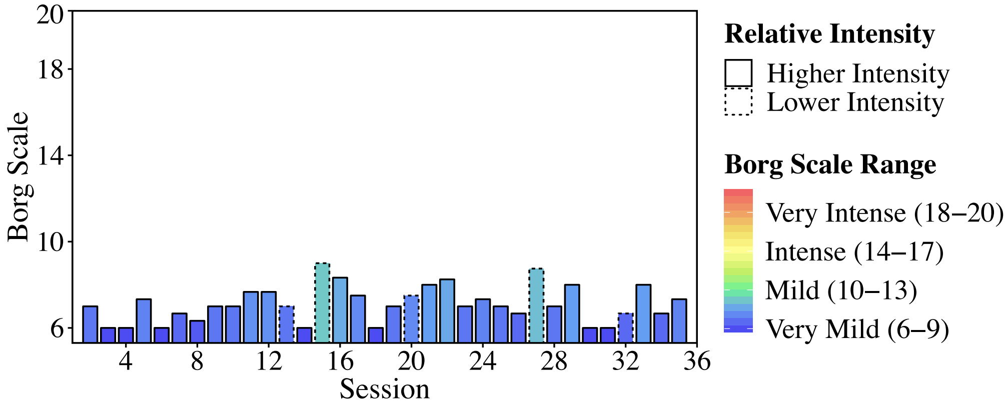


Figure 9.6: Exertion levels (Borg scale) and relative intensity of sessions during the cardiac rehabilitation programme. The physiatrists aim to achieve very mild or mild levels of Borg scale.

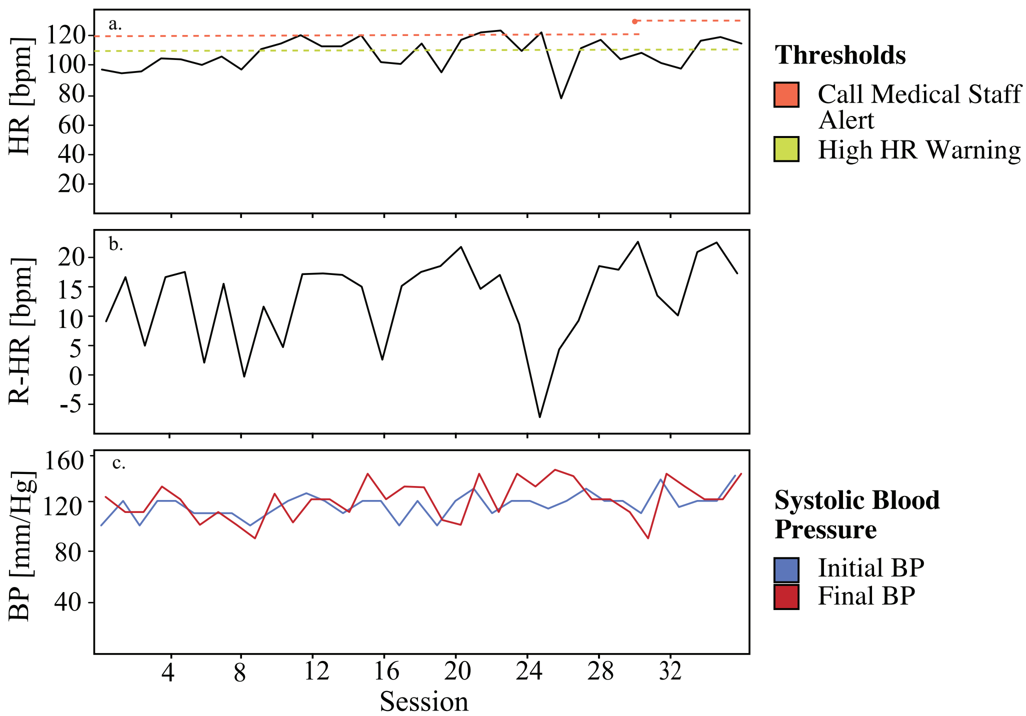


Figure 9.7: Physiological evolution of the patient during 35 sessions: (a) Average heart rate (HR) during *training*, (b) Recovery heart rate (R-HR) and (c) Systolic blood pressure (BP). The patient mostly stayed below the critical heart rate (corresponding to the call medical staff alert). The recovery heart rate was higher at the end of the programme than the first session, which shows an improvement on the patient's health. The systolic BP was maintained in a safe range (110-130 mmHg) in most of the sessions.

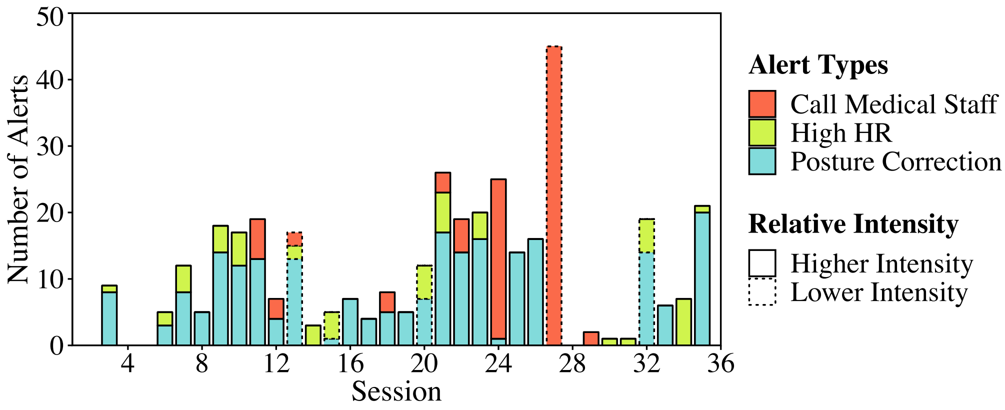


Figure 9.8: Robot alerts for the heart rate and cervical posture during the cardiac rehabilitation programme. These alerts show that continuous monitoring is vital in cardiac rehabilitation programme.

9.4.1.2 The Critical Session

The 13th session is an important example of the interest of continuous monitoring of patients and social robots. In this session, the patient had a higher heart rate compared to the previous ones, crossing a critical threshold set by physiatrists on two instances which resulted in two calls to the medical staff (red dotted points in Figure 9.9). The prompt alert of the robot helped the physiatrists as a medical tool to immediately detect the complication, such that they can instantly intervene by decreasing the intensity of the exercise. In addition, the alerts in the previous sessions may have increased the awareness of the physiatrists to detect the complication. Upon the intervention, the patient reported to the physiatrists, feeling dizzy and very tired and continued the session with low intensity exercises to progressively decrease the heart rate, as required. Following this session, the patient was referred to the Emergency Room and a percutaneous transluminal coronary angioplasty plus stent was performed.

Relying on objective data is important in such situations as self-reports might be biased and hide underlying conditions: here the patient reported a *very mild* Borg scale (7), which contradicts his high heart rate and what he told the physiatrists at the intervention. Throughout this session, the number of posture corrections was also relatively high, which may also originate from the high exertion level and dizziness of the patient.

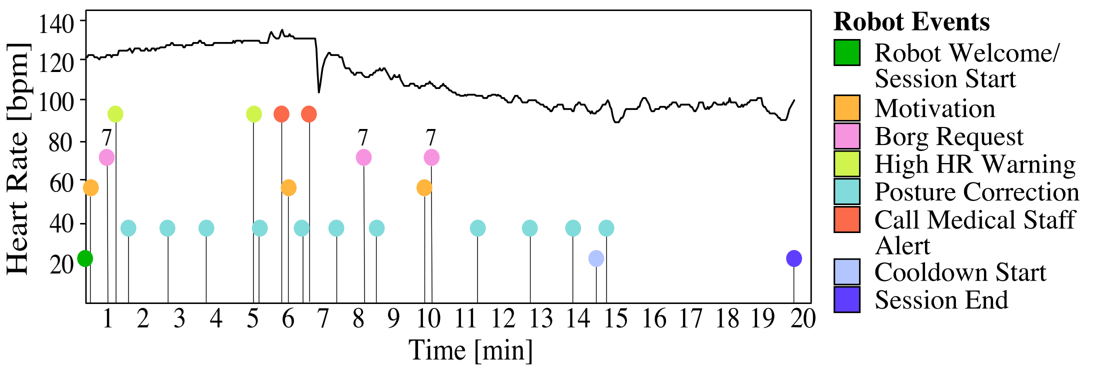


Figure 9.9: Patient's heart rate and robot's feedback during the "critical" session. The physiatrists intervened after the *call medical staff* alert to reduce the exercise intensity, which decreased the heart rate of the patient. The patient was feeling very tired and dizzy, which may have resulted in a high number of posture corrections. Note that the exertion level (Borg scale) is reported to be *very mild* (7) in contrast to the alerts.

9.4.1.3 Long-term Perception of the Robot

Qualitative data was collected using the WAI questionnaire at the middle and end of the study, in order to analyse the long-term perception of the robot by the patient (Figure 9.10). Results showed that the perception for *Goal*, *Task* and *Bond* was maintained highly positive throughout the outpatient phase of the CR programme, supporting our objectives with personalisation of the robot.

In the case of negative formulation, the results show a decrease in *Task* and *Goal* category. These outcomes show that the negative perception of the robot decreases with time.

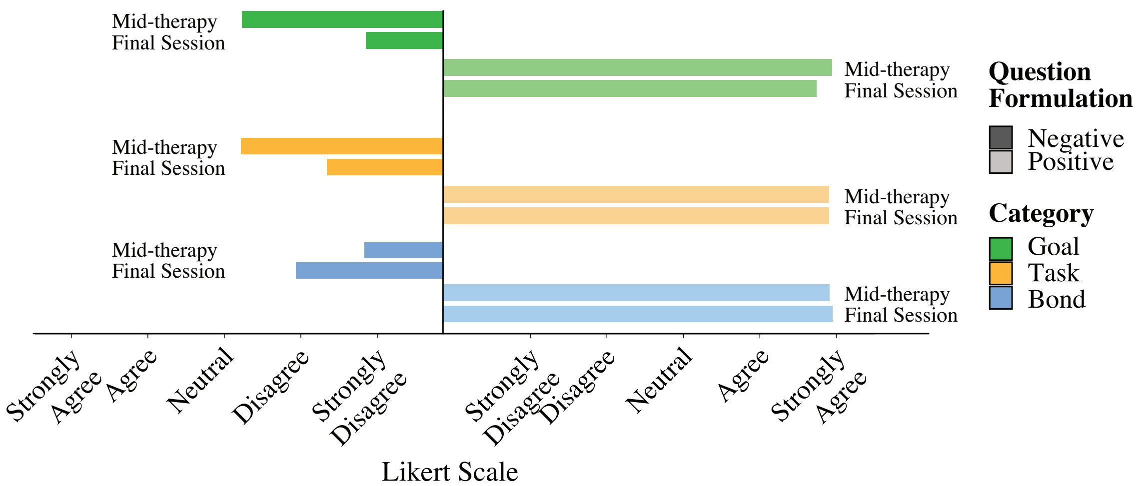


Figure 9.10: Working Alliance Inventory (WAI) responses evaluated at the middle of the therapy and the final session. The result show that the perception for *Goal*, *Task* and *Bond* was maintained highly positive throughout the cardiac rehabilitation programme. The negative perception of *Task* and *Goal* decreased with time, whereas the negative perception of *Bond* increased. The patient noted feeling that the robot would not cooperate with the patient if he did not comply to its requests, which decreased his bond with the robot.

The patient is less confused by the tasks (corrections) made by the robot during the rehabilitation, and the patient believes that the time spent with the robot is of increasing value over the duration of the CR programme. In contrast, the negative perception of *Bond* increased. Detailed analysis of the open-question responses showed that the patient felt that the robot would no longer cooperate with him during therapy if he did not make the corrections requested by the robot (e.g., cervical posture corrections and heart rate warnings).

The UTAUT questionnaire results show that the patient perceived the robot and the therapy as highly positive: strongly agreeing with 92.9% questions and agreeing with the rest. The responses of the patient to the open questions in UTAUT support that the robot was perceived very positively: *"I would recommend using the robot as it is a great help during the cardiac rehabilitation programme."*, *"I would not change anything about this system. The robot interacts in a positive way with me, it helps me along with the medical staff, and it is also a good tool for them."*.

9.4.1.4 Interaction with the Robot

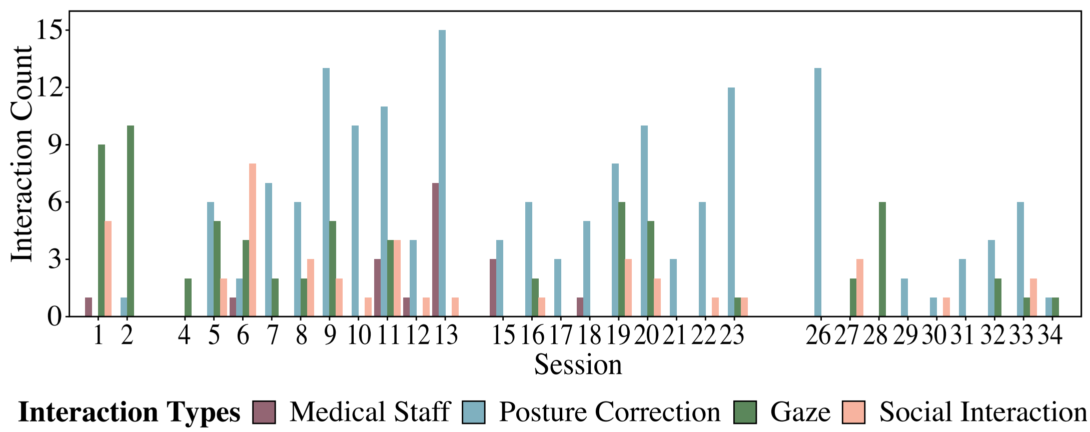


Figure 9.11: Interaction results of the video analysis based on 30 sessions (note the explanation of the missing sessions[2]). The results suggest that the patient socially interacted and looked at the robot throughout the cardiac rehabilitation programme. The patient corrected his posture without fail upon a simple prompt by the robot.

We analysed 30 recorded sessions[2] to observe the interactions with the robot throughout the CR programme (Figure 9.11). Four types of interactions were analysed by an independent coder:

- *Medical staff*: the medical staff interacts with the robot either through responding to

[2]Results for sessions 3, 14, 24, 25 and 35 are missing due to technical problems with the recordings.

the robot's request for intervention during a critical high heart rate, or talks to the robot when checking the patient,

- *Posture correction*: the patient corrects posture upon the robot's request,
- *Gaze*: when the patient looks at the robot to pay attention,
- *Social interaction*: the patient's verbal (e.g., thanking the robot after the personalised progress feedback given at the end of the session) and non-verbal (e.g., gesture to the robot after motivational feedback, or touch the head of the robot) communications with the robot, and talking about the robot to other patients.

The medical staff worked collaboratively with the robot to intervene in the sessions when necessary, such as for the 13th session which is discernible in Figure 9.11, or to change the exercise intensity. The infrequent interactions indicate that the medical staff found the robot reliable and trusted it as a tool of monitoring the patient adequately and supporting the CR programme, which is in line with the perceptions of the clinicians from the focus group (Casas et al., 2019).

Regarding the number of alerts for cervical posture correction (in Figure 9.8), the video analysis showed that the patient corrected his posture without fail upon a simple prompt by the robot.

While the *gaze* and *social interaction* were higher at the initial sessions due to possibly the novelty effect and the adjustment process to the technology, these interactions also occurred throughout the later sessions, which suggest that the patient did not lose interest in the robot throughout the long-term CR programme. Moreover, the video analysis showed that the patient is very focused on the exercise as expected, hence, the patient could mostly look at the robot at the beginning of the treadmill exercise and at the end of the *cooldown* stage, which could have resulted in the low number of *gaze* and *social interactions* depending on the session intensity. The patient actively socially interacted with the robot or talked to the others about the robot through several ways:

- talking to other patients about the robot's role and its benefits,
- mirroring the robot's gesture to the *Call medical staff* alert,
- reacting positively (e.g., smiling or thanking the robot) to the motivational feedback of the robot,

- touching (caressing) the robot, which has occurred (once) at the end of a session after the robot "sighed" going into sleep mode,
- reacting negatively to the robot, in the case of misidentifications from user recognition, posture correction (once) and alert to the physiatrist (once).

A detailed analysis of the perceived personalised aspects of the system showed that the patient positively reacted to the positive progress motivation given at the end of the session for having a lower number of alerts in the session with *"Let's keep up the good work"*. The patient verbally thanked the robot in three of these cases, which did not happen towards the other motivations of the robot derived from the non-personalised robot.

The user recognition performed poorly (14 out of 38[7] times), due to the malfunctioning face recognition (13 times) due to blurry images and cropped faces in the picture, including the first (enrolment) session of the user. Since the user recognition (MMIBN:OL) uses online learning, this negatively affected the performance of the system. The video analysis shows that out of 9 misidentifications, the patient reacted negatively only twice (i.e., he did not react the other times). The negative reactions of the patient to the misidentifications in user recognition may further support the social agent perspective since they indicate that the patient disapproves the behaviour of the robot. It is important to note that the patient strongly agreed with the question "The robot recognises me correctly." regardless of the low performance of the user recognition, which is suggestive of our previous finding in Chapter 7 that personalisation mitigates the negative user experience.

Additionally, the video recordings showed that the patient responded to all the *Borg scale* and *High HR* warnings of the robot through the tablet, because of the compulsory structure of the requests. Figure 9.12 shows that initially, the patient had difficulty in interacting with the system, but he quickly adjusted to the system.

Finally, the analysis of the patient's attendance shows that the patient attended the CR sessions twice per week, as required, in 13 out of 18 weeks, missing more sessions in the beginning, which suggest that the personalised attendance tracking of the robot may have improved the adherence to the CR programme.

[7] The patient was recognised more than the number of sessions due to the system restarts to overcome sensor connection problems during the session.

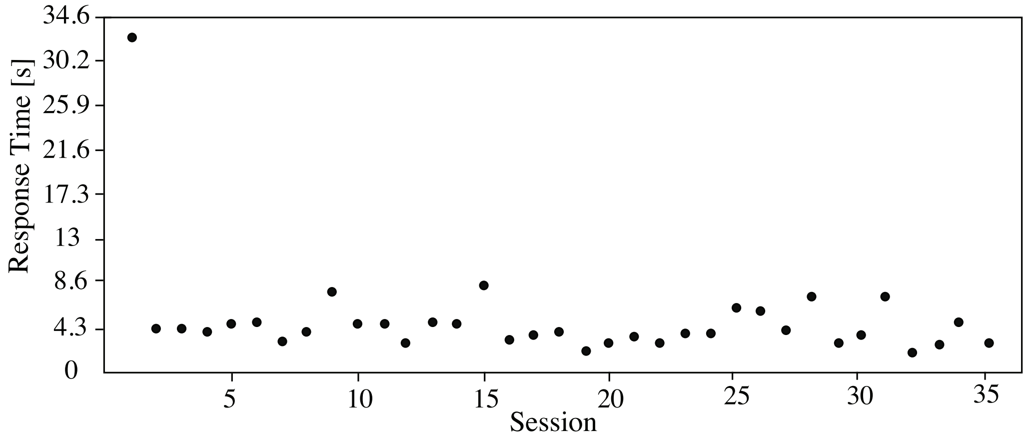


Figure 9.12: Exertion level (Borg scale) request response time throughout the cardiac rehabilitation programme, showing that the patient quickly adjusted to the system, after the initial difficulty.

9.4.2 Personalised Robot: Comprehensive Analysis

While the in-depth analysis of a patient's complete CR programme allowed us to understand the factors that affected the perception of the patient, a comprehensive analysis of the six patients that finished the study will help understanding whether these findings persist in other patients. In order to reduce the variability between the sessions due to confounding factors that affect the patients' health during the sessions, we grouped the sessions by four. The patient that we analysed in detail in the previous section is referred to as the patient 3 (P3) in this section.

9.4.2.1 Physiological Progress

Figure 9.13 shows that the training heart rate tended to increase for all patients due to the increasing intensity of the sessions. The Bayesian analysis supports the decisive evidence

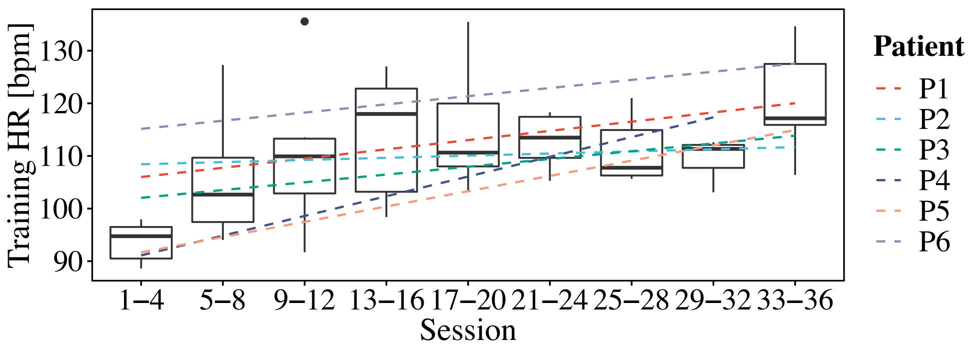


Figure 9.13: The training heart rate throughout the cardiac rehabilitation programme for patients in the *personalised robot* condition. Linear regression lines for each patient suggest that the training heart rate mainly increased due to the increasing intensity of the sessions.

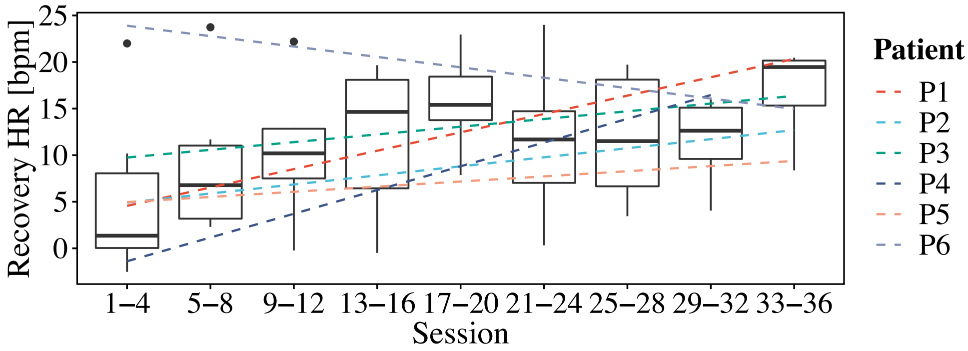


Figure 9.14: The recovery heart rate (R-HR) throughout the cardiac rehabilitation programme for patients in the *personalised robot* condition. Linear regression is applied to the recovery heart rate progress for each patient. Higher recovery heart rate is better. The results show that 5 out of 6 patients fully improved their recovery heart rate compared to their initial values, which correspond to a successful completion of the cardiac rehabilitation programme.

for the main effect of the session ($\mathsf{BF} = 110.1$) and very strong evidence for the effect of the patient ($\mathsf{BF} = 65.3$), suggesting that the training heart rate varies from patient to patient.

Figure 9.14 shows the corresponding recovery heart rate (R-HR). This result shows that the majority of patients (5 out of 6) fully improved their recovery heart rate compared to their initial values, which corresponds to a successful completion of the CR programme. There is very strong evidence for main effects of patients ($\mathsf{BF} = 58.3$), however, anecdotal (weak) evidence against an effect of the programme stage (session) ($\mathsf{BF} = 0.79$). This result arises because of the sixth patient (P6), as the plot shows. In addition, an analysis of the data of the remaining patients shows that there is moderate evidence ($\mathsf{BF} = 6.4$) for the main effect of stage and anecdotal evidence ($\mathsf{BF} = 0.34$) against an effect of the patient, suggesting that the 5 patients similarly recovered throughout the CR programme. The decrease of recovery heart rate for P6 may arise from the relation between the (higher) training heart rate, which may have caused more difficulties in the CR programme and for retaining the recovery heart rate.

Similar to the P3, the *mild* exertion levels (Borg scale) of each patient (Figure 9.15) and the corresponding number of alerts (Figure 9.16) do not fully agree. While the Borg scale never reached the *intense* level (>13) in any session and there is strong evidence ($\mathsf{BF} = 0.08$) against an effect of the stage, in contrast to the increasing session intensity. Moreover, high heart rate warnings and critical heart rate alerts happened frequently in some of the sessions, which allowed the medical to immediately intervene by changing the exercise intensity. It is important to note that while the average high heart rate warnings ($\mathsf{M} = 1.14$, $\mathsf{SD} = 2.02$) and critical heart rate alerts ($\mathsf{M} = 1.47$, $\mathsf{SD} = 4.93$) are very low throughout the

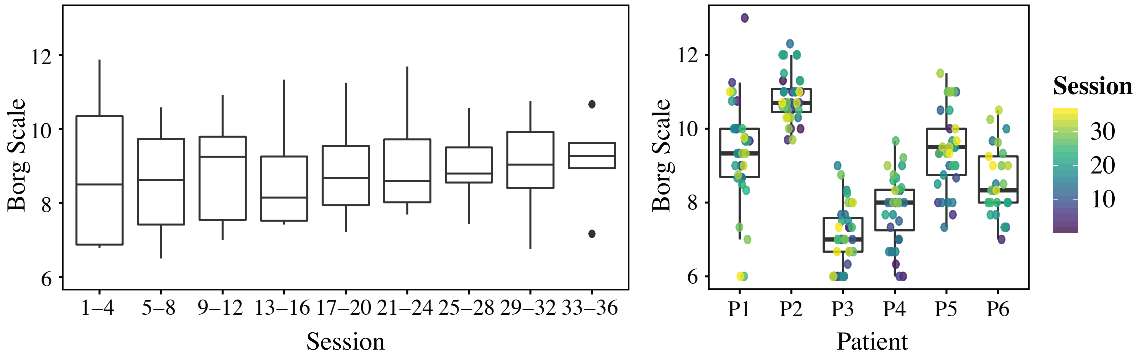


Figure 9.15: Exertion level (Borg scale) of patients in the *personalised robot* condition, which show that the patients had *very mild* (6-9) or *mild* (10-13) levels throughout the cardiac rehabilitation programme.

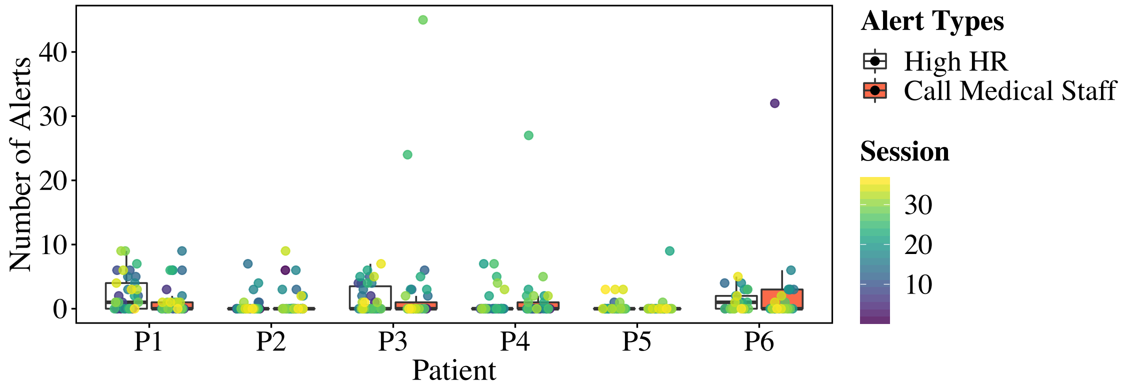


Figure 9.16: Number of high training heart rate (HR) and critical heart rate (call medical staff) alerts of the patients in the *personalised robot* condition. In contrast to low exertion levels, there are quite a few number of critical alerts in some of the sessions, which allowed the medical team to immediately intervene.

CR programme, even a single alert is vital for detecting a critical condition, as we have previously shown in Section 9.4.1.2.

9.4.2.2 Long-term Perception of the Robot

We can analyse how the patients' overall perception of the robot and the therapy changed over time through the WAI questionnaire (presented in Figure 9.17). Since the data is ordinal and violates the normality assumption (and has a low sample size), we apply a dependent two-group Wilcoxon signed rank test with Bonferroni correction, because the patients responded to the WAI at different times (i.e., paired measures, testing non-equality and "greater" hypotheses). There is a significant improvement between the perceived goal construct in the positive formulation ($p = 0.003$, $r = 0.38$ - moderate effect size, $V = 42$) from mid-therapy test to the final session. In the other constructs, there is no significant

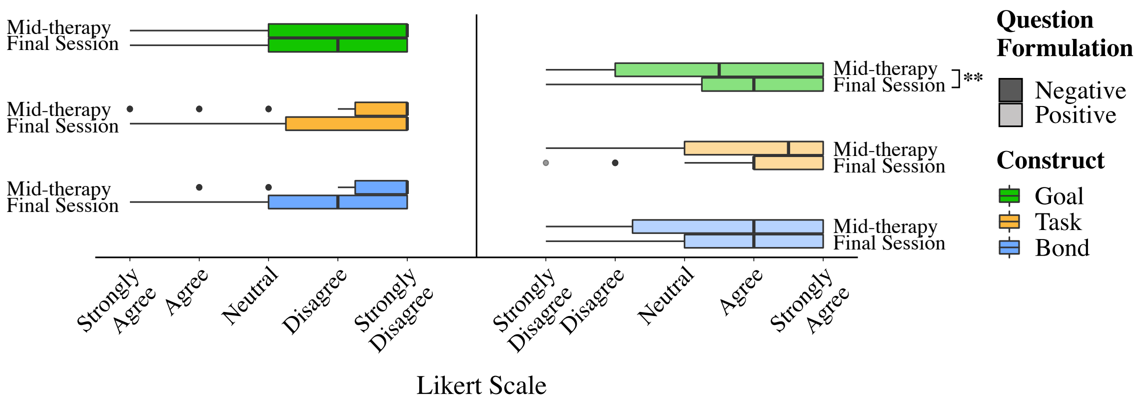


Figure 9.17: Working Alliance Inventory (WAI) responses for the *personalised robot* condition, evaluated at the middle of the CR programme (18th session) and the final session. The results suggest that the patients' positive perception of the robot and the therapy was maintained over the cardiac rehabilitation programme, in addition to a significant improvement of the perceived goal construct in positive formulation ($p = 0.003$, $r = 0.38$ - moderate effect size, $V = 42$).

difference between the tests ($p > 0.05$). The positive formulation of *Bond*, *Goal* and *Task* all show that the robot and the therapy were generally positively perceived, and the patients kept their bond with the robot over the duration of the CR programme. Moreover, the patients generally disagree with the negative formulations (e.g., "I feel uncomfortable with the robot."), showing that in addition to highly positively perceiving the robot, the majority does not negatively perceive it.

UTAUT questionnaire (Figure 9.18) confirms our previous findings that the personalised robot was mostly positively perceived by the patients, in terms of perceived usefulness (U: M = 4.31, SD = 0.63), ease of use (EU: M = 4.27, SD = 0.71), utility (PU: M = 3.83, SD = 0.99), safety (S: M = 4.5, SD = 0.45), trust (PT, M = 4.25, SD = 0.71), enjoyment (PE: M = 4.28, SD = 0.71) and attitude (A: M = 3.83, SD = 1.6). The adaptivity (PA: M = 3.5, SD = 1.38), sociability (PS: M = 3.57, SD = 1.05), and social presence (SP: M = 3.05, SD = 1.1) was mostly perceived as neutral. Note that some of these constructs presented here include specific personalisation questions, as presented in Table 9.2. For instance, the perceived enjoyment construct evaluates the enjoyment (e.g., "I am pleased that the robot remembers me.") of the personalisation features (i.e., recognition, referring to the patient with the name, tracking and referring to the therapy progress, remembering the user), hence, the very high score of the patients shows that the personalisation was perceived very positively. Moreover, the questions evaluating user engagement (M = 4.83, SD = 0.41), motivation for adherence (M = 4.83, SD = 0.41) and motivation arising from referral to the patient's therapy progress (M = 4.17, SD = 1.17) also show the robot was able to facilitate user engagement and adherence.

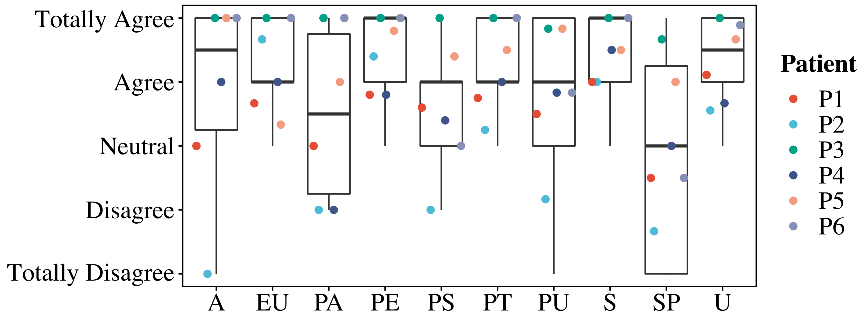
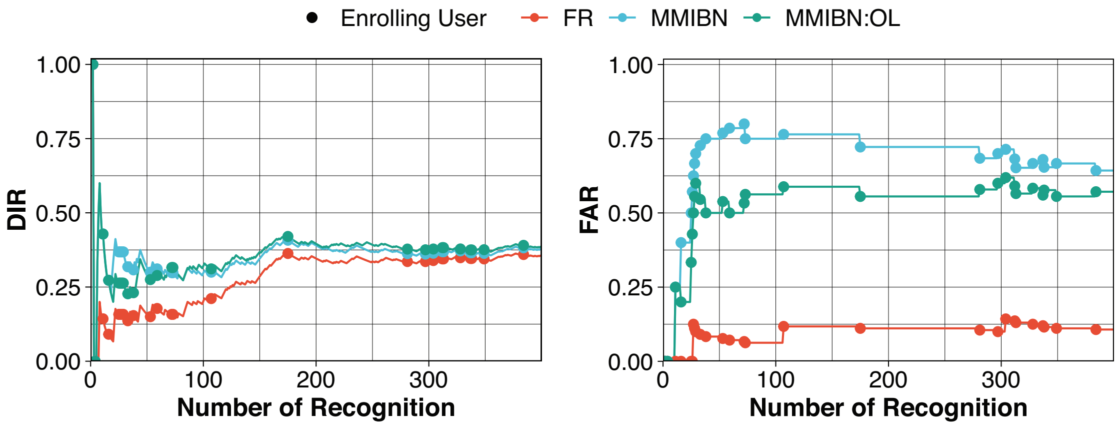


Figure 9.18: UTAUT questionnaire results for the *personalised robot* condition. The average of the questions per construct for each patient is denoted with a coloured point on the plots. The questions are formulated positively. The results show that the *personalised robot* was positively perceived by the patients, in terms of high perceived usefulness (U), ease of use (EU), utility (PU), safety (S), trust (PT), and enjoyment (PE). Patients responded mostly neutrally for perceived adaptivity (PA), perceived sociability (PS) and social presence (SP) questions.

Within open questions, the patients noted the need for improving the robustness of the user recognition and sensors. The user recognition performance was indeed very low (overall DIR= 0.38, FAR= 0.56 for MMIBN:OL), which, as we previously noted, was due to the failures arising from face recognition (DIR= 0.35, FAR= 0.11) that caused online learning to decrease the performance. However, in contrast to the patient in the case study, the MMIBN:OL performed better overall than face recognition (FR) throughout the duration of the study, as can be seen in Figure 9.19. The figure also shows how the non-adaptive model, Multi-modal Incremental Bayesian Network (MMIBN), would have performed over the data (DIR= 0.36, FAR= 0.67). The results are in contrast to our study in Chapter 5, showing that MMIBN:OL performs slightly better than MMIBN in identifying known users, and notably better in identifying new users. FR performs considerably better in False Alarm Rate (FAR) than our approaches because it identified most (63%) of known users as new. Both of our proposed approaches perform better in recognising known users than FR, supporting that our proposed user recognition is suitable for real-world interactions, and improves the recognition even when the identifiers are malfunctioning.

Other improvements for the system that the patients suggested include repetitiveness of the phrases, which was also addressed in the previous study with the *social robot*. In addition, one patient found the appearance and the sound of the robot to be childish. Regardless, all patients recommended the system for future patients, and commented on its usefulness, personalisation and effects on user motivation: *"The cardiac rehabilitation*



(a) (b)

Figure 9.19: Detection and Identification Rate (DIR) and FAR of user recognition for the patients in the *personalised robot* condition. There are more number of enrolled users (represented with a dot) than the recruited users for the study due to re-enrolment of some users to the system with a different ID at a later time to overcome the face recognition (FR) errors encountered initially. Higher DIR and lower FAR are better. The results show that online learning (MMIBN:OL) performs better than the non-adaptive model (MMIBN) in both aspects, and both of our proposed approaches outperform FR. FR performs better in FAR due to estimating most users as unknown.

programme with the robot will help you to recover as quickly as possible, and you will be able to progress by being linked to the robot.", "I feel confident in doing the rehabilitation with the robot, because I know that it is personalised and constantly monitoring my performance and progress.", "Working with the robot motivates me.", "Working with the robot makes me feel happy.".

9.4.2.3 Interaction with the Robot

Similar to the patient 3, the compliance of all the patients to the posture corrections is very high, supported with moderate evidence (BF = 0.16), as can be seen from Figure 9.20. In addition, there is decisive evidence (BF = 0.003) against an interaction between session and compliance, showing that the patients comply to the robot's request throughout the CR programme, even though each patient has a different number of posture correction feedback (BF = 4958687, decisive evidence).

Figure 9.21 shows that, similar to the patient 3, while the patients gazed at the robot more frequently initially, in general, gaze has occurred throughout the programme, and there is anecdotal evidence against an effect for the patient (BF = 0.42) suggesting that each patient similarly gazes at the robot. On the other hand, there is very strong evidence (BF = 39.97) that patients differ in their social interactions (as evident in Figure 9.21), but there is moderate evidence against an effect of the session (BF = 0.21), which suggests that

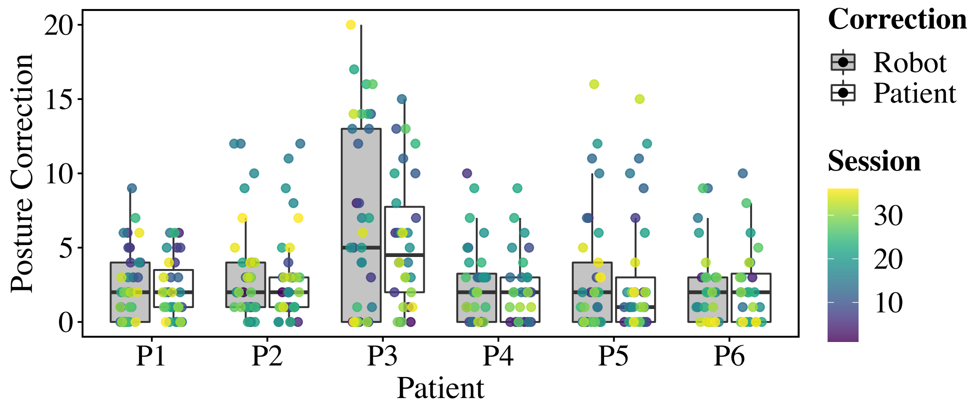


Figure 9.20: Number of posture correction requests by the *personalised robot* and the patients' active posture corrections (i.e., compliance to the request). The results show that all patients complied with the robots requests, regardless of the session.

the patients maintain their social interaction frequency throughout the CR programme. Similar to the patient 3, other patients interacted with the robot in different ways, such as talking to the robot, smiling at being recognised or for motivational or attendance feedback, giving a negative response for an incorrect recognition, or saying "Bye!" to the robot on the patient's last session. The video recordings showed that some of the medical staff also interacted verbally with the robot, such as joking with it, thanking the robot, or

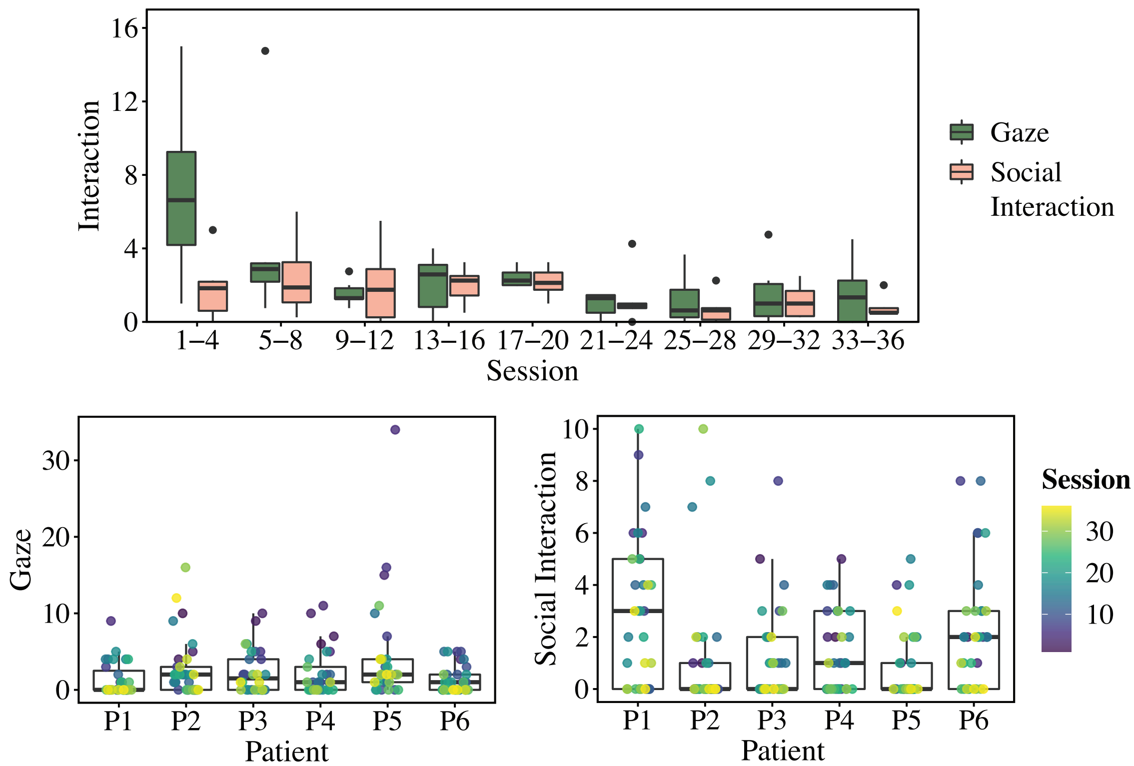


Figure 9.21: *Gaze* and *social interaction* of the patients with the *personalised robot* throughout the cardiac rehabilitation programme. The results indicate that the interactions were maintained in the long-term programme.

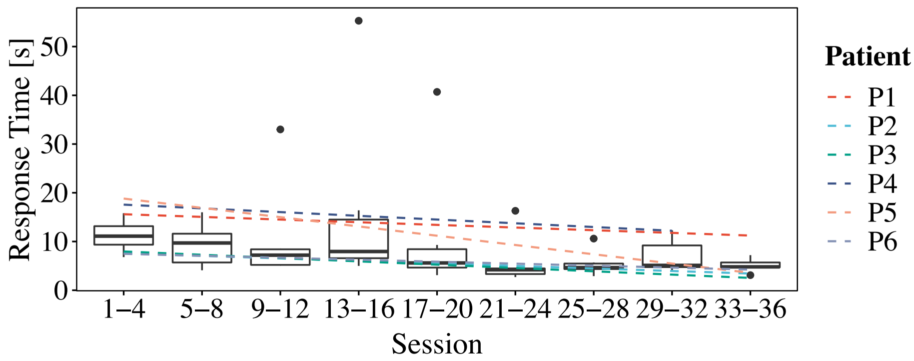


Figure 9.22: Response time of the patients to the Borg scale request of the robot throughout the cardiac rehabilitation programme for the *personalised robot* condition. Linear regression lines per patient suggest that the response time generally decreases throughout the programme.

talking about its benefits to other medical staff.

While the regression lines suggest that the patients adapt to the system over time and give faster responses, as shown in Figure 9.22, there is moderate evidence ($\mathsf{BF} = 0.26$) against an effect of the stages. There is also anecdotal evidence ($\mathsf{BF} = 0.36$) against an effect of the patient, suggesting that the patients interact similarly with the system.

9.4.3 Comparison of All Conditions

In this section, we compare the results of all the conditions in terms of adherence to the CR programme, physiological parameters, cervical posture progress, and the perceptions of the patients. As a result of the dropouts and incomplete therapies, we had a low group size in conditions, which caused violations to the ANOVA assumptions. Thus, we applied robust ANOVA (Wilcox, 2017) for comparisons between conditions, as described in Section 9.3.2.

9.4.3.1 Adherence

This study aimed to improve user motivation and, in turn, adherence to the cardiac rehabilitation programme, which is vital for the complete recovery of a patient with a cardiovascular disease. However, as previously described in Section 9.3.1, there were several patients who could not complete the CR programme due to the reasons beyond the control of the patients (e.g., funding, medical condition, outbreak of COVID-19). Especially the outbreak of COVID-19 prevented 6 patients in the *personalised robot* condition and one patient in the *social robot* condition from continuing their CR programme. Thus, we cannot

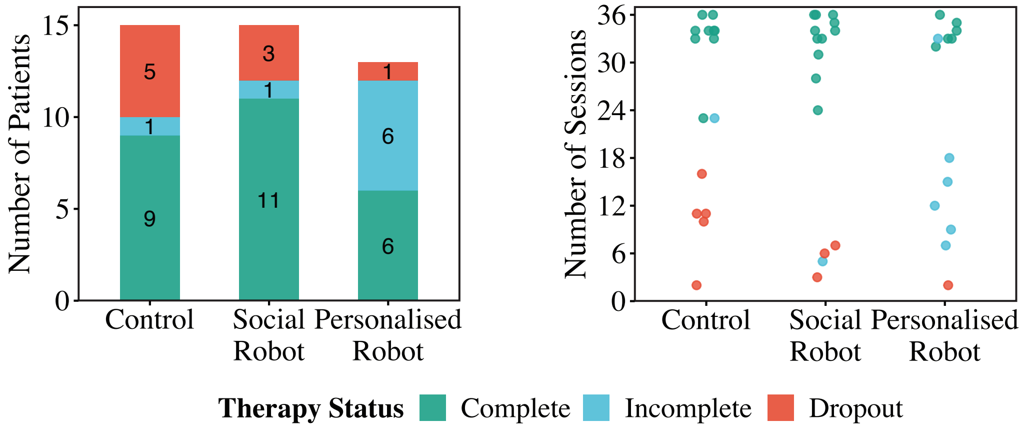


Figure 9.23: The therapy status of the users in the *control*, *social robot* and *personalised robot* conditions: "complete" refers to the completed cardiac rehabilitation programme as determined by the clinicians; "incomplete" is when patients need to stop the CR programme due to reasons beyond their control (e.g., funding, medical condition, outbreak of COVID-19), and "dropout" refers to dropping out of the study or not attending 3 sessions in a row without a justification.

conclude the effects of the *personalised robot* on adherence. The attended sessions per condition (as shown in Figure 9.23) shows that the dropouts occur at earlier stages of the CR programme. Nonetheless, the *social robot* condition has fewer dropouts compared to the *control* condition. Furthermore, the patients in the *personalised robot* condition attended their sessions mostly regularly (twice a week), and have rated the motivation for adherence question highly, as presented in Section 9.4.2.2.

9.4.3.2 Physiological Progress

As mentioned in the previous sections, the training heart rate (HR) and the recovery heart rate (R-HR) are the most important physiological parameters of the CR programme that

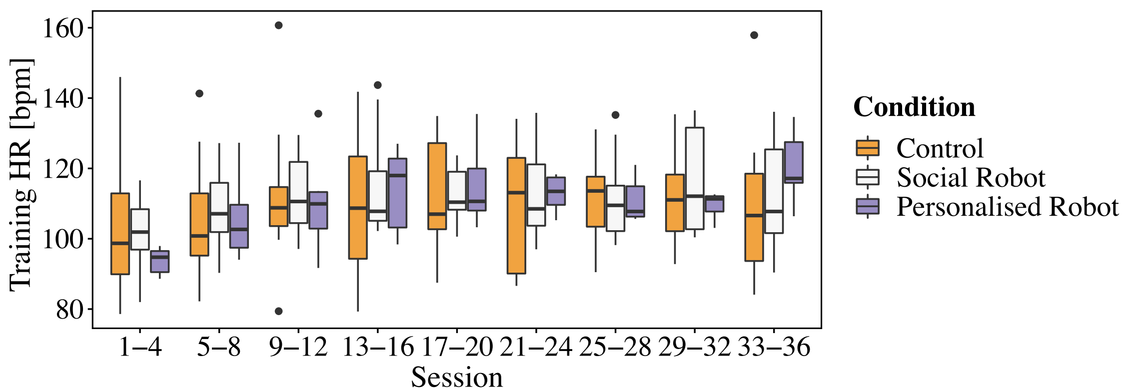


Figure 9.24: The training heart rate (HR) of the patients increased due to the increase in session intensity throughout the CR programme in all conditions. No significant differences were found between the conditions.

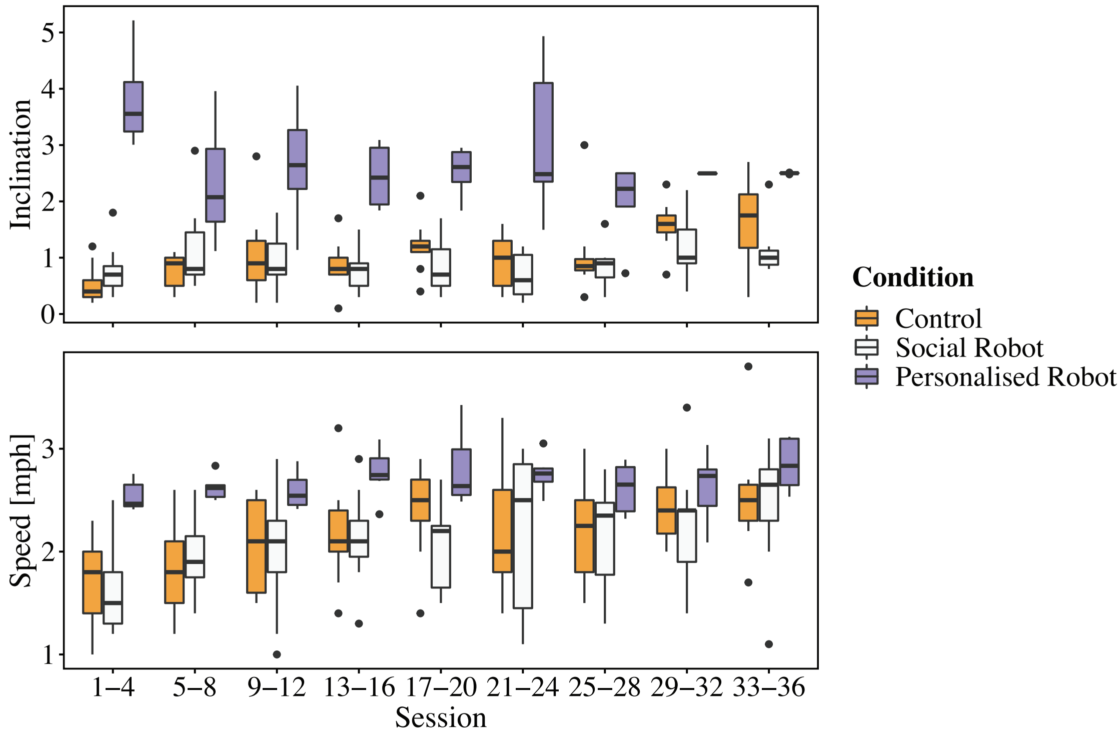


Figure 9.25: Treadmill speed (mph) and inclination (measured in angles) increased throughout the cardiac rehabilitation programme within all conditions. There are significant differences ($p < 10^{-8}$ for speed, $p < 2 \times 10^{-16}$ for inclination) between the *personalised robot* condition and the other conditions, suggesting that the clinicians applied high intensity training in the *personalised robot* condition.

determine the patient's health progress. Figure 9.24 shows the progress of the training heart rate throughout the CR programme for all conditions. As expected, the average heart rate increased throughout the programme ($p = 0.02$) in all the conditions (no significant differences), because the session (exercise) intensity, as determined by the treadmill speed and inclination (Figure 9.25), increased throughout the rehabilitation to improve the health of the patient. In fact, the *personalised robot* condition had a significantly higher session intensity than the other conditions ($p < 10^{-8}$ for speed and $p < 2 \times 10^{-16}$ for inclination). This shows that the (*personalised*) robot allowed high intensity training. This type of training did not have a negative effect on the training heart rate because the medical team could intervene when the value reaches a critical level, as shown in Section 9.4.1.2. While this aspect was present in both robot conditions, the high intensity training was only applied in the *personalised robot* condition. We think that the initial results of the *social robot* condition improved the trust of the clinicians, which facilitated applying this strategy in the *personalised robot* condition that started towards the end of the *social robot* condition.

Figure 9.26 shows that in all conditions, the recovery heart rate (R-HR) of the patients have improved over the duration of the CR programme, which is supported with significant

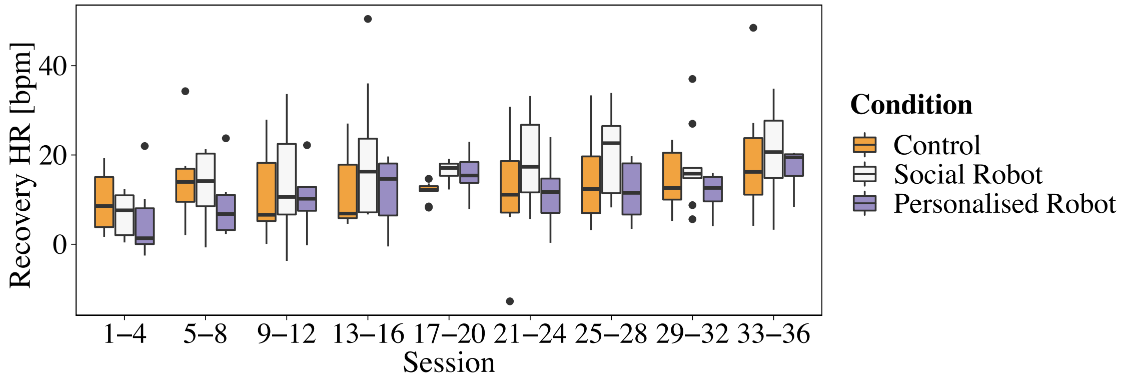


Figure 9.26: The recovery heart rate (R-HR) of the patients improved throughout the cardiac rehabilitation programme for all conditions with $p = 0.005$ for differences between sessions, showing the success of the cardiac rehabilitation programme. No significant differences were found between the conditions.

differences ($p = 0.005$) between sessions. No significant differences were found between the conditions, showing that neither robot negatively affected cardiac rehabilitation programme. On the one hand, this suggests the task performance has not changed with the presence of a robot, against the findings in the literature for different rehabilitation scenarios. On the other hand, this also suggests that the robot did not negatively affect the patients' health, which is a very important finding, because it shows that the robot does not take away from the success of the CR programme (e.g., through distractions).

Despite the higher session intensity throughout the CR programme, the perceived exertion level (Borg scale) stayed within the healthy range (6-12) for all patients, and no significant differences were observed between the conditions or the sessions. In contrast, as previously observed in the *personalised robot* condition, patients may have high heart rate warnings and critical heart rate alerts (*Call medical staff*) during the sessions, as shown in Figure 9.28.

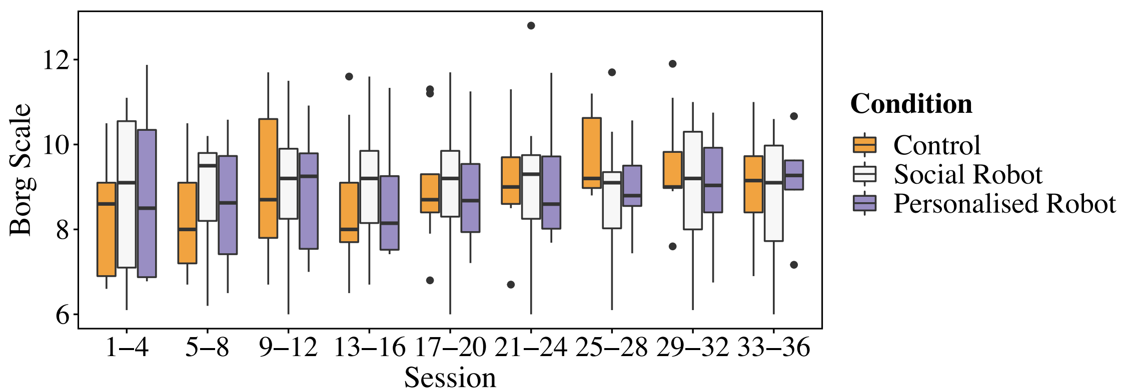


Figure 9.27: The perceived exertion level (Borg scale) of the patients throughout the cardiac rehabilitation programme within all conditions stayed within the healthy range (6-12). No significant differences were found between the conditions or the sessions.

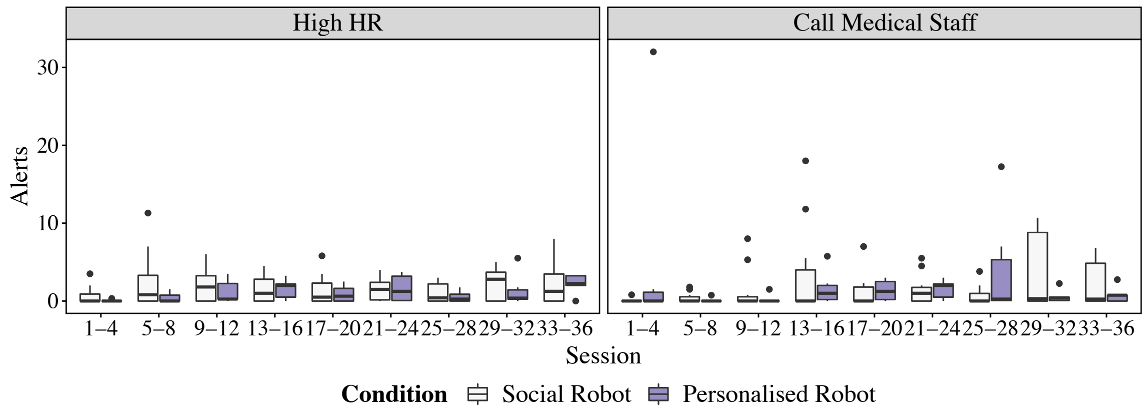


Figure 9.28: Number of high training heart rate (HR) and critical heart rate (*Call medical staff*) alerts of the patients throughout the CR programme. The results show that in contrast to the low perceived exertion levels (Borg scale), warning and critical heart rate values may arise in the sessions throughout the CR programme.

No significant differences were found in the alerts between the *social robot* and *personalised robot* conditions, however, critical heart rate depends on the session ($p = 0$).

9.4.3.3 Cervical Posture Progress

In Section 9.4.2.3, we noticed that the patients complied fully to the robot's requests for cervical posture correction. Due to the lack of video data for the *social robot* condition, we cannot analyse the compliance between both robots. Nonetheless, we can compare the number of requests that the robot made to evaluate if the patients' posture improved over time. While no significant differences were found between conditions or sessions, Figure 9.29 shows that number of requests were lower for patients in the *personalised robot* condition, suggesting that the patients in this condition generally maintained a better posture throughout the CR programme.

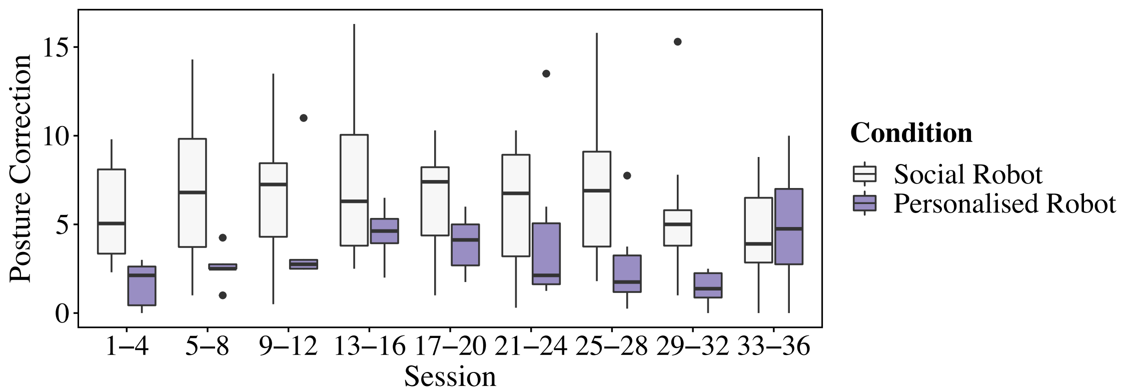


Figure 9.29: The number of cervical posture correction requests by the *social* and the *personalised robot* throughout the cardiac rehabilitation programme. The results show that the corrections were generally less in the *personalised robot* condition.

9.4.3.4 Interaction with the Robot

Figure 9.30 shows that patients generally adjust to the robot and the system over time, which is supported by a significant difference between the sessions ($p = 0.007$). There are no significant differences for the response times ($p = 0.58$) between the robot conditions. However, there is a significant interaction between the condition and sessions ($p = 0.02$), which indicate that the performance within the session depends on the condition.

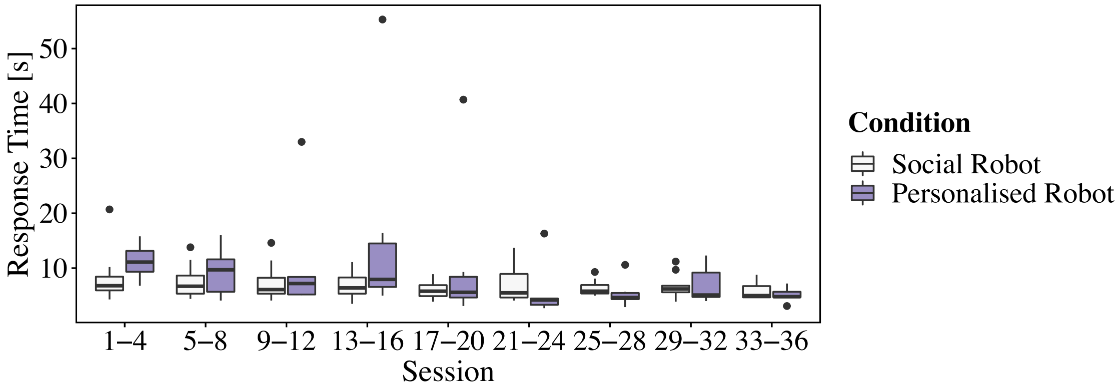


Figure 9.30: Response time of the patients to the Borg scale request of the robot throughout the cardiac rehabilitation programme for *social* and *personalised robot* conditions. Patients in both conditions generally adjust to the robot and the system over time.

9.4.3.5 Perception of the Robot

As previously described in Section 9.3.2, in order to compare the expectations to the experience with the robot, instead of the patients in the *control* condition, the Unified Theory of Acceptance and the Use of Technology (UTAUT) questionnaire was applied to 20 patients at their early outpatient or maintenance phase without any prior experience with the

Table 9.3: Mann-Whitney U-test results for the Unified Theory of Acceptance and the Use of Technology (UTAUT) questionnaire for the *focus group*, the *social robot* and *personalised robot* conditions. The significant differences ($p < 0.05$) are highlighted in bold.

Construct	Focus Group/ Social Robot	Focus Group/ Personalised Robot	Social Robot/ Personalised Robot
Perceived Usefulness (U)	**0.002**	0.35	0.07
Perceived Utility (PU)	**0.0007**	0.49	**0.04**
Safety (S)	0.22	**0.02**	0.28
Ease of Use (EU)	**0.03**	0.13	0.7
Perceived Trust (PT)	**0.0001**	**0.03**	0.17
Perceived Sociability (PS)	0.1	0.26	0.65
Social Presence (SP)	0.17	0.34	0.78

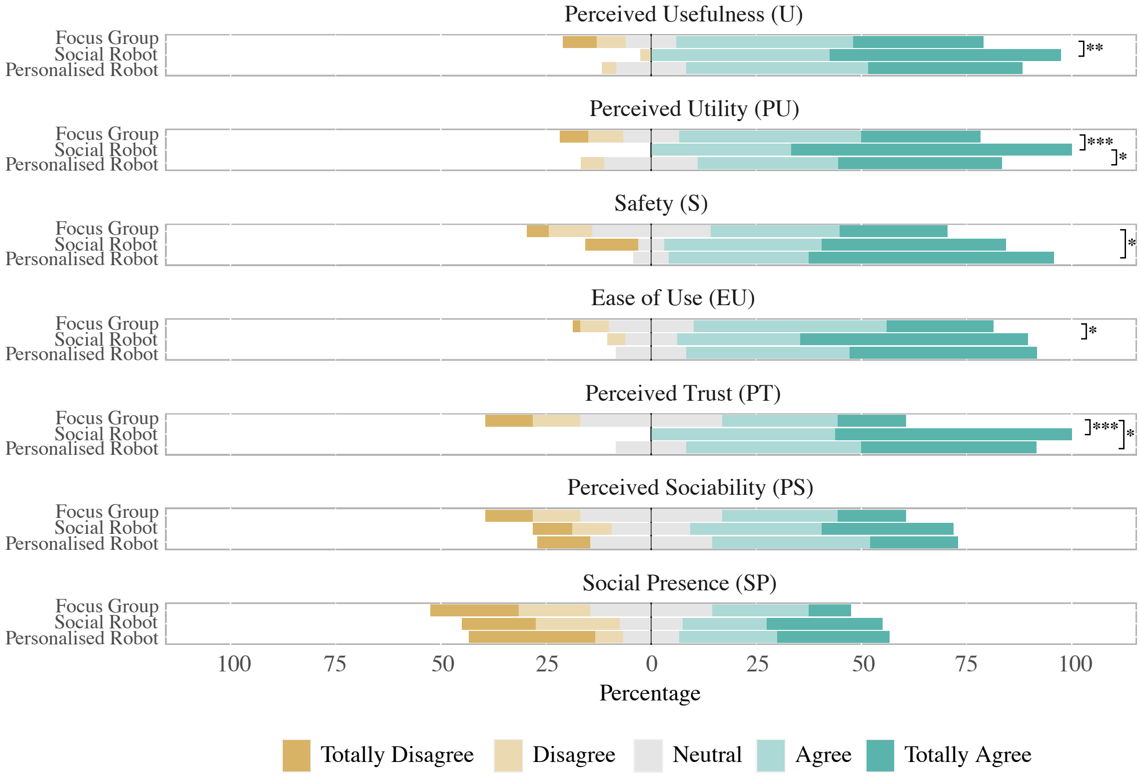


Figure 9.31: Unified Theory of Acceptance and the Use of Technology (UTAUT) questionnaire results for the *focus group*, the *social robot* and *personalised robot* conditions. The patients in the focus group did not have prior experience of the system, and completed the questionnaire after the debriefing and video demonstrations of the *social robot*. The patients in the robot conditions completed the questionnaire after their last session of the outpatient phase of the cardiac rehabilitation programme (i.e., after completing the study). Significant differences are denoted with $p < 0.05$:*, $p < 0.01$:**, $p < 0.001$:***, as presented in Table 9.3.

robot or our system. A debriefing was made about the systems and the parameters measured in the system, followed by a video presentation of the *social robot* condition, before applying the UTAUT questionnaire to the *focus group*. For the patients in the *social robot* and *personalised robot* condition, the UTAUT was applied after the completion of their outpatient phase of the cardiac rehabilitation programme with the robot (i.e., after completing the study). Figure 9.31 and Table 9.3 presents the UTAUT questionnaire (Table F.1 in Appendix F) results and the significant differences between the conditions. Note that the personalisation questions in the UTAUT questionnaire, which was additionally developed for the *personalised robot* condition, are not included in this analysis for fair comparison between the conditions.

There are significant differences between the expectations of the *focus group* and the perceptions of the patients that completed the CR programme with the *social robot*, in terms of perceived usefulness, utility, ease of use, and trust. On the other hand, the patients in the *personalised robot* condition perceived the robot significantly more safe and trusted it more than the *focus group*. The *personalised robot* was perceived more positively than the *social robot* in terms of the perceived sociability, ease of use, safety and social presence constructs, however, no significant differences were found. On the other hand, the usefulness, utility and trust were less positive for the *personalised robot* than the *social robot*. We believe this may be due to the user recognition and recall problems that we experienced within the sessions, which may have caused negative experience, as previously mentioned in Section 9.4.2.2. Nonetheless, both *social robot* and *personalised robot* conditions improved the expectations about the robot and the system.

The additional feedback (through open questions) of the patients in the *social robot* condition was similar to that of the *personalised robot* condition (Section 9.4.2.2). The patients reported that the robot increased their confidence in the CR programme (*"I was very insecure at the beginning of the rehabilitation. Thanks to the robot I got confidence."*), as well as improved their compliance and adherence (e.g., *"I want to come to my rehabilitation. I have the advantage that the robot watches over my health status every second and I feel more secure."*), and their therapy progress (e.g., *"The robot was beneficial to the development of the CR programme."*). In contrast, the patients in the *focus group* had lower confidence in using the robot (e.g., *"I would trust more in human physiatrist."*), as can be observed from the UTAUT results. However, some of the patients acknowledged the potential benefits of continuous monitoring, as observed within both the *social* and *personalised robot* conditions.

Furthermore, as described in Section 9.3.2, the UTAUT questionnaire was also applied to a focus group of 15 clinicians (e.g., nurses, occupational therapists, physiatrists) without prior experience with our system or the robot (Casas et al., 2019). Following the questionnaire, the clinicians were debriefed, similar to the *focus group*, about the capabilities of systems and that of our robot, and a demonstration of the *social robot* condition was made. Afterwards, a discussion group was formed that focuses on the challenges, modifications, and improvements that can be developed in the interface. The results showed that prior to the system's demonstration, the medical team perceived the robot as a threat (i.e., potential replacement of clinicians), however, the explanation of the technology and emphasis on the robot's use as a tool to improve their efficiency during the therapy, in addition to improving patients' engagement, motivation and adherence, the perception of the clinicians positively changed, similar to the results in (Winkle et al., 2018). The clinicians also noted that continuous monitoring and feedback mechanism of the robot can allow high intensity training in the cardiac rehabilitation that would beneficial for the programme, as indeed conducted in the *personalised robot* condition (Section 9.4.3.2). However, similar to the patients in the *personalised robot* condition, they noted that social capabilities (i.e., sociability and social presence) should be improved. The clinicians also suggested adding a feature for providing statistics of the patients' performance as the system was capable to perform online monitoring during the complete outpatient phase. Since this feature is already implemented in the *personalised robot* condition to provide feedback to the patients based on their therapy progress, a graphical user interface (GUI) can be developed in the future to report this information to the clinicians.

9.5 Discussion

The main challenges of CR are close monitoring of the patient's progress and assuring adherence to the long-term program to ensure that the patient recovers fully and retains healthy habits. Our study aimed to improve user motivation and adherence, by recognising users, providing personalised continuous monitoring and feedback, recalling the patients' previous session progress and tracking attendance.

While the CR programme could be successfully completed without the presence of a personalised robot, continuous monitoring and immediate feedback provided additional benefits to the patient and the medical team, as highlighted by the alerts during the sessions, the corrections of the posture, the "critical" session and the reported perceptions of the robot. The results also showed that each patient has different physiological parameters, showing the importance of individualised and personalised care. In addition, continuous monitoring enabled the medical staff to perform high-intensity training without a negative effect on the training heart rate. In fact, the involvement and trust of the medical staff was the key to the success of the CR programme and the reliableness of our proposed solution.

The patients in all conditions reported having low exertion levels throughout the CR programme regardless of the session intensity, however, these results did not correlate with the alerts received during the sessions in the robot conditions, which supported the significance of continuous monitoring. The low reported results may arise from the *self-presentation* (Bond, 1982) effect, which refers to conforming to normative behaviours to gain the approval of another individual (e.g., doctors) (Ganster et al., 2010; Irfan et al., 2018a). Hence, it is important to combine self-reports with sensory data to detect any problems, which the patient may not realise or admit, and which allows correct intervention during the therapy.

Through the personalisation of the robot, we aimed to increase the perceived sociability and social presence of the robot, and increase the user motivation, engagement and adherence. While we cannot derive conclusions on the adherence of the patients due to the outbreak of the COVID-19, the patients in both *social* and *personalised robot* conditions reported that they felt motivated to come to the CR sessions because of the robot. The patients also acknowledged and remarked the usefulness of the robot over the conventional CR programme and that of its personalisation features, and they remarked that they would recommend the CR programme with the robot. In addition, the video analysis in the *personalised robot* condition showed that posture compliance, gaze and social interaction with the robot was maintained throughout the CR programme, which is a valuable result showing that the patient did not lose interest in the robot throughout the long-term therapy and acknowledged the robot as a social agent, by verbally and non-verbally interacting with it.

The long-term perception of the robot and the therapy was generally positively perceived, and the patients kept their bond with the robot over the duration of the CR programme. In addition, both the *social robot* and *personalised robot* improved the expectations about the robot and the system. The *personalised robot* was perceived slightly more positively (but not significantly) than the *social robot* in terms of the perceived sociability, ease of use, safety and social presence. In contrast, the usefulness, utility and trust to the robot were less positive, which we believe was due to the user recognition problems that arose from the face recognition failures, that the patients also remarked. Nevertheless, our multi-modal user recognition with online learning (MMIBN:OL) performed better throughout the *personalised robot* condition than face recognition and the non-adaptive model, showing that the proposed user recognition is suitable for real-world interactions, and improves the recognition even when the identifiers are malfunctioning.

While our proposed robot needs improvements regarding user recognition and repetitiveness of the feedback, the results of the study showed that a personalised socially assistive robot will be beneficial in cardiac rehabilitation to help the patient and the medical team for monitoring the therapy progress within and throughout the sessions, maintaining the motivation and adherence, and achieving compliance for the corrective measures.

9.6 Summary

This chapter described the design and deployment of a personalised socially assistive robot for cardiac rehabilitation programme in a real-world long-term study (2.5 years) conducted at a hospital in Colombia. Cardiac rehabilitation (CR) is a long-term programme lasting 18 weeks (or 36 sessions), which is offered to the patients that suffered a cardiovascular event, in order to accelerate recovery and reduce the risk of suffering recurrent events. We conducted three conditions within the study: (1) *control* condition that closely resembles the conventional CR programme, with the additional sensory interface to obtain physiological (i.e., heart rate and blood pressure) and spatiotemporal parameters (i.e., gait, cadence, step length) and exertion level, (2) *social robot* for providing non-personalised continuous monitoring of these parameters and the cervical posture to provide immediate feedback based on the sensory interface, (3) *personalised robot* that recognises patients, personalises the feedback, recalls the patients' previous session progress and tracks their adherence, with the aim to improve user motivation and adherence to the CR programme. While we could not analyse the adherence to the programme due to the outbreak of COVID-19 that prevented some of the patients to complete the CR programme, the patients in both *social* and *personalised robot* conditions reported an increase in motivation to come to the sessions because of the robot. The majority of the patients in all conditions that actively participated and successfully completed the CR programme, as established by the medical team, had an improvement in their health, as measured by the recovery heart rate. Both the *social robot* and *personalised robot* improved the expectations about the robot and the system. The *personalised robot* was perceived more positively than the *social robot* in terms the perceived sociability, ease of use, safety and social presence constructs, but no significant differences were found. In contrast, we noticed a drop in perceived usefulness, utility and trust, which may have originated from the user recognition errors due to the face recognition failures. Nevertheless, our multi-modal user recognition with online learning (MMIBN:OL) performed better than the face recognition and the non-adaptive model, supporting its suitability for real-world interactions even in the presence of malfunctioning identifiers. The *personalised robot* was also perceived very positively by the patients throughout the CR programme, and the patients reported that they would recommend the CR programme with the robot. The patients maintained gaze and social interaction with the robot, often in response to the personalised behaviours of the robot. In addition, they fully complied to the robot's posture correction requests throughout the CR programme. Continuous monitoring was found to be prominent to cardiac rehabilitation, as it allowed high-intensity training and facilitated immediate intervention by the medical team in critical situations, which may not have been detected otherwise, due to the low reported exertion levels. Overall, this study demonstrated the potential and benefits of long-term personalised interaction in socially assistive robotics.

Chapter 10

Discussion and Conclusions

The main thesis that this work sought to put forward is as follows:

> User experience in long-term human-robot interactions can be improved by personalising the interaction through recognising users and recalling previously learned information.

Correspondingly, we developed a novel user recognition method in Chapter 3 and validated its reliability and robustness in Chapter 4 and Chapter 5. Following that, we designed applications and explored methods for personalisation of the interaction, within customer-oriented service robotics to improve user experience relying on dialogue in Chapter 6 to Chapter 8, and within socially assistive robotics to improve user motivation and adherence to the cardiac rehabilitation (CR) programme in Chapter 9.

This chapter summarises our findings, explains how our research questions were addressed, restates the contributions within our work, discusses the limitations of our work and suggests future directions for research that could build on the findings from this thesis. Following this, the thesis is concluded with the primary outcomes of this work.

10.1 Summary

Initially, Chapter 1 provided an introduction to the motivation, the underlying research questions and the contributions of this thesis. Following that, Chapter 2 presented a background for the three main topics of this thesis: personalisation in long-term Human-Robot Interaction (HRI), user recognition and conversational agents. Our review showed that long-term HRI necessitates fully autonomous robots, especially in real-world interactions, however, user expectations and interest may still wane over time. Several long-term studies showed that personalising the interaction by recalling user's personal attributes, preferences and behaviour patterns, along with previous shared history with users, can improve user interest and engagement and facilitate building rapport with users, particularly in customer-oriented service robotics and socially assistive robotics for healthcare applications. The first step towards personalisation is recognising the user, and in long-term HRI, this requires learning and identifying users autonomously and incrementally, possibly starting from a state without any known users, and adapting to the changes in the user appearances and behaviours. In addition, relying on a uni-modal identifier, such as face recognition, may result in misidentifications due to the noise in the data, such as a blurry image or bad lighting conditions. Combining a primary biometric (e.g., face recognition) with other types of biometric information, such as soft biometrics (e.g., gender, age), was found to improve user recognition. However, our review showed that there are no existing or commercial user recognition systems with such capabilities. Interactions in service robotics and socially assistive robotics are generally closed-domain and task-oriented, which means the conversations contain a limited range of responses and contain a task to achieve, such as ordering a drink or giving feedback to a patient. Our research in conversational agents revealed that most task-oriented dialogue systems in HRI use rule-based approaches due to their robustness, however, they are restricted to the hand-crafted set of rules, which may not contain all the set of phrases that users may use, and none of the proposed rule-based approaches has been evaluated autonomously in real-world studies. In contrast, data-driven dialogue models offer flexibility and reduce the costs of laboriously hand-crafting rules, because they can learn these rules from data itself. However, they have not been applied to user-specific personalisation in long-term interactions with chatbots or robots.

The lack of an available solution for long-term user recognition led us to build a novel method in Chapter 3 using a Multi-modal Incremental Bayesian Network with non-adaptive weights (MMIBN) or with *online learning* (i.e., learning and updating data sequentially), which is the first user recognition method that can continuously and incrementally learn users, without the need for any preliminary training. Our approach is also the first method that combines a primary biometric (i.e., face recognition) with weighted soft biometrics (i.e., gender, age, height and time of interaction) for improving open world user identification in real-time Human-Robot Interaction. Moreover, the proposed approach can be extended with other biometrics and applied to any commercially available robot due to its computationally lightweight structure. We introduced methods to allow incremental and online learning in Bayesian networks, building upon previous literature. We also introduced a loss function for long-term user recognitions and a quality parameter for improving identification.

We evaluated the reliability and robustness of the proposed user recognition approach in a long-term (4 weeks) real-world HRI study with a Pepper robot (SoftBank Robotics Europe, France) and 14 users in Chapter 4. The results showed that the proposed approaches and the designed recognition architecture are suitable for real-time user recognition in long-term human-robot interactions, as they allow fully autonomous user enrolment and recognition, in addition to improving the identification rate compared to face recognition up to 4.4%. Online learning was able to learn the behaviour patterns of the users correctly and decrease the incorrect identifications for new users, however, it did not improve the known user identification rate. While this study was valuable to evaluate its reliability in real-world interactions and optimise the parameters of the network, we believed that the results might be biased due to the small population size and the characteristics of the population.

Correspondingly, in Chapter 5, we artificially generated a Multi-modal Long-Term User Recognition Dataset with a large number of users (200) that has varying characteristics, based on a commonly used image recognition dataset, which contains images of celebrities in events and movies with age and gender information (IMDB-WIKI in Rothe et al. (2015, 2018)). We simulated two types of time of interactions, with users appearing at random times during the day (i.e., uniform distribution) similar to a coffee shop interaction, and with users coming repeatedly at the same time of the day/week, similar to a rehabilitation appointment, and we artificially generated heights. The remaining biometrics are obtained through the Pepper robot's proprietary algorithms using the images of the chosen users in IMDB-WIKI dataset, thereby providing real signals to our Bayesian Network. We compared our non-adaptive and online learning model with various normalisation methods on this dataset to a state-of-the-art open world recognition approach, Extreme Value Machine (Rudd et al., 2018), and the base face recognition and soft biometrics identifiers on the Pepper robot. The results show that the proposed Multi-modal Incremental Bayesian Network (MMIBN) models with hybrid normalisation decrease the long-term recognition performance loss (L) significantly and improve the identification rate significantly and substantially compared to all the baselines, in exchange for a higher number of incorrect estimations of new users. Both MMIBN models perform significantly equivalent for both random and patterned timing and scale well to larger datasets. However, similar to the previous study, online learning did not improve the user recognition performance, on the other hand, decreased the bias in the system caused by face recognition (FR) by equalising the performance between the users. Both models perform better with the increasing number of recognitions and outperform FR and Extreme Value Machine models after only a small number of recognitions. These findings showed that our MMIBN models are suitable to be applied to robots for real-world long-term human-robot interactions as an initial step towards personalising the interaction.

We identified coffee shop interactions to be suitable for personalisation in long-term interactions under the customer-oriented service robotics domain. Similar to user recognition, personalisation requires incremental learning of users requests and adapting the interaction accordingly, which is known as *continual (or lifelong) learning* in machine learning. While this is a trivial problem for rule-based approaches, it poses a challenge to data-driven approaches due to the *catastrophic forgetting* problem, which refers to forgetting previously learned classes upon learning new classes (Parisi et al., 2019).In addition, it is desirable for a data-driven model to learn users preferences from a few samples of interactions, known as *few-shot learning*, which is also a grand challenge in machine learning (Triantafillou et al., 2017). Due to the lack of available corpora for human-human interaction or HRI and the challenges in collecting thousands of interactions, we artificially generated the text-based Barista Datasets to evaluate the rule-based and data-driven approaches on generic and personalised long-term interactions, as described in Chapter 6.

Based on these datasets, in Chapter 7, we designed fully autonomous generic and personalised Barista Robots with our multi-modal user recognition method, an online automatic speech recognition (ASR), and a rule-based dialogue management system (RBDMS). We conducted the first real-world study that explores fully autonomous personalisation in dialogue for long-term (5-days) human-robot interactions in an international student campus for five days with non-native English speakers. During the experiments, we experienced several challenges due to speech recognition failures, arising from the foreign accent of non-native speakers, latency due to connection problems, quietly speaking users, user’s distance from the robot, low accuracy due to the robot’s microphones, and the delay between voice activity detection and recording. Nonetheless, these failures showed that personalisation can overcome a negative user experience. Our study also showed that a rule-based dialogue manager lacked flexibility in responding to various user responses.

Subsequently, in Chapter 8, we explored the potential of data-driven approaches for personalisation in long-term interactions, based on the Barista Datasets. We used the state-of-the-art dialogue models, namely the variants of Memory Networks (Bordes et al., 2016; Joshi et al., 2017; Zhang et al., 2018), Supervised Embeddings (Bordes et al., 2016; Joshi et al., 2017), and Sequence-to-Sequence (Sutskever et al., 2014), which are strong baselines in other domains of personalisation based on single interactions, such as adapting to general user attributes in task-oriented dialogue (Joshi et al., 2017) or "person"alising an open-domain dialogue by maintaining a given personality (Zhang et al., 2018). Our evaluations showed that while the generative Sequence-to-Sequence (Seq2Seq) model and the retrieval-based End-to-End Memory Networks (MemN2N) performed exceedingly well in generic task-oriented dialogue, none of the state-of-the-art data-driven dialogue models performed sufficiently well to be deployed in personalised long-term interactions in the real world. The prominent underlying reason was found to be the lack of capability to learn new customer names or new order items for most models. Hence, providing the user preferences information at the beginning of a dialogue, similar to a knowledge-base extraction, or the separate user profile memory architectures did not markedly improve the performance. The models had high dialogue state tracking errors in the few-shot learning scenario where generative models performed best, whereas, a larger training set improved accuracy in generic task-oriented dialogue, however, no marked improvement was found on suggesting personalised phrases. Moreover, all models were found to be suitable for real-time interaction, but Memory Network and Split Memory Network take the minimum time for training and responding to user queries.

Lastly, in Chapter 9, we described the design and deployment of a personalised socially assistive robot for cardiac rehabilitation programme (of 18 weeks) in a real-world long-term study that ran for 2.5 years at the Fundación Cardioinfantil-Instituto de Cardiología (FCI-IC) hospital in Colombia. The robot used MMIBN with online learning for user recognition, and provided personalised and immediate feedback based on patients' attendance, session and therapy progress, using a sensory interface for continuously measuring physiological (i.e., heart rate, blood pressure) and spatiotemporal (i.e., speed, cadence, step length), and session intensity (i.e., exertion level, and treadmill inclination) parameters. The personalised robot was compared to using a generic (non-personalised) robot and to conventional cardiac rehabilitation programme without a robot. While half of the patients in the personalised robot condition could not complete their therapies due to the outbreak of COVID-19, our initial findings suggest that the personalised robot was perceived very positively throughout the programme, and the patients complied with the robot and the social interaction was maintained for 18 weeks, suggesting that the user motivation and engagement was maintained in the long-term interaction. Both robot conditions improved the perceptions about the therapy with the robot and improved motivation to attend the rehabilitation sessions. The personalised robot was perceived slightly more positively than the generic robot in terms of sociability, ease of use, safety and social presence. The face recognition failed frequently due to the noise in the data (e.g., blurry images and cropped faces), which resulted in a drop for perceived usefulness, utility and trust, however, MMIBN with online learning was found to perform better than the non-adaptive model and face recognition. The fact that the robot monitored the patients continuously, and alerted the medical staff when the values reached a critical level, enabled trust in the clinicians to provide high intensity training, and allowed immediate intervention in critical situations.

Overall, our findings in this work provide conclusive evidence for our main thesis, that is, personalisation improves user experience in long-term human-robot interactions.

10.2 Contributions

Based on the models we developed for evaluating the research questions and their conclusions, the main scientific contributions of this thesis can be summarised as follows, as stated in Chapter 1:

- **Design and implementation of a multi-modal user identification system with incremental and online learning to enable personalisation in long-term interactions.** This contribution is one of the cornerstones of this work and constitutes a fundamental step towards personalising the long-term human-robot interactions in the real world. It is the first method for sequential and incremental learning in open world user recognition, without any need for pre-training. In addition, this proposed approach is the first in combining soft biometrics with a primary biometric for open world user identification in real-time in human-robot interaction (Chapter 3; Irfan et al. (2018b, under review)).

- **Extension of an online learning method for Bayesian networks** based on Voting Expectation Maximization (EM) (Cohen et al., 2001a,b) and Maximum Likelihood estimation was proposed for modelling the noise in the modalities and adapting the learning rate based on the frequency of user appearances (Chapter 3; Irfan et al. (under review)).
- **Introduction of quality of the estimation and long-term recognition performance loss** enabled improving recognition within long-term interactions (Chapter 3; Irfan et al. (2018b, under review)).
- **Evaluation of MMIBN in a real-world long-term user study**, which showed that the proposed model is applicable for user recognition in real-world human-robot interactions. Moreover, the proposed model outperformed base face recognition based on higher user identification rate (Chapter 4; Irfan et al. (2018b)).
- **Creation of a multi-modal long-term user recognition dataset** with 200 users of varying characteristics based on the IMDB-WIKI dataset (Rothe et al., 2015, 2018) enabled evaluating our user recognition model with a large number of users (Chapter 5; Irfan et al. (under review)).
- **Evaluation of MMIBN with the multi-modal long-term user recognition dataset** showed that the proposed model significantly outperforms base face recognition, soft biometrics and a state-of-the-art approach in open world recognition. In contrast to our initial expectations, the model with online learning was found to decrease the recognition performance in comparison to a non-adaptive model with fixed likelihoods. However, it was shown to decrease the bias in face recognition and equalise the performance between users (Chapter 5; Irfan et al. (under review)).
- **Creation of the text-based simulated Barista Datasets for generic and personalised task-oriented closed-domain dialogue** enabled creating a rule-based dialogue manager as an order-taking barista in a coffee shop, in addition to serving as a baseline to train and evaluate data-driven dialogue models (Chapter 6).
- **Design of a fully autonomous barista robot** with MMIBN, automatic speech recognition and a rule-based dialogue management system for generic and personalised

long-term barista interactions (Chapter 7; Irfan et al. (2020b)).

- **Evaluation of the barista robot in a real-world long-term study with non-native English speakers** showed that personalisation can mitigate interaction failures and the negative user experience. This study is the first study for fully autonomous personalisation in dialogue for long-term HRI conducted in the real-world (Chapter 7; Irfan et al. (2020b)).
- **Exploration of the potential of the state-of-the-art data-driven dialogue models using the Barista Datasets** showed that Seq2Seq model achieves near-perfect accuracy in generic long-term interactions, whereas, no model was found to be suitable to be applied to personalisation in real-world long-term interactions. Nevertheless, End-to-End Memory Networks model also performed well in generic long-term interactions and showed potential for personalised long-term interactions (Chapter 8).
- **Design and evaluation of a personalised socially assistive robot for cardiac rehabilitation to improve user motivation and adherence in the real-world long-term clinical therapy of patients**. The results showed that the personalised robot was perceived positively throughout the CR programme. Moreover, the gaze, social interaction and the compliance to the robot's requests were maintained in the long-term and the personalisation features were appreciated by the patients. The patients in both robot conditions reported that working with a robot improved motivation to attend the rehabilitation sessions. Moreover, continuous monitoring of the patient facilitated immediate intervention by the medical team in critical situations and enabled high-intensity training. In addition, our multi-modal user recognition model with online learning was found to perform better than the non-adaptive model, when the identifiers are malfunctioning (Chapter 9; Lara et al. (2017a,b); Casas et al. (2018a,b,c); Irfan et al. (2020a)).

10.3 Experimental Limitations and Future Work

Long-term human-robot interaction in the real world is a very challenging and complex problem, which requires a high level of autonomy and adaptability. Correspondingly, the designed solutions need to be a combination of various features for verbal and non-verbal interactions. In this thesis, we explored various topics as initial stepping stones towards personalised interactions that can meet and maintain user expectations in the long-term. However, each topic we have undertaken is a wide research area by itself, hence, we had to limit our vision, and thus, our experimental work, based on the restrictions of resources and time. In this section, we will touch upon these limitations, and highlight areas where future work could be performed based on our findings.

10.3.1 Multi-modal User Recognition

Our proposed Multi-modal Incremental Bayesian Network (MMIBN) is the first method for sequential and incremental learning in open world user recognition, without any need for pre-training. In addition, this proposed approach is the first in combining soft biometrics with a primary biometric for open world user identification in real-time in human-robot interaction. However, in our experiments, we used only one primary biometric (i.e., face recognition), hence, in the absence of facial information, the recognition could not be made. In such a case, the image was discarded, and another image is taken, or multiple images were taken to increase the reliability, as used in Chapter 4. However, the frequent presence of face recognition failures resulted in low user recognition performance, which negatively affected the user experience in Chapter 7 and Chapter 9. Hence, other primary biometrics, such as voice for speaker identification, can be used. However, there is no reliable open-set speaker identification (i.e., recognising new users, as well as previous users) method (Togneri & Pullella, 2011). While fingerprint identification is a reliable method, especially in combination with soft biometrics (Jain et al., 2004), it has an intrusive nature which is not desirable in HRI. Similarly, QR codes or access cards can be used, however, these are external devices which the user can easily lose, and they reduce the naturalness of the interaction. Therefore, other soft biometrics that do not rely on intrusive methods can be used such as posture, hair colour, location (e.g., in an office), eye colour, gait, clothing and facial marks (Jain & Park, 2009; Scheirer et al., 2011; Arigbabu et al., 2015; Park & Jain, 2010; Zewail et al., 2004; Ouellet et al., 2014; Al-Qaderi & Rad, 2018). Our proposed models allow extension with other biometric traits, they can be applied to any (robot) platform and we are releasing the code of our models, thus, we suggest combining other modalities within our models to improve incremental recognition in real-world HRI.

Our proposed models assume that there is a single person in the image. This can be ensured by using a face detection algorithm prior to using our model to detect the faces in the image, and then align and extract them to apply identification on each separate user. For instance, NAOqi face recognition, used in this work, has that feature. Many other face detection algorithms also provide positions of the boundary of the face in the image, allowing cropping images prior to recognition. After the recognition of the user, a face tracking method can be used to identify if the user has changed for HRI or video-based recognition with multiple users.

We have not compared MMIBN to a state-of-the-art deep learning approach, because there does not exist a method for open world user recognition due to the catastrophic forgetting problem (McClelland et al., 1995; McCloskey & Cohen, 1989; Parisi et al., 2019), which refers to forgetting previously learned classes when a new class is introduced. In addition, our method allows starting from a state without any known users, which deep learning methods cannot offer. Furthermore, our method is suitable to be applied to low-computational power systems, such as robots. Deep learning structures are not optimised for such systems, and open-set methods with re-training can take a considerable amount of time (e.g., 6-7 seconds for the encode-train-recognise cycle per user within a small dataset[1]), which is not suitable for real-time interactions, and the previous data may not be available for re-training.

Online learning was found to decrease the bias in face recognition and improve the recognition in the presence of identifier failures, however, it did not have a marked improvement over the non-adaptive MMIBN. We suggest three possible solutions for improving online learning: (a) identifiers with lower noise can be used, which can be difficult to achieve in real-world scenarios, (b) similar to the work in (Cohen et al., 2001b; Liu & Liao, 2008), increase the learning rate when there is a large error between the estimated parameter and its mean value, and decrease when convergence is reached, (c) confidence value of the identifiers or the quality of the estimation (Q) can be used to determine whether the likelihoods should be updated at each iteration, to avoid updating when the noise is high. We believe that online learning is beneficial for user recognition, thus, strongly suggest MMIBN:OL to be explored with these suggestions. Online learning can also allow detecting anomalies in the data (e.g., the user seen at 1 am in Figure 4.6 in Chapter 4), for security purposes, such as detecting intruders in a house, which may a future application for robots. However, our online learning method requires supervised learning through a "human-in-the-loop" system. There are approaches that allow unsupervised online

[1] `https://www.pyimagesearch.com/2018/06/18/face-recognition-with-opencv-python-and-deep-learning/`

learning (e.g., clustering), but this may result in worse performance due to inaccuracies in the data. In addition, the privacy concern is a point that would need to be addressed in that case and in user recognition, in general. Currently, the model only allows removing the last user, since online learning in the Bayesian network adapts the likelihoods accordingly, hence, a previously added user cannot be simply removed as this would cause errors in likelihoods. However, if all the previous user data is available, the model can be re-trained without the user in question. Our implementation provides a function for this, and the training is relatively fast even when the dataset is large.

The presented results are dependent on the noise level of the identifiers, characteristics of the population (i.e., the distribution of parameters within the population), the defined loss function and α (i.e., the long-term recognition performance loss function parameter that determines the importance of identifying previous users, Detection and Identification Rate (DIR), to identifying new users, False Alarm Rate (FAR), in long-term interactions). However, the Multi-modal Long-Term User Recognition Dataset encapsulates a diverse set of characteristics (i.e., soft biometrics) for a large number of users, thus, we believe that the optimised parameters (i.e., weights, normalisation method, the quality of the estimation) are suitable for deployment to real-world applications, as we have demonstrated in Chapter 9, with the identifiers that we used (i.e., the proprietary algorithms of the Pepper and NAO robots). However, by using other algorithms for the identifiers or by setting a desired FAR depending on the application, a different set of weights can be achieved (using the Multi-modal Long-Term User Recognition Dataset) with lower/higher FAR and consequently lower/higher DIR. Moreover, we suggest using a variable threshold for the quality of the estimation (θ_Q) based on the number of users in the dataset to ensure that the quality is higher when the number of users is low. In addition, we use a fixed set of weights (of likelihoods) for the modalities, however, using adaptable weights per user may provide a better performance, as the performance of the identifiers may change depending on the user. For instance, for a very tall user, the height would be more powerful in identifying the user, as we have seen in Chapter 4.

10.3.2 Personalised Service Robot

In Chapter 7, we presented the first real-world study that explores fully autonomous personalisation in dialogue for long-term human-robot interactions, which was conducted at an international student campus with non-native English speakers. Our study showed that ASR is not reliable in real-world human-robot interactions, even in a structured task-oriented dialogue. Due to a variety of factors (e.g., unreliable internet connection, noise in the environment, quietly speaking users, short sentences that prevented voice detection failures) 69.8% of the utterances were not processed and only 55.4% of the processed ones were correctly recognised, which caused users to repeat their phrases several times and resulted in failed interactions, and in turn, a negative user experience. Nonetheless, our findings also suggested that personalisation mitigates interaction failures and the negative user experience. However, while the experiments were conducted at the coffee bar of the campus, the location was reserved for the experiment and a schedule was created to avoid delays in the experiment and to evaluate the perceptions of the users without being affected by another user's experience. Because of this aspect, the participants had to sign up prior to the experiment, which may have caused participation by people that are interested in interacting with a robot (Kanda et al., 2010). In addition, we had a low number of users due to the challenges in recruiting subjects in long-term experiments, as described in Section 2.1.1. In addition, the duration of the experiment (5 days) may not have been long enough to decrease the novelty effect. We acknowledge that these restrictions may have caused a decrease in the reliability of our findings and their applicability for real-world interactions. In the future, when the technology is robust enough to provide reliable ASR and robust dialogue management, we suggest the experiment to be repeated in a field trial, that is, at a busy coffee shop to confirm our findings. Note that all our findings within real-world HRI studies apply to adults, mostly because of the application areas, thus, they may not apply to interactions with children.

Natural language interaction is a grand challenge in HRI, not only due to the necessity of a high level of automatic speech recognition and language understanding, but also because it contains multi-modal and mixed-initiative interaction (i.e., the robot or human can start or change the interaction) and necessitates cognitive modelling (e.g., understanding the state of the user) (Goodrich & Schultz, 2007). Moreover, user expectations vary between robots and spoken dialogue systems (e.g., Alexa). For instance, if the user says "I would like to order something", the agent should further explore the item the user desires, which could result in a whole chain of to-and-fro questions on the exact nature of the order, which is a more challenging problem than receiving a command such as "Play Bob Marley" (Bartneck et al., 2019). However, the users expect the same level of capability in dialogue with robots, which result in a decrease in user experience, as seen in Chapter 7. Thus, it is important to clearly state the capabilities of the robot (Forlizzi & DiSalvo, 2006), however, the information forms are not sufficient to overcome these expectations, as revealed by our study.

In order to improve the accuracy of ASR, we suggest constraining grammar of ASR (Kennedy et al., 2017), ensuring a reliable internet connection or using an onboard ASR, using high-quality microphones, and adapting the interaction to the native language of the user. However, even when these factors are taken into consideration, there may still be inaccuracies in ASR due to a noisy environment. While rule-based approaches are the most common dialogue managers in HRI, our results showed that they are not suitable in the presence of inaccuracies (Chapter 7). Moreover, rule-based approaches require users to respond in a particular manner, which causes frustrations and loss of time as in our real-world study (Williams et al., 2018; Bartneck et al., 2019; Irfan et al., 2020b). Similar to (Williams et al., 2018), we have observed that users repeat their phrases several times, adapt their speech and even accept wrong orders, if the robot fails to understand them. However, in a real-world scenario, in case of failures, the customers are likely to walk away and never return to the shop. Thus, in contrast to our motivation, it would instead decrease user visits. In addition, instead of decreasing the time for taking an order, it will increase the time, lower the satisfaction (Giuliani et al., 2013), and increase the number of people waiting in the line. On the other hand, Chapter 8 demonstrated that data-driven approaches are not suitable for personalisation in long-term interactions either. Thus, a probabilistic approach, such as a Partially Observable Markov Decision Process (POMDP), can be used which would allow the system to recover from incorrect states due to speech recognition errors, and allow user adaptation (Young et al., 2013; Mo et al., 2016). In addition, user emotions (e.g., negative reactions in Figure 7.6 in Chapter 7) can be detected and modelled (e.g., with a POMDP as in Yuan (2015)) to evaluate the user satisfaction with the bot response, which can also allow recovering from speech recognition errors. In fact, most of the publicly available and commercial task-oriented chatbot systems (e.g., Alexa, Siri, Google Assistant) are often a combination of hand-crafted components, which allow extracting information through common queries, and statistical methods that provide robustness to noise and ambiguity and allow learning through data (Gao et al., 2019). Thus, we suggest comparing a rule-based dialogue manager with a probabilistic one for personalised long-term HRI in a real-world experiment.

Our study showed that explicit confirmation of the user identity was necessary before suggesting the user preference, because users do not pay enough attention to the robot's responses or do not understand it well, which causes further errors in recognition and recall. As previously discussed in Chapter 6, while it is uncommon to give a full name at a coffee shop, using only the first name could result in confusing the users, and thus the most preferred order of the customers with that of another customer with the same first name, because the verification of the identity is based only on the customer's name. This is also undesirable in terms of data sensitivity and privacy, especially in the presence of others (Hedaoo et al., 2019).

Additionally, an interesting research direction might be designing a bartender in multi-party scenarios, similar to (Giuliani et al., 2013; Foster et al., 2012), and personalising the subsequent interactions based on the customers' orders and behaviours towards the robot.

10.3.3 Data-Driven Approaches in Personalised Long-Term Interaction

Using our text-based Barista and Personalised Barista Datasets, we evaluated the potential of the state-of-the-art data-driven dialogue models for user-specific personalisation in task-oriented dialogue for long-term interactions. For these evaluations, we selected the models that are strong baselines for other domains of personalisation based on single interactions, such as adapting to general user attributes in task-oriented dialogue or "person"alising an open-domain dialogue by maintaining a given personality, namely Supervised Embeddings (Bordes et al., 2016; Joshi et al., 2017), Sequence-to-Sequence (Seq2Seq) (Sutskever et al., 2014), End-to-End Memory Networks (MemN2N) (Weston et al., 2014; Sukhbaatar et al., 2015; Bordes et al., 2016), Split Memory (Joshi et al., 2017), Key-Value Profile Memory Networks (Zhang et al., 2018), and Generative Profile Memory Networks (Zhang et al., 2018) based on the Seq2Seq model. The results showed that Seq2Seq and MemN2N performed very well for generic task-oriented dialogue, however, no model could perform sufficiently well for user-specific personalised task-oriented dialogue in long-term interactions.

While our experiments evaluated the performance of data-driven approaches in long-term interactions from a variety of perspectives (i.e., style of the interaction, dataset size, out-of-vocabulary (OOV) entities, inaccuracies and time complexities), there are by no means extensive. We could only train and evaluate each model once, due to the limited available computational power and the vast amount of time required to train some of the methods. Moreover, we used the hyperparameters from the original work, with slight modifications for improved performance or due to the restrictions on resources, however, we have not conducted additional tuning of these parameters, which may provide different results than those presented in this thesis.

Another limitation of our experiments was using implementations based on different definitions of *out-of-vocabulary (OOV)*: (a) words that are not seen in the training set, but are included in the vocabulary (Bordes et al., 2016; Joshi et al., 2017) for MemN2N, Split Memory and Supervised Embeddings models, (b) words are not seen in the training set or included in the vocabulary (Zhang et al., 2018) and ConvAI challenge[2]. We have done so to remain faithful to the original work and also because the latter definition caused erroneous accuracy calculation in the former methods. Including the words in the *OOV* set within the vocabulary for Seq2Seq, Profile Memory and Key-Value may improve the performance, however, based on the timing limitations and the vast amount of time required to train these methods, we could not compare that aspect. Nevertheless, we believe that the definition of out-of-vocabulary should, in fact, not include these words, because in a real-world scenario, a barista can come across new customer names (e.g., foreign names), which they cannot know beforehand. Moreover, the detailed analysis of the inaccuracies, as presented in Section 8.5.4, showed that data-driven approaches are not capable of dealing with that scenario even if they have this information in their vocabulary, whereas this is a trivial *slot-filling* task for rule-based approaches.

Our Barista Datasets simulate interactions with a barista at a coffee shop and contain a variety of utterances to make or change an order, based on the author's personal experiences interacting with baristas. However, as our real-world study showed, users can use a variety of other (similar) phrases. Hence, a Wizard-of-Oz (WoZ) method can be used to collect further data, as well as add the additional phrases that appeared in our study for improving the dataset. Google very recently released a crowd-sourced dataset, Taskmaster[3] (Byrne et al., 2019), collected with WoZ (i.e., a person responded to the user queries, acting in place for the Google Assistant) or self-dialogues (i.e., crowd-sourcers imagine responding to the assistant and write responses for both sides), which contains ordering a drink from coffee shops, among other tasks. This dataset does not contain

[2] http://convai.io/
[3] `https://github.com/google-research-datasets/Taskmaster`

personalised long-term interactions, does not include accompaniments in the order, nor is it based on direct interactions with baristas (i.e., the user interacts with the Google Assistant, which orders the drink for them). However, it contains additions to the drinks such as milk types, other flavours or whipped cream, however, these options are not widely available in all coffee shops. Moreover, it contains combined orders (e.g., a small breakfast tea and a chocolate cookie) and multiple orders (e.g., a large coffee and a small breakfast tea), which our datasets lack, since our main goal was to evaluate personalisation in long-term interactions, which is already a challenging task without these components, as we saw in Chapter 8. Nonetheless, our datasets can be extended with these components, and new tasks can be correspondingly designed for generic task-oriented dialogue, as a future direction. In addition, while the Barista datasets contain a large variety of menu items obtained from the menu of a coffee shop chain, if researchers would like to evaluate or use a barista robot for a different range of menu items, we would be happy to offer them an adapted barista dataset according to their need. In addition, *transfer learning*, i.e., transferring skills learned in one task to another, can be evaluated based on the incremental aspect of our tasks.

We strongly suggest conducting a real-world experiment with the best (trained) models in our work (i.e., Seq2Seq for generic task-oriented dialogue and MemN2N for personalised task-oriented dialogue) to evaluate whether these approaches are, in fact, flexible for responding to various user queries and can deal with automatic speech recognition errors, in comparison to a rule-based approach. A real-world experiment can also allow using human-evaluation to determine the models' fluency, consistency and task performance, which may provide different results than the automated metrics (Zhang et al., 2018). This would also be the first study to explore data-driven approaches in long-term HRI. Note that in our datasets each user interacts the same number of times with the barista, since we desired to eliminate any bias in learning for the data-driven approaches. However, we acknowledge that in a real-world scenario, data-driven approaches may perform worse for some users that interact less with the barista, because of this learned bias.

Due to the "black box" structure of deep learning methods, we do know the exact reason behind some of the patterns we observed in our analysis. For instance, using the model's own (incorrect) responses in the conversation context performed better than using the correct labels for Profile Memory, which does not correlate with our expectations. We may speculate that the model learned to pay more attention to the user responses, and this hypothesis may be partially tested because the attention probabilities are available. However, because this cannot be automatised, the analysis on a specific example cannot be generalised (i.e., it would be "cherry-picking"). Similarly, while we identified that the vanilla MemN2N can learn new entities in the OOV datasets, we do not know why it does not use them, especially for new customer names. This is referred to as the *explainability* (or *interpretability*) problem that the deep learning approaches generally face, whereas, rule-based and probabilistic approaches, such as Bayesian networks and POMDPs, are explainable. In order to understand why such a phenomenon occurs, we need new automated metrics that can evaluate the underlying reasons behind the inaccuracies, similar to our methodology in this work, such that new models can be developed to overcome these deficiencies. In addition, we need explainable methods that can learn from feedback during deployment, similar to (Hancock et al., 2019).

Our results showed that Memory Network (MemN2N) shows potential for personalised long-term interactions. Moreover, since our experiments demonstrated that MemN2N and Split Memory models could train and respond to users in a short amount of time, they are suitable to learn new examples sequentially through re-training, such that (potentially) they can recall new customer names in the subsequent interactions. Correspondingly, for a future direction, we propose developing "Episodic Memory Networks" with a generic pre-trained memory for responding to new users, and a separate memory structure, similar to the Split Memory, for each user that contains the previous dialogue history with the user, learned progressively through re-training or online learning. We believe that this can improve both continual and few-shot learning because pre-trained information will be available for generic dialogue and new information will be learned sequentially per user. In addition, this model could improve *attention* and preference recalling because the previous user history will be contained in a separate memory. Moreover, a *forgetting* mechanism or other biologically-inspired methods (Kirkpatrick et al., 2016; Wang et al., 2020) can be introduced to remove very old memories for increasing the efficiency of memory retrieval and reducing catastrophic forgetting.

Another future direction, based on our findings, is combining a generative (e.g., Sequence-to-Sequence) model with a retrieval-based (e.g., Memory Network) approach, similar to (Madotto et al., 2018), or combining data-driven models with probabilistic approaches, e.g., Markov Decision Processes as in Microsoft's XiaoIce (Zhou et al., 2018) chatbot, or combining Seq2Seq with transformers (Vaswani et al., 2017) as used in (Roller et al., 2020), to compensate for the deficiencies in the models. In addition, Memory Network models can be improved with methods to improve retrieval, such as word-based hashing (Weston et al., 2014; Dodge et al., 2015), clustering word embeddings (Weston et al., 2014), organising memories in a hierarchical system (e.g., Maximum Inner Search Product (MIPS) in (Chandar et al., 2016)), using match-type entities (Bordes et al., 2016) (especially for OOV entities) to help access relevant memories efficiently.

We would like to emphasise that our Barista Datasets and the modified baselines for our experiments are available online[4] for academic use, and we provided all the information about the dataset statistics, hyperparameters and the computational power. Correspondingly, all our results are reproducible, according to the Machine Learning Reproducibility Checklist[5]. We encourage future researchers to evaluate the performance of their baselines using our datasets to develop reliable data-driven approaches for personalised long-term interactions and move the state-of-the-art forward in continual learning.

10.3.4 Personalisation in Socially Assistive Robotics

Our long-term clinical study for cardiac rehabilitation programme with a personalised socially assistive robot demonstrated the positive impact that a robot can have on the patients' health, motivation and perceptions of the programme, which are invaluable findings for both rehabilitation and robotics. However, some of these findings were not statistically significant, hence, making them difficult to generalise. The reason was, of course, the small group size, which mostly originated due to the outbreak of COVID-19 which prevented patients from completing the programme, in addition to other reasons (e.g., funding, medical condition) for dropout, and in turn, the study could not be concluded. We hope to continue the personalised robot condition (with new patients) in the future after the cardiac rehabilitation programme restarts in Fundación Cardioinfantil-Instituto de Cardiología, however, this depends on our project funding, the clinic and external circumstances. Nonetheless, our initial findings demonstrated that a personalised robot is promising for improving motivation in long-term interactions and is suitable for real-world clinical cardiac rehabilitation programme.

Incorporating a robot in a clinical therapy poses several challenges, such as user recognition failures and technical failures in the other sensors (e.g., incorrect measurements

[4] `https://github.com/birfan/BaristaDatasets`

[5] `https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf`

in heart rate or gait, and malfunctioning tablet interface), which led to slow response times, incorrect reactions, as well as low interactivity, which may have resulted in a drop the user's interest and trust in the system in the long-term interaction, as we have also observed in our study (Section 9.4.3). While these challenges will always be present due to the noise in the environment or the sensors, they should be addressed with the best available methods before deploying robots to clinical therapy without the supervision of the experimenter. Correspondingly, our multi-modal user recognition method should be improved with previously suggested approaches, and robust additional sensors should be added to the system.

Based on the suggestions of the patients and the clinicians, novel and a larger variety of robot responses should be added to decrease the repetitiveness of the robot, especially in long-term interactions. Moreover, a future research direction can be to adapt these responses over time based on the patient's sensory values to keep the interactions engaging and interesting in the long-term (Matarić & Scassellati, 2016). For instance, if the user has a high heart rate or becomes tired, this can be addressed in the feedback. An emotion detection method can be used (Tapus et al., 2007), however, this may not be robust because the patient is mostly focused and tired during the exercise, hence, the detected emotions may not reliably model the user.

In addition, as suggested by the clinicians, a feature to display the patients' performance can be added to help the clinicians in closely monitoring the progress of the patient. This is partially implemented in the personalised robot condition, but a graphical user interface (GUI) needs to be developed for presenting the results to the clinicians.

Our study was brought to a halt in May 2020 because of the COVID-19 pandemic, which stopped the CR programme in Fundación Cardioinfantil-Instituto de Cardiología clinic because the patients with cardiovascular diseases carry a high risk in mortality. While we fully agree with the risks and the decisions of the medical staff, our proposed solution can allow reducing the necessary physical contact with the medical staff, hence, it is an ideal solution for continuing the cardiac rehabilitation programme, which is vital for patients with cardiovascular diseases. However, as we previously stated, its components need to be improved for reliability and robustness for deploying it without the supervision of the experimenters. Nonetheless, we would like to emphasise that the success of the robot and the study relies heavily on the adoption of the method by the clinicians and their immediate interventions in critical situations. Prior to their experience with the robot, the clinicians were reluctant to use a robot, mostly out of fear that the robot would replace their jobs, as presented in Chapter 9. However, the demonstrations of the robot positively changed their perceptions and allowed them to understand that the robot is there as a tool to help them improve the CR programme, but relies fundamentally on their collaboration. Moreover, the continuous monitoring aspect of the robot allowed high intensity training, as observed in the personalised robot condition, without any negative consequences. However, the robot might have introduced additional duties for the clinicians, hence, it would be beneficial to interview the medical staff (e.g., occupational therapists, physiatrists, nurses) involved in the study to evaluate their long-term perspectives of the robot.

In all our real-world experiments, we used the Pepper and NAO robots from SoftBank Robotics Europe. The Pepper robot was chosen for the user recognition study (Chapter 4) because of its noticeable height and integrated tablet. We used the Adapted Pepper for the barista robot study (Chapter 7) due to its improved microphones and height appropriate for a barista. In contrast, we used the NAO robot for cardiac rehabilitation (Chapter 9), because the literature suggests using a robot with a child-like appearance for assistive tasks (Tapus et al., 2007). However, one of the patients in the personalised robot condition considered its appearance and voice as childish and not appropriate for rehabilitation therapy, while other patients in the robot conditions did not express this. It could be interesting to repeat the study with a more adult-like robot (e.g., Pepper) to evaluate whether that improves the perceived social presence. Thus, it is unclear as to whether the findings of this thesis would translate to robots of a different size, appearance, or morphology. Nonetheless, due to the current wide availability and the common use of these robots, the findings in this thesis are still directly relevant to many other researchers regardless of whether the results generalise to other platforms. Moreover, in general, our findings agree with that of the previous research conducted with other robots and supports our main thesis, that is, personalisation improves user experience in long-term human-robot interactions.

10.4 Conclusion

While a robot may seem engaging and interesting at first sight or in short-term interactions, the generic and repetitive behaviours of the robot may cause the user engagement and interest to wane over time. The thesis presented here is that user experience in long-term human-robot interactions can be improved by personalising the interaction through recognising users and recalling previously learned information. Correspondingly, we developed the first multi-modal incremental user recognition algorithm suitable for real-world interactions and designed personalised robots for long-term real-world applications, namely, a barista robot for a coffee shop and a socially assistive robot for a clinical cardiac rehabilitation programme. On the one hand, our experiments showed that fully autonomous personalisation is not sufficiently robust for real-world interactions with the current state-of-the-art architectures, such as user or speech recognition, or rule-based or data-driven dialogue models. On the other hand, personalisation mitigated the interaction failures and the negative user experience within diverse fields, and helped maintain motivation and engagement in long-term interactions, supporting our thesis. These are promising findings to further design and deploy personalised robots in health-care and service domains, building on top of our proposed solutions, as well as potentially for education and domestic applications with long-term interactions. However, reliable and adaptable fully autonomous solutions that can handle the variability in user behaviours, learn incrementally and sequentially, and recover from failures need to be developed in parallel to real-world studies, in order to attain personalised robots that can meet user expectations and provide long-lasting and pleasant interactions.

Bibliography

Aamot, I. L., Forbord, S. H., Karlsen, T., & Støylen, A. (2014). Does rating of perceived exertion result in target exercise intensity during interval training in cardiac rehabilitation? A study of the Borg scale versus a heart rate monitor. *Journal of Science and Medicine in Sport*, *17*(5), 541–545.

Aarts, A. A. e. a. O. S. C. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Abreu, M. C. D. C., & Fairhurst, M. (2011). Enhancing identity prediction using a novel approach to combining hard- and soft-biometric information. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *41*(5), 599–607.

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., & Le, Q. V. (2020). Towards a human-like open-domain chatbot. `https://arxiv.org/abs/2001.09977`.

Ahmad, M. I., Mubin, O., & Orlando, J. (2017). A systematic review of adaptivity in human-robot interaction. *Multimodal Technologies and Interaction*, *3*(14).

Ahmad, M. I., Mubin, O., Shahid, S., & Orlando, J. (2019). Robot's adaptive emotional feedback sustains children's social engagement and promotes their vocabulary learning: a long-term child-robot interaction study. *Adaptive Behavior*, *27*(4), 24–266.

Al-Qaderi, M. K., & Rad, A. B. (2018). A multi-modal person recognition system for social robots. *Applied Sciences*, *8*(3).

Allen, J. F., Schubert, L. K., Ferguson, G., Heeman, P. A., Hwang, C. H., Kato, T., Light, M., Martin, N. G., Miller, B. W., Poesio, M., & Traum, D. R. (1994). The trains project: a case study in building a conversational planning agent. *J. Exp. Theor. Artif. Intell.*, *7*, 7–48.

Altenhoener, T., Leppin, A., Grande, G., & Romppel, M. (2005). Social inequality in patients' physical and psychological state and participation in rehabilitation after myocardial infarction in germany. *International Journal of Rehabilitation Research*, *28*(3), 251–257.

Arigbabu, O. A., Ahmad, S. M. S., Adnan, W. A. W., & Yussof, S. (2015). Integration of multiple soft biometrics for human identification. *Pattern Recogn. Lett.*, *68*(P2), 278–287.

Aryananda, L. (2001). Online and unsupervised face recognition for humanoid robot: toward relationship with people. In *IEEE-RAS International Conference on Humanoid Robots*. Tokyo, Japan: IEEE.

Aryananda, L. (2009). Learning to recognize familiar faces in the real world. In *2009 IEEE International Conference on Robotics and Automation (ICRA)*, (pp. 1991–1996). Kobe, Japan: IEEE.

Atkinson, R., & Shiffrin, R. (1968). Human memory: A proposed system and its control processes. In K. W. Spence, & J. T. Spence (Eds.) *Psychology of Learning and Motivation*, vol. 2, (pp. 89–195). Academic Press.

Aust, H., Oerder, M., Seide, F., & Steinbiss, V. (1995). The philips automatic train timetable information system. *Speech Communication*, *17*(3), 249–262. Interactive Voice Technology for Telecommunication Applications.

Austin, J. L. (1962). *How to Do Things with Words*. Harvard University Press.

Bach, K., & Harnish, R. (1979). *Linguistic communication and speech acts*. MIT Press.

Baddeley, A., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.) *The psychology of learning and motivation: Advances in research and theory*, vol. 8, (pp. 47–89). New York, NY, USA: Academic Press.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. `https://arxiv.org/abs/1409.0473`.

Bai, B., Weston, J., Collobert, R., & Grangier, D. (2009). Supervised semantic indexing. In M. Boughanem, C. Berrut, J. Mothe, & C. Soule-Dupuy (Eds.) *Advances in Information Retrieval. ECIR 2009*, vol. 5478 of *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer.

Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008). The effect of presence on human-robot interaction. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, (pp. 701–706).

Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., & Sabanovic, S. (2019). *Human-Robot Interaction - An Introduction*. Cambridge: Cambridge University Press.

Bauer, E., Koller, D., & Singer, Y. (1997). Update rules for parameter estimation in Bayesian networks. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, UAI'97, (pp. 3–13). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Baxter, P., & Belpaeme, T. (2014). Pervasive memory: the future of long-term social hri lies in the past.

Beleites, C., & Salzer, R. (2008). Assessing and improving the stability of chemometric models in small sample size situations. *Analytical and Bioanalytical Chemistry*, *390*(5), 1261–1271.

Belpaeme, T., Baxter, P., Read, R., Wood, R., Cuayáhuitl, H., Kiefer, B., Racioppa, S., Kruijff-Korbayová, I., Athanasopoulos, G., Enescu, V., Looije, R. M., Neerincx, M. A., Demiris, Y., Ros-Espinoza, R., Beck, A., Cañamero, L., Hiolle, A., Lewis, M., Baroni, I., Nalin, M., Cosi, P., Paci, G., Tesser, F., Sommavilla, G., & Humbert, R. (2013). Multimodal child-robot interaction: Building social bonds. *J. Hum.-Robot Interact.*, *1*(2), 33–53.

Bendale, A., & Boult, T. (2015). Towards open world recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1893–1902). Boston, MA, USA: IEEE.

Bendale, A., & Boult, T. E. (2016). Towards open set deep networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1563–1572). Las Vegas, NV, USA: IEEE.

Bennett, C., & Rudnicky, A. I. (2002). The carnegie mellon communicator corpus. URL `http://www.speech.cs.cmu.edu/Communicator/Corpus/`

Beswick, A. D., Rees, K., West, R. R., Taylor, F. C., Burke, M., Griebsch, I., Taylor, R. S., Victory, J., Brown, J., & Ebrahim, S. (2005). Improving uptake and adherence in cardiac rehabilitation: literature review. *Journal of Advanced Nursing*, *49*(5), 538–555.

Bethel, C. L., & Murphy, R. R. (2008). Survey of non-facial/non-verbal affective expressions for appearance-constrained robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(1), 83–92.

Bethell, H., Lewin, R., & Dalal, H. (2009). Cardiac rehabilitation in the united kingdom. *Heart*, *95*(4), 271–275.

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.*, *12*(2), 293–327.

Biel, A. L., & Friedrich, E. V. C. (2018). Why you should report bayes factors in your transcranial brain stimulation studies. *Frontiers in Psychology*, *9*, 1125.

Bigün, E. S., Bigün, J., Duc, B., & Fischer, S. (1997). Expert conciliation for multi modal person authentication systems by Bayesian statistics. In J. Bigün, G. Chollet, & G. Borgefors (Eds.) *Audio- and Video-based Biometric Person Authentication. AVBPA 1997. Lecture Notes in Computer Science*, vol. 1206. Berlin, Heidelberg: Springer.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc.

Blom, J. (2000). Personalization: A taxonomy. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '00, (pp. 313–314). New York, NY, USA: Association for Computing Machinery.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., & Winograd, T. (1977). Gus, a frame-driven dialog system. *Artificial Intelligence*, *8*(2), 155–173.

Bond, C. F. (1982). Social Facilitation: A self-presentational view. *Journal of Personality and Social Psychology*, *42*, 1042–1050.

Bonnard, Q., Lemaignan, S., Zufferey, G., Mazzei, A., Cuendet, S., Li, N., Özgür, A., & Dillenbourg, P. (2013). Chilitags 2: Robust fiducial markers for augmented reality and robotics.
URL `http://chili.epfl.ch/software`

Bordes, A., Boureau, Y.-L., & Weston, J. (2016). Learning end-to-end goal-oriented dialog. `https://arxiv.org/abs/1605.07683`.

Borg, G. (1998). *Borg's perceived exertion and pain scales*. Human Kinetics.

Bratman, M. E. (1999). *Intention, Plans, and Practical Reason*. CSLI Publications.

Breazeal, C. (2002). *Designing Sociable Robots*. Cambridge, MA, USA: MIT Press.

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Duckworth, D., Yavuz, S., Goodrich, B., Dubey, A., Kim, K.-Y., & Cedilnik, A. (2019). Taskmaster-1: Toward a realistic and diverse dialog dataset. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP)*.

Campos, J., Kennedy, J., & Lehman, J. F. (2018). Challenges in exploiting conversational memory in human-agent interaction. In *International Conference on Autonomous Agents and Multiagent Systems*.

Cao, H., Van de Perre, G., Kennedy, J., Senft, E., Gómez Esteban, P., De Beir, A., Simut, R., Belpaeme, T., Lefeber, D., & Vanderborght, B. (2019). A personalized and platform-independent behavior control system for social robots in therapy: Development and applications. *IEEE Transactions on Cognitive and Developmental Systems*, *11*(3), 334–346.

Carlson, J., Johnson, J., Franklin, B., & VanderLaan, R. (2000). Program participation, exercise adherence, cardiovascular outcomes, and program cost of traditional versus modified cardiac rehabilitation. *Am J Cardiol*, *86*(1), 17–23.

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The robotic social attributes scale (rosas): Development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 254–262). ACM.

Casas, J., Céspedes, N., Cifuentes, C., Gutierrez, L. F., Rincón-Roncancio, M., & Múnera, M. (2019). Expectation vs. reality: Attitudes towards a socially assistive robot in cardiac rehabilitation. *Appl. Sci.*, *9*, 4651.

Casas, J., Gomez, N. C., Senft, E., Irfan, B., Gutiérrez, L. F., Rincón, M., Múnera, M., Belpaeme, T., & Cifuentes, C. A. (2018a). Architecture for a social assistive robot in cardiac rehabilitation. In *2018 IEEE 2nd Colombian Conference on Robotics and Automation (CCRA)*, (pp. 1–6).

Casas, J., Irfan, B., Senft, E., Gutiérrez, L., Rincon-Roncancio, M., Munera, M., Belpaeme, T., & Cifuentes, C. A. (2018b). Social assistive robot for cardiac rehabilitation: A pilot study with patients with angioplasty. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 79–80). ACM.

Casas, J., Irfan, B., Senft, E., Gutiérrez, L., Rincon-Roncancio, M., Munera, M., Belpaeme, T., & Cifuentes, C. A. (2018c). Towards a SAR system for personalized cardiac rehabilitation: A patient with PCI. In *2018 ACM/IEEE International Conference on Human-Robot Interaction Personal Robots for Exercising and Coaching workshop*. ACM.

Casas, J., Senft, E., Gutiérrez, L., Rincón-Roncancio, M., Múnera, M., Belpaeme, T., & Cifuentes, C. A. (2020). Social assistive robots: Assessing the impact of a training assistant robot in cardiac rehabilitation. *International Journal of Social Robotics*.

Castellano, G., Aylett, R., Dautenhahn, K., Paiva, A., McOwan, P. W., & Ho, S. (2008). Long-term affect sensitive and socially interactive companions. In *Fourth International Workshop on Human-Computer Conversation*. Bellagio, Italy: L. Erlbaum Associates Inc.

Chan, J., & Nejat, G. (2011). A learning-based control architecture for an assistive robot providing social engagement during cognitively stimulating activities. In *2011 IEEE International Conference on Robotics and Automation*, (pp. 3928–3933).

Chandar, S., Ahn, S., Larochelle, H., Vincent, P., Tesauro, G., & Bengio, Y. (2016). Hierarchical memory networks. `https://arxiv.org/abs/1605.07427`.

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, *19*(2), 25–35.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, *abs/1409.1259*. URL `http://arxiv.org/abs/1409.1259`

Chung, J., Gülçehre, Ç., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, *abs/1412.3555*. URL `http://arxiv.org/abs/1412.3555`

Churamani, N., Anton, P., Brügger, M., Flieundefinedwasser, E., Hummel, T., Mayer, J., Mustafa, W., Ng, H. G., Nguyen, T. L. C., Nguyen, Q., & et al. (2017). The impact of personalisation on human-robot interaction in learning scenarios. In *Proceedings of the 5th International Conference on Human Agent Interaction*, (pp. 171–180). New York, NY, USA: Association for Computing Machinery.

Clabaugh, C., Mahajan, K., Jain, S., Pakkar, R., Becerra, D., Shi, Z., Deng, E., Lee, R., Ragusa, G., & Matarić, M. (2019). Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders. *Frontiers in Robotics and AI*, *6*, 110.

Cohen, I., Bronstein, A., & Cozman, F. G. (2001a). Adaptive online learning of bayesian network parameters. Tech. Rep. HPL-2001-156, HP Laboratories.

Cohen, I., Bronstein, A., & Cozman, F. G. (2001b). Online learning of Bayesian network parameters. Tech. Rep. HPL-2001-55 (R.1), HP Laboratories.

Cohen, P. R., & Levesque, H. J. (1990). Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.) *Intentions in Communication*. MIT Press.

Cohen, P. R., & Perrault, C. R. (1986). Elements of a plan-based theory of speech acts. In *Communication in Multiagent Systems*.

Conn, K., Changchun Liu, Sarkar, N., Stone, W., & Warren, Z. (2008). Affect-sensitive assistive intervention technologies for children with autism: An individual-specific approach. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, (pp. 442–447).

Cooper, A. F., Weinman, J., Hankins, M., Jackson, G., & Horne, R. (2007). Assessing patients' beliefs about cardiac rehabilitation as a basis for predicting attendance after acute myocardial infarction. *Heart*, *93*(1), 53–58.
URL `https://heart.bmj.com/content/93/1/53`

Core, M. G., & Allen, J. F. (1997). Coding dialogs with the damsl annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, (pp. 28–35).

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*, 163–191.

Cruz, C., Sucar, L. E., & Morales, E. F. (2008). Real-time face recognition for human-robot interaction. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, (pp. 1–6). Amsterdam, Netherlands: IEEE.

Dantcheva, A., Elia, P., & Ross, A. (2016). What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, *11*(3), 441–467.

Dautenhahn, K. (2004). Robots We Like to Live With ?! - A Developmental Perspective on a Personalized , Life-Long Robot Companion. In *2004 IEEE International Workshop on Robot and Human Interactive Communication*, (pp. 17–22).

de Graaf, M. M. A., Allouch, S. B., & Klamer, T. (2015). Sharing a life with harvey: Exploring the acceptance of and relationship-building with a social robot. *Computers in Human Behavior*, *43*, 1–14.

De Rosa, R., Mensink, T., & Caputo, B. (2016). Online open world recognition. `https://arxiv.org/abs/1604.02275`.

de Ruyter, B., Saini, P., Markopoulos, P., & van Breemen, A. (2005). Assessing the effects of building social intelligence in a robotic interface for the home. *Interacting with Computers*, *17*(5), 522 – 541. Social Impact of Emerging Technologies.

Deng, E., Mutlu, B., & Mataric, M. J. (2019). Embodiment in socially interactive robots. *Foundations and Trends in Robotics*, *7*(4), 251–356.

Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I., Lowe, R., Prabhumoye, S., Black, A. W., Rudnicky, A., Williams, J., Pineau, J., Burtsev, M., & Weston, J. (2019). The second conversational intelligence challenge (convai2). `https://arxiv.org/abs/1902.00098`.

Dodge, J., Gane, A., Zhang, X., Bordes, A., Chopra, S., Miller, A. H., Szlam, A., & Weston, J. (2015). Evaluating prerequisite qualities for learning end-to-end dialog systems. `https://arxiv.org/abs/1511.06931`.

Doering, M., Glas, D. F., & Ishiguro, H. (2019a). Modeling interaction structure for robot imitation learning of human social behavior. *IEEE Trans. Hum.-Mach. Syst.*, *49*(3), 219–231.

Doering, M., Kanda, T., & Ishiguro, H. (2019b). Neural-network-based memory for a social robot: Learning a memory model of human behavior from data. *J. Hum.-Robot Interact.*, *8*(4).

Doering, M., Liu, P., Glas, D. F., Kanda, T., Kuliundefined, D., & Ishiguro, H. (2019c). Curiosity did not kill the robot: A curiosity-based learning system for a shopkeeper robot. *J. Hum.-Robot Interact.*, *8*(3).

Dondrup, C., Baillie, L., Broz, F., & Lohan, K. (2018). How can we transition from lab to the real world with our hci and hri setups? In *2018 ACM/IEEE International Conference on Human-Robot Interaction, 4th Workshop on Public Space Human-Robot Interaction: MobileHRI: Mobile human-robot interaction*. Chicago, IL, USA.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3), 177–190. Socially Interactive Robots.

El Asri, L., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., & Suleman, K. (2017). Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, (pp. 207–219). Saarbrücken, Germany: Association for Computational Linguistics.

Eriksson, J., Matarić, M. J., & Winstein, C. J. (2005). Hands-off assistive robotics for post-stroke arm rehabilitation. In *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.*, (pp. 21–24).

Fasola, J., & Matarić, M. (2013a). Socially assistive robot exercise coach: Motivating older adults to engage in physical exercise. In J. Desai, G. Dudek, O. Khatib, & V. Kumar (Eds.) *Experimental Robotics*, vol. 88 of *Springer Tracts in Advanced Robotics*, (pp. 463–479). Heidelberg: Springer.

Fasola, J., & Matarić, M. J. (2013b). A socially assistive robot exercise coach for the elderly. *J. Hum.-Robot Interact.*, *2*(2), 3–32.

Fayin Li, & Wechsler, H. (2005). Open set face recognition using transduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(11), 1686–1697.

Fei, G., Wang, S., & Liu, B. (2016). Learning cumulatively to become more knowledgeable. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (pp. 1565–1574). San Francisco, CA: ACM.

Feil-Seifer, D., & Matarić, M. J. (2005). Defining socially assistive robotics. In *9th International Conference on Rehabilitation Robotics (ICORR) 2005*, (pp. 465–468).

Fernaeus, Y., Håkansson, M., Jacobsson, M., & Ljungblad, S. (2010). How do you play with a robotic toy animal? a long-term study of pleo. In *Proceedings of the 9th International Conference on Interaction Design and Children*, IDC '10, (pp. 39–48). New York, NY, USA: Association for Computing Machinery.

Fernaeus, Y., Jacobsson, M., Ljungblad, S., & Holmqvist, L. E. (2009). Are we living in a robot cargo cult? In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (pp. 279–280).

Filliat, D. (2007). A visual bag of words method for interactive qualitative localization and mapping. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, (pp. 3921–3926). Rome, Italy: IEEE.

Forlizzi, J., & DiSalvo, C. (2006). Service robots in the domestic environment: A study of the roomba vacuum in the home. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, HRI '06, (pp. 258–265). New York, NY, USA: Association for Computing Machinery.

Foster, M. E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., & Petrick, R. P. (2012). Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, (pp. 3–10). New York, NY, USA: Association for Computing Machinery.

François, D., Powell, S., & Dautenhahn, K. (2009). A long-term study of children with autism playing with a robotic pet: Taking inspirations from non-directive play therapy to encourage children's proactivity and initiative-taking. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, *10*(3), 324–373.

Gaisser, F., Rudinac, M., Jonker, P. P., & Tax, D. (2013). Online face recognition and learning for cognitive robots. In *2013 16th International Conference on Advanced Robotics (ICAR)*, (pp. 1–9). Montevideo, Uruguay: IEEE.

Ganhotra, J., & Polymenakos, L. (2018). Knowledge-based end-to-end memory networks. `https://arxiv.org/abs/1804.08204`.

Ganster, T., Eimler, S., Rosenthal-von der Pütten, A. M., Hoffmann, L., & Krämer, N. (2010). Methodological considerations for long-term experience with robots and agents. In *European Meetings on Cybernetics and Systems Research (EMCSR)*.

Gao, J., Galley, M., & Li, L. (2019). Neural approaches to conversational ai. *Foundations and Trends in Information Retrieval*, *13*(2-3), 127–298.

Gašić, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., Tsiakoulis, P., & Young, S. (2013). POMDP-based dialogue manager adaptation to extended domains. In *Proceedings of the SIGDIAL 2013 Conference*, (pp. 214–222). Metz, France: Association for Computational Linguistics.
URL `https://www.aclweb.org/anthology/W13-4035`

Ge, Z., Demyanov, S., Chen, Z., & Garnavi, R. (2017). Generative openmax for multi-class open set classification. `https://arxiv.org/abs/1707.07418`.

Genevay, A., & Laroche, R. (2016). Transfer learning for user adaptation in spoken dialogue systems. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, (pp. 975–983). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Giuliani, M., Petrick, R. P., Foster, M. E., Gaschler, A., Isard, A., Pateraki, M., & Sigalas, M. (2013). Comparing task-based and socially intelligent behaviour in a robot bartender. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, (pp. 263–270). New York, NY, USA: Association for Computing Machinery.

Giuliano, C., Parmenter, B. J., Baker, M., Mitchell, B. L., Williams, A. D., Lyndon, K., Mair, T., Maiorana, A., Smart, N. A., & Levinger, I. (2017). Cardiac rehabilitation for patients with coronary artery disease: A practical guide to enhance patient outcomes through continuity of care. *Clin Med Insights Cardiol.*, *11*.

Glas, D. F., Minato, T., Ishi, C. T., Kawahara, T., & Ishiguro, H. (2016). Erica: The erato intelligent conversational android. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, (pp. 22–29).

Glas, D. F., Wada, K., Shiomi, M., Kanda, T., Ishiguro, H., & Hagita, N. (2017). Personal greetings: Personalizing robot utterances based on novelty of observed behavior. *International Journal of Social Robotics*, *9*(2), 181–198.

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A. C., & Jue Wang (2005). Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, (pp. 1338–1343).

Goddeau, D., Meng, H., Polifroni, J., Seneff, S., & Busayapongchai, S. (1996). A form-based dialogue manager for spoken language applications. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, vol. 2, (pp. 701–704).

Gonzales, C., Torti, L., & Wuillemin, P.-H. (2017). aGrUM: a Graphical Universal Model framework. In *International Conference on Industrial Engineering, Other Applications of Applied Intelligent Systems*, Proceedings of the 30th International Conference on Industrial Engineering, Other Applications of Applied Intelligent Systems. Arras, France: Springer.

Goodrich, M. A., & Schultz, A. C. (2007). Human-robot interaction: A survey. *Found. Trends Hum.-Comput. Interact.*, *1*(3), 203–275.

Gordon, G., & Breazeal, C. (2015). Bayesian active learning-based robot tutor for children's word-reading skills. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, (pp. 1343–1349). AAAI Press.

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., Das, M., & Breazeal, C. (2016). Affective personalization of a social robot tutor for children's second language skills. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, (pp. 3951–3957). AAAI Press.

Graves, A. (2013). Generating sequences with recurrent neural networks. `https://arxiv.org/abs/1308.0850`.

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. `https://arxiv.org/abs/1410.5401`.

Green, A., & Eklundh, K. S. (2003). Designing for learnability in human-robot communication. *IEEE Transactions on Industrial Electronics*, *50*(4), 644–650.

Green, A., Huttenrauch, H., & Eklundh, K. S. (2004). Applying the wizard-of-oz framework to cooperative service discovery and configuration. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, (pp. 575–580).

Gunther, M., Hu, P., Herrmann, C., Chan, C. H., Jiang, M., Yang, S., Dhamija, A. R., Ramanan, D., Beyerer, J., Kittler, J., & et al. (2017). Unconstrained face detection and open-set face recognition challenge. *2017 IEEE International Joint Conference on Biometrics (IJCB)*.

Gwinner, K., Gremler, D., & Bitner, M. (1998). Relational benefits in services industries: The customer's perspective. *Journal of the Academy of Marketing Science*, *26*(2), 101–114.

Halpern, Y., Hall, K., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., & Baeuml, M. (2016). Contextual prediction models for speech recognition. In *Proceedings of Interspeech 2016*.

Hammill, B. G., Curtis, L. H., Schulman, K. A., & Whellan, D. J. (2010). Relationship between cardiac rehabilitation and long-term risks of death and myocardial infarction among elderly medicare beneficiaries. *Circulation*, *121*(1), 63–70.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley.

Hancock, B., Bordes, A., Mazaré, P.-E., & Weston, J. (2019). Learning from dialogue after deployment: Feed yourself, chatbot! `https://arxiv.org/abs/1901.05415`.

Hanheide, M., Hebesberger, D., & Krajnik, T. (2017). The when, where, and how: An adaptive robotic info-terminal for care home residents - a long-term study. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, (pp. 341–349).

Hanheide, M., Wrede, S., Lang, C., & Sagerer, G. (2008). Who am I talking with? A face memory for social robots. In *2008 IEEE International Conference on Robotics and Automation, ICRA*, (pp. 3660–3665). Pasadena, California, USA: IEEE.

Hayashi, K., Shiomi, M., Kanda, T., & Hagita, N. (2010). Who is appropriate? a robot, human and mascot perform three troublesome tasks. In *19th International Symposium in Robot and Human Interactive Communication*, (pp. 348–354).

Hedaoo, S., Williams, A., Wadgaonkar, C., & Knight, H. (2019). A robot barista comments on its clients: Social attitudes toward robot data use. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '19, (pp. 66–74). IEEE Press.

Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: the almere model. *Int J of Soc Robotics*, *2*, 361–375.

Hemphill, C. T., Godfrey, J. J., & Doddington, G. R. (1990). The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.
URL `https://www.aclweb.org/anthology/H90-1021`

Henderson, M. (2015). Machine learning for dialog state tracking: A review. In *Proceedings of The First International Workshop on Machine Learning in Spoken Language Processing*.

Henderson, M., Thomson, B., & Williams, J. D. (2014). The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, (pp. 263–272). Philadelphia, PA, U.S.A.: Association for Computational Linguistics.

Ho, W. C., Dautenhahn, K., Lim, M. Y., & Casse, K. D. (2010). Modelling human memory in robotic companions for personalisation and long-term adaptation in hri. In A. V. Samsonovich, K. R. Johannsdottir, A. Chella, & B. Goertzel (Eds.) *BICA*, vol. 221 of *Frontiers in Artificial Intelligence and Applications*, (pp. 64–71). IOS Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, *9*(8), 1735–1780.

Honkela, A., & Valpola, H. (2003). On-line variational bayesian learning. In *In Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003*, (pp. 803–808).

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the working alliance inventory. *Journal of Counseling Psychology*, *36*(2), 223–233.

Huttenrauch, H., & Eklundh, K. S. (2002). Fetch-and-carry with cero: observations from a long-term user study with a service robot. In *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*, (pp. 158–163).

Ingrand, F., & Ghallab, M. (2017). Deliberation for autonomous robots. *Artif. Intell.*, *247*(C), 10–44.

Irfan, B. (2019a). Multi-modal personalisation in long-term human-robot interaction. In *9th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL- EpiRob 2019), Workshop on Personal Robotics and Secure Human-Robot Collaboration*.

Irfan, B. (2019b). Multi-modal personalisation in long-term human-robot interaction. In *Behaviors.AI workshop*.

Irfan, B., Céspedes Gomez, N., Casas, J., Senft, E., Gutiérrez, L. F., Rincon-Roncancio, M., Munera, M., Belpaeme, T., & Cifuentes, C. A. (2020a). Using a personalised socially assistive robot for cardiac rehabilitation: A long-term case study. In *29th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, (pp. 124–130). IEEE.

Irfan, B., Garcia Ortiz, M., Lyubova, N., & Belpaeme, T. (under review). Multi-modal incremental Bayesian network with online learning for open world user identification. *ACM Transactions on Human-Robot Interaction (THRI)*.

Irfan, B., Hellou, M., Mazel, A., & Belpaeme, T. (2020b). Challenges of a real-world HRI study with non-native english speakers: Can personalisation save the day? In *Companion of the 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM.

Irfan, B., Kennedy, J., Lemaignan, S., Papadopoulos, F., Senft, E., & Belpaeme, T. (2018a). Social psychology and human-robot interaction: An uneasy marriage. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 13–20). ACM.

Irfan, B., Lyubova, N., Garcia Ortiz, M., & Belpaeme, T. (2018b). Multi-modal open-set person identification in HRI. In *2018 ACM/IEEE International Conference on Human-Robot Interaction Social Robots in the Wild workshop*.

Irfan, B., Narayanan, A., & Kennedy, J. (2020c). Dynamic emotional language adaptation in multiparty interactions with agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20. New York, NY, USA: Association for Computing Machinery.

Irfan, B., Ramachandran, A., Spaulding, S., Glas, D. F., Leite, I., & Koay, K. L. (2019). Personalization in long-term human-robot interaction. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (pp. 685–686). IEEE.

Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*, *38*(12), 2270–2285.

Jain, A. K., Dass, S. C., & Nandakumar, K. (2004). Soft biometric traits for personal recognition systems. In *International Conference on Biometric Authentication*, no. 3072 in LNCS, (pp. 731–738). Hong Kong, China: Springer.

Jain, A. K., & Park, U. (2009). Facial marks: Soft biometric for face recognition. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, (pp. 37–40). Cairo, Egypt: IEEE Press.

Jain, A. K., Ross, A. A., & Nandakumar, K. (2011). *Introduction to Biometrics*, chap. Multibiometrics, (pp. 209–258). Springer Publishing Company, Incorporated.

Jain, L. P., Scheirer, W. J., & Boult, T. E. (2014). Multi-class open set recognition using probability of inclusion. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.) *Computer Vision – ECCV 2014*, (pp. 393–409). Zurich, Switzerland: Springer International Publishing.

Jeffreys, H. (1939). *Theory of probability*. Oxford: The Clarendon Press, 1 ed.

Jolly, K., Taylor, R., Lip, G., Greenfield, S., Raftery, J., Mant, J., Lane, D., Jones, M., Lee, K. W., & Stevens, A. (2007). The birmingham rehabilitation uptake maximisation study (brum). home-based compared with hospital-based cardiac rehabilitation in a multi-ethnic population: cost-effectiveness and patient adherence. *Health Technology Assessment*, *11*(35).

Joshi, C. K., Mi, F., & Faltings, B. (2017). Personalization in goal-oriented dialog. *arXiv*, *abs/1706.07503*.
URL `http://arxiv.org/abs/1706.07503`

Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing*, chap. Dialog Systems and Chatbots. 3 ed.

Kachouie, R., Sedighadeli, S., Khosla, R., & Chu, M.-T. (2014). Socially assistive robots in elderly care: A mixed-method systematic literature review. *International Journal of Human-Computer Interaction*, *30*(5), 369–393.

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Hum.-Comput. Interact.*, *19*(1), 61–84.

Kanda, T., Sato, R., Saiwaki, N., & Ishiguro, H. (2007). A two-month field trial in an elementary school for long-term human-robot interaction. *IEEE Transactions on Robotics*, *23*(5), 962–971.

Kanda, T., Shiomi, M., Miyashita, Z., Ishiguro, H., & Hagita, N. (2010). A communication robot in a shopping mall. *IEEE Transactions on Robotics*, *26*(5), 897–913.

Kasap, Z., & Magnenat-Thalmann, N. (2012). Building long-term relationships with virtual and robotic characters: The role of remembering. *Vis. Comput.*, *28*(1), 87–97.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015). The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, (pp. 67–74). New York, NY, USA: Association for Computing Machinery.

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., & Belpaeme, T. (2017). Child speech recognition in human-robot interaction: Evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 82–90). ACM.

Khosla, R., Nguyen, K., & Chu, M.-T. (2016). Socially assistive robot enabled personalised care for people with dementia in australian private homes. In *25th International Conference on Information Systems Development (ISD2016)*. Poland.

Kidd, C., & Breazeal, C. (2008). Robots at home: understanding long-term human-robot interaction. In *Intelligent robots and systems. IEEE/RSJ international conference on IROS 2008*, (pp. 3230–3235). New York, NY, USA: IEEE.

Kidd, C. D., & Breazeal, C. (2007). A robotic weight loss coach. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, (pp. 1985–1986). AAAI Press.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. `https://arxiv.org/abs/1412.6980`.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2016). Overcoming catastrophic forgetting in neural networks. `https://arxiv.org/abs/1612.00796`.

Kitashov, F., Svitanko, E., & Dutta, D. (2018). Foreign english accent adjustment by learning phonetic patterns. `https://arxiv.org/abs/1807.03625`.

Kocaballi, A. B., Berkovsky, S., Quiroz, J. C., Laranjo, L., Tong, H. L., Rezazadegan, D., Briatore, A., & Coiera, E. (2019). The personalization of conversational agents in health care: Systematic review. *J Med Internet Res*, *21*(11).

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*, chap. Parameter Estimation, (pp. 717–782). The MIT Press.

Kraus, W., & Keteyian, S. (2007). *Cardiac Rehabilitation*. Totowa, NJ, USA: Humana Press.

Kumar, A., Irsoy, O., Su, J., Bradbury, J., English, R., Pierce, B., Ondruska, P., Gulrajani, I., & Socher, R. (2015). Ask me anything: Dynamic memory networks for natural language processing. *CoRR*, *abs/1506.07285*.
URL `http://arxiv.org/abs/1506.07285`

Kunze, L., Hawes, N., Duckett, T., Hanheide, M., & Krajník, T. (2018). Artificial intelligence for long-term robot autonomy: A survey. *IEEE Robotics and Automation Letters*, *3*(4), 4023–4030.

Lara, J. S., Casas, J., Aguirre, A., Munera, M., Rincon-Roncancio, M., Irfan, B., Senft, E., Belpaeme, T., & Cifuentes, C. A. (2017a). Human-robot sensor interface for cardiac rehabilitation. In *2017 International Conference on Rehabilitation Robotics (ICORR)*, (pp. 1013–1018).

Lara, J. S., Casas, J., Munera, M., Senft, E., Irfan, B., Gutiérrez, L. F., Pinzón, L., Belpaeme, T., Rincon, M., & Cifuentes, C. A. (2017b). Sensor interface for cardiac rehabilitation monitoring: Pilot clinical study. In *Proceedings of the IX Congreso Iberoamericano de Tecnologías de Apoyo a la Discapacidad (IBERDISCAP 2017)*.

Lee, K.-C., & Kriegman, D. (2005). Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, (pp. 852–859). San Diego, CA, USA: IEEE.

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human-robot interaction. *International Journal of Human-Computer Studies*, *64*(10), 962–973.

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lee, M. K., Forlizzi, J., Kiesler, S., Rybski, P., Antanitis, J., & Savetsila, S. (2012). Personalization in hri: A longitudinal field experiment. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (pp. 319–326).

Lee, S. (2013). Structured discriminative model for dialog state tracking. In *Proceedings of the SIGDIAL 2013 Conference*, (pp. 442–451). Metz, France: Association for Computational Linguistics.
URL `https://www.aclweb.org/anthology/W13-4069`

Lee, S., & Eskenazi, M. (2013). Recipe for building robust spoken dialog state trackers: Dialog state tracking challenge system description. In *Proceedings of the SIGDIAL 2013 Conference*, (pp. 414–422). Metz, France: Association for Computational Linguistics.
URL `https://www.aclweb.org/anthology/W13-4066`

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2014). Empathic robots for long-term interaction. *Int J of Social Robotics*, *6*, 329–341.

Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: A survey. *International Journal of Social Robotics*, *5*(2), 291–308.

Leite, I., Pereira, A., & Lehman, J. F. (2017). Persistent memory in repeated child-robot conversations. In *Proceedings of the 2017 Conference on Interaction Design and Children*, (pp. 238–247). New York, NY, USA: ACM.

Li, J. (2015). The benefit of being physically present. *Int. J. Hum.-Comput. Stud.*, *77*(C), 23–37.

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., & Dolan, B. (2016a). A persona-based neural conversation model. `https://arxiv.org/abs/1603.06155`.

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., & Jurafsky, D. (2016b). Deep reinforcement learning for dialogue generation. `https://arxiv.org/abs/1606.01541`.

Liappas, N., Teriús-Padrón, J., Machado, E., Loghmani, M. R., García-Betances, R., Vincze, M., Quero, I., & Cabrera-Umpierrez, M. (2019). Best practices on personalization and adaptive interaction techniques in the scope of smart homes and active assisted living. In *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*. IEEE.

Lim, M. Y., Aylett, R., Vargas, P. A., Ho, W. C., & Dias, J. a. (2011). Human-like memory retrieval mechanisms for social companions. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '11, (pp. 1117–1118). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Lim, S., & Cho, S.-B. (2006). Online learning of Bayesian network parameters with incomplete data. In D.-S. Huang, K. Li, & G. W. Irwin (Eds.) *Computational Intelligence*, (pp. 309–314). Berlin, Heidelberg: Springer Berlin Heidelberg.

Liu, J., & Liao, Q. (2008). Online learning of Bayesian network parameters. In *2008 Fourth International Conference on Natural Computation*, vol. 3, (pp. 267–271). Jinan, China: IEEE.

Liu, P., Glas, D. F., Kanda, T., & Ishiguro, H. (2016). Data-driven HRI: learning social behaviors by example from human-human interaction. *IEEE Trans. Robotics*, *32*(4), 988–1008.

Liu, P., Glas, D. F., Kanda, T., & Ishiguro, H. (2018). Learning proactive behavior for interactive social robots. *Autonomous Robots*, *42*, 1067–1085.

Liu, P., Glas, D. F., Kanda, T., & Ishiguro, H. (2019). Two demonstrators are better than one-a social robot that learns to imitate people with different interaction styles. *IEEE Transactions on Cognitive and Developmental Systems*, *11*(3), 319–333.

Liu, P., Glas, D. F., Kanda, T., Ishiguro, H., & Hagita, N. (2014). How to train your robot - teaching service robots to reproduce human social behavior. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, (pp. 961–968).

Lowe, R., Pow, N., Serban, I., & Pineau, J. (2015). The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, (pp. 285–294). Prague, Czech Republic: Association for Computational Linguistics.

Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., & Pineau, J. (2016). On the evaluation of dialogue systems with next utterance classification. `https://arxiv.org/abs/1605.05414`.

Lucas, J. M., Fernandez, F., Salazar, J., Ferreiros, J., & Segundo, R. G. S. (2009). Managing speaker identity and user profiles in a spoken dialogue system. In *Procesamiento del Lenguaje Natural*, 43, (pp. 77–84).

Luo, L., Huang, W., Zeng, Q., Nie, Z., & Sun, X. (2019). Learning personalized end-to-end goal-oriented dialog. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. Honolulu, Hawaii.

Maclean, N., & Pound, P. (2000). A critical review of the concept of patient motivation in the literature on physical rehabilitation. *Social Science & Medicine*, *50*(4), 495–506.

Madotto, A., Wu, C.-S., & Fung, P. (2018). Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 1468–1478). Association for Computational Linguistics.
URL `http://aclweb.org/anthology/P18-1136`

Martinson, E., Lawson, W., & Trafton, G. (2013). Identifying people with soft-biometrics at fleet week. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, HRI '13, (pp. 49–56). Tokyo, Japan: IEEE Press.

Matarić, M., & Scassellati, B. (2016). Socially assistive robotics. In B. Siciliano, & O. Khatib (Eds.) *Springer Handbook of Robotics*, (pp. 1973–1993). Springer, Cham.

Matarić, M., Tapus, A., & Feil-Seifer, D. (2007). Personalized socially assistive robotics. In *Workshop on Intelligent Systems for Assisted Cognition*.

Matsumoto, T., Satake, S., Kanda, T., Imai, M., & Hagita, N. (2012). Do you remember that shop? computational model of spatial memory for shopping companion robots. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, (pp. 447–454). New York, NY, USA: Association for Computing Machinery.

Mavridis, N. (2015). A review of verbal and non-verbal human-robot interactive communication. *Robotics and Autonomous Systems*, *63*, 22–35.

Mavridis, N., & Petychakis, M. (2010). Human-like memory systems for interactive robots: desiderata and two case studies utilizing grounded situation models and online social networking. In *Remembering Who We Are - Human Memory for Artificial Agents Symposium*, (pp. 46–51). Leicester, UK.

Mavridis, N., & Roy, D. (2006). Grounded situation models for robots: Where words and percepts meet. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, (pp. 4690–4697).

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.) *Psychology of Learning and Motivation*, vol. 24, (pp. 109 – 165). Academic Press.

McKee, G., Biddle, M., Donnell, S. O., Mooney, M., Brien, F. O., & Moser, D. K. (2014). Cardiac rehabilitation after myocardial infarction: What influences patients' intentions to attend? *European Journal of Cardiovascular Nursing*, *13*(4), 329–337.

Mendes Júnior, P. R., de Souza, R. M., Werneck, R. d. O., Stein, B. V., Pazinato, D. V., de Almeida, W. R., Penatti, O. A. B., da S. Torres, R., & Rocha, A. (2017). Nearest neighbors distance ratio open-set classifier. *Machine Learning*, *106*, 359–386.

Mensink, T., Verbeek, J., Perronnin, F., & Csurka, G. (2013). Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(11), 2624–2637.

Michaud, F., & Théberge-Turmel, C. (2002). Mobile robotic toys and autism. In K. Dautenhahn, A. Bond, L. Cañamero, & B. Edmonds (Eds.) *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, (pp. 125–132). Boston, MA: Springer US.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. `https://arxiv.org/abs/1301.3781`.

Milch, B., & Russell, S. (2010). Extending bayesian networks to the open-universe case. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.) *Heuristics, Probability and Causality. A Tribute to Judea Pearl (Tributes)*, (pp. 217–239). College Publication.

Miller, A., Fisch, A., Dodge, J., Karimi, A.-H., Bordes, A., & Weston, J. (2016). Key-value memory networks for directly reading documents. `https://arxiv.org/abs/1606.03126`.

Miller, A. H., Feng, W., Fisch, A., Lu, J., Batra, D., Bordes, A., Parikh, D., & Weston, J. (2017). Parlai: A dialog research software platform. *CoRR*, *abs/1705.06476*.
URL `http://arxiv.org/abs/1705.06476`

Mo, K., Li, S., Zhang, Y., Li, J., & Yang, Q. (2016). Personalizing a dialogue system with transfer learning. `https://arxiv.org/abs/1610.02891`.

Moore, R. K. (2012). A bayesian explanation of the 'uncanny valley' effect and related psychological phenomena. *Scientific Reports*, *864*.

Morey, R. D., & Rouder, J. N. (2018). Bayesfactor: Computation of bayes factors for common designs.
URL `http://CRAN.R-project.org/package=BayesFactor`

Mori, M. (1970). The uncanny valley. *Energy*, *7*, 33–35.

Moro, C., Nejat, G., & Mihailidis, A. (2018). Learning and personalizing socially assistive robot behaviors to aid with activities of daily living. *J. Hum.-Robot Interact.*.

Müller, S., Sprenger, S., & Gross, H. (2014). Online adaptation of dialog strategies based on probabilistic planning. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, (pp. 692–697).

Mutlu, B., Osman, S., Forlizzi, J., Hodgins, J., & Kiesler, S. (2006). Task structure and user attributes as elements of human-robot interaction design. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, (pp. 74–79).

Navarro, D. (2018). *Learning Statistics with R: A tutorial for psychology students and other beginners*, chap. Bayesian statistics. Danielle Navarro, 0.6 ed.
URL `https://bookdown.org/ekothe/navarro26/bayes.html`

Neal, R. M., & Hinton, G. E. (1999). *A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants*, (pp. 355–368). Cambridge, MA, USA: MIT Press.

Neururer, M., Schlögl, S., Brinkschulte, L., & Groth, A. (2018). Perceptions on authenticity in chat bots. *Multimodal Technologies Interact*, *2*(3), 60–78.

Ng, H. G., Anton, P., Brügger, M., Churamani, N., Fließwasser, E., Hummel, T., Mayer, J., Mustafa, W., Nguyen, T. L. C., Nguyen, Q., Soll, M., Springenberg, S., Griffiths, S., Heinrich, S., Navarro-Guerrero, N., Strahl, E., Twiefel, J., Weber, C., & Wermter, S. (2017). Hey robot, why don't you talk to me? In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, (pp. 728–731).

Niemelä, M., Heikkilä, P., Lammi, H., & Oksman, V. (2019). A social robot in a shopping mall: Studies on acceptance and stakeholder expectations. In O. Korn (Ed.) *Social Robots*, Human-Computer Interaction Series, (pp. 119–144). Germany: Springer.

Okamura, A. M., Matarić, M. J., & Christensen, H. I. (2010). Medical and health-care robotics. *IEEE Robotics Automation Magazine*, *17*(3), 26–37.

Opper, M., & Winther, O. (1999). A bayesian approach to on-line learning. In D. Saad (Ed.) *On-line learning in neural networks*, Publications of the Newton Institute, (pp. 363–378). Cambridge: Cambridge University Press. Copyright of Cambridge University Press Available on Google Books.

Oravecz, Z., Huentelman, M., & Vandekerckhove, J. (2016). Sequential bayesian updating for big data. In M. Jones (Ed.) *Big Data in Cognitive Science*, Frontiers of Cognitive Psychology, (pp. 13–33). Taylor & Francis.
URL `https://books.google.fr/books?id=Dy1uDQAAQBAJ`

Ouellet, S., Grondin, F., Leconte, F., & Michaud, F. (2014). Multimodal biometric identification system for mobile robots combining human metrology to face recognition and speaker identification. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, (pp. 323–328). Edinburgh, Scotland: IEEE.

Pandey, A. K., & Gelin, R. (2018). A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics and Automation Magazine*, *25*(3), 40–48.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Pargellis, A. N., Kuo, H.-K. J., & Lee, C.-H. (2004). An automatic dialogue generation platform for personalized dialogue applications. *Speech Communication*, *42*(3), 329–351.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, *113*, 54–71.

Park, U., & Jain, A. K. (2010). Face matching and retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security*, *5*(3), 406–415.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In X. Xie, M. W. Jones, & G. K. L. Tam (Eds.) *British Machine Vision Conference*, (pp. 41.1–41.12). Swansea, UK: BMVA Press.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics.

Phillips, P. J., Grother, P., & Micheals, R. (2011). Evaluation methods in face recognition. In S. Z. Li, & A. K. Jain (Eds.) *Handbook of Face Recognition*, (pp. 553–556). Springer Publishing Company, Incorporated, 2nd ed.

Piepoli, M. F., Corrà, U., Benzer, W., Bjarnason-Wehrens, B., Dendale, P., Gaita, D., McGee, H., Mendes, M., Niebauer, J., Olsen Zwisler, A.-D., Schmid, J.-P., & Cardiac Rehabilitation Section of the European Association of Cardiovascular Prevention and Rehabilitation (2010). Secondary prevention through cardiac rehabilitation: From knowledge to implementation. a position paper from the cardiac rehabilitation section of the european association of cardiovascular prevention and rehabilitation. *Eur J Cardiovasc Prev Rehabil*, *17*(1), 1–17.

Pieraccini, R., Levin, E., & Eckert, W. (1997). Amica: The at&t mixed initiative conversational architecture. In *Proceedings of the European Conference on Speech Communication and Technology*, (pp. 1875–1878).

Prassler, E., Munich, M. E., Pirjanian, P., & Kosuge, K. (2016). Domestic robotics. In B. Siciliano, & O. Khatib (Eds.) *Springer Handbook of Robotics*, (pp. 1729–1758). Springer Handbooks.

Prescott, T., & Caleb-Solly, P. (2017). Robotics in social care: A connected care ecosystem for independent living. *UK-RAS Network Robotics & Autonomous Systems*.
URL `https://www.ukras.org/wp-content/uploads/2018/10/UK_RAS_wp_social_spread_low_res_ref.pdf`

Rajendran, J., Ganhotra, J., Singh, S., & Polymenakos, L. (2018). Learning end-to-end goal-oriented dialog with multiple answers. `https://arxiv.org/abs/1808.09996`.

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., King, E., Bland, K., Wartick, A., Pan, Y., Song, H., Jayadevan, S., Hwang, G., & Pettigrue, A. (2018). Conversational ai: The science behind the alexa prize. `https://arxiv.org/abs/1801.03604`.

Read, R., & Belpaeme, T. (2012). How to use non-linguistic utterances to convey emotion in child-robot interaction. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, (pp. 219–220). New York, NY, USA: Association for Computing Machinery.

Read, R., & Belpaeme, T. (2016). People interpret robotic non-linguistic utterances categorically. *International Journal of Social Robotics*, *8*(1), 31–50.

Ren, H., Xu, W., Zhang, Y., & Yan, Y. (2013). Dialog state tracking using conditional random fields. In *Proceedings of the SIGDIAL 2013 Conference*, (pp. 457–461). Metz, France: Association for Computational Linguistics.
URL `https://www.aclweb.org/anthology/W13-4071`

Rich, E. (1979). User modeling via stereotypes. *Cognitive Science*, *3*(4), 329–354.

Richardson, K., Coeckelbergh, M., Wakunuma, K., Billing, E., Ziemke, T., Gomez, P., Vanderborght, B., & Belpaeme, T. (2018). Robot enhanced therapy for children with autism (DREAM): A social model of autism. *IEEE Technology and Society Magazine*, *37*(1), 30–39.

Rickert, M., Foster, M. E., Giuliani, M., By, T., Panin, G., & Knoll, A. (2007). Integrating language, vision and action for human robot dialog systems. In C. Stephanidis (Ed.) *Universal Access in Human-Computer Interaction. Ambient Interaction. UAHCI 2007.*, vol. 4555 of *Lecture Notes in Computer Science*, (pp. 987–995). Springer, Berlin, Heidelberg.

Riek, L. D. (2012). Wizard of oz studies in hri: A systematic review and new reporting guidelines. *J. Hum.-Robot Interact.*, *1*(1), 119–136.

Riek, L. D. (2017). Healthcare robotics. *Commun. ACM*, *60*(11), 68–78.

Robins, B., Dautenhahn, K., Boekhorst, R., Billard, A., Keates, S., Clarkson, P., Langdon, P., & Robinson, P. (2004). Effects of repeated exposure to a humanoid robot on children with autism. In *Designing a More Inclusive World*. London: Springer.

Robins, B., Dautenhahn, K., Boekhorst, R. T., & Billard, A. (2005). Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society*, *4*(2), 105–120.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y.-L., & Weston, J. (2020). Recipes for building an open-domain chatbot. `https://arxiv.org/abs/2004.13637`.

Rosa, R. D., Orabona, F., & Cesa-Bianchi, N. (2015). The abacoc algorithm: A novel approach for nonparametric classification of data streams. In *2015 IEEE International Conference on Data Mining*, (pp. 733–738). Atlantic City, NJ, USA: IEEE.

Ross, A., & Jain, A. K. (2007). Human recognition using biometrics: an overview. *Annales Des Télécommunications*, *62*(1), 11–35.

Rothe, R., Timofte, R., & Gool, L. V. (2015). Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

Rothe, R., Timofte, R., & Gool, L. V. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, *126*(2-4), 144–157.

Roy, N., Baltus, G., Fox, D., Gemperle, F., Goetz, J., Hirsch, T., Margaritis, D., Montemerlo, M., Pineau, J., Schulte, J., & Thrun, S. (2000). Towards personal service robots for the elderly. In *Workshop on Interactive Robots and Entertainment (WIRE 2000)*.

Ruano-Ravina, A., Pena-Gil, C., Abu-Assi, E., Raposeiras, S., van 't Hof, A., Meindersma, E., Bossano Prescott, E. I., & González-Juanatey, J. R. (2016). Participation and adherence to cardiac rehabilitation programs. a systematic review. *Int J Cardiol*, *223*, 436–443.

Rudd, E. M., Jain, L. P., Scheirer, W. J., & Boult, T. E. (2018). The extreme value machine. *IEEE transactions on pattern analysis and machine intelligence*, *40*(3), 762–768.

Rudovic, O., Lee, J., Dai, M., Schuller, B., & Picard, R. W. (2018). Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, *3*(19).

Sabelli, A. M., Kanda, T., & Hagita, N. (2011). A conversational robot in an elderly care center: An ethnographic study. In *Proceedings of the 6th International Conference on Human-Robot Interaction*, HRI '11, (pp. 37–44). New York, NY, USA: Association for Computing Machinery.

Sadek, M. D., Bretier, P., & Panaget, F. (1997). Artimis: Natural dialogue meets rational agency. In *Proceedings of the Fifteenth International Joint Conference on Artifical Intelligence - Volume 2*, IJCAI'97, (pp. 1030–1035). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Sadhya, D., Pahariya, P., Yadav, R., Rastogi, A., Kumar, A., Sharma, L., & Singh, S. K. (2017). Biosoft - a multimodal biometric database incorporating soft traits. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, (pp. 1–6). New Delhi, India: IEEE.

Salter, T., Dautenhahn, K., & Bockhorst, R. (2004). Robots moving out of the laboratory - detecting interaction levels and human contact in noisy school environments. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, (pp. 563–568).

Saunderson, S., & Nejat, G. (2019). How robots influence humans: A survey of nonverbal communication in social human-robot interaction. *International Journal of Social Robotics*, *11*(4), 575–608.

Scane, K., Alter, D., Oh, P., & Brooks, D. (2012). Adherence to a cardiac rehabilitation home program model of care: a comparison to a well-established traditional on-site supervised program. *Appl Physiol Nutr Metab*, *37*(2), 206–213.

Scassellati, B. (2007). How social robots will help us to diagnose, treat, and understand autism. In S. Thrun, R. Brooks, & H. Durrant-Whyte (Eds.) *Robotics Research*, (pp. 552–563). Berlin, Heidelberg: Springer Berlin Heidelberg.

Scassellati, B., Boccanfuso, L., Huang, C.-M., Mademtzi, M., Qin, M., Salomons, N., Ventola, P., & Shic, F. (2018). Improving social skills in children with asd using a long-term, in-home social robot. *Science Robotics*, *3*(21).

Scheeff, M., Pinto, J., Rahardja, K., Snibbe, S., & Tow, R. (2002). Experiences with sparky, a social robot. In K. Dautenhahn, A. Bond, L. Cañamero, & B. Edmonds (Eds.) *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, (pp. 173–180). Boston, MA: Springer US.

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., & Boult, T. E. (2013). Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(7), 1757–1772.

Scheirer, W. J., Jain, L. P., & Boult, T. E. (2014). Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(11), 2317–2324.

Scheirer, W. J., Kumar, N., Ricanek, K., Belhumeur, P. N., & Boult, T. E. (2011). Fusing with context: A Bayesian approach to combining descriptive attributes. In *2011 International Joint Conference on Biometrics (IJCB)*, (pp. 1–8). Washington, DC, USA: IEEE.

Schneider, S., Goerlich, M., & Kummert, F. (2017). A framework for designing socially assistive robot interactions. *Cognitive Systems Research*, *43*, 301–312.

Schodde, T., Bergmann, K., & Kopp, S. (2017). Adaptive robot language tutoring based on bayesian knowledge tracing and predictive decision-making. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, (pp. 128–136).

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Senft, E., Baxter, P., Kennedy, J., & Belpaeme, T. (2015). SPARC: Supervised Progressively Autonomous Robot Competencies. In *Proceedings of the International Conference on Social Robotics (ICSR)*, (pp. 603–612). Springer.

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., & Pineau, J. (2015). A survey of available corpora for building data-driven dialogue systems. *CoRR*.

Severinson-Eklundh, K., Green, A., & Hüttenrauch, H. (2003). Social and collaborative aspects of interaction with a service robot. *Robotics and Autonomous Systems*, *42*(3), 223–234. Socially Interactive Robots.

Shahsavari, H., Shahriari, M., & Alimohammadi, N. (2012). Motivational factors of adherence to cardiac rehabilitation. *Iranian journal of nursing and midwifery research*, *17*(4), 318–324.

She, J., Nakamura, H., Makino, K., Ohyama, Y., & Hashimoto, H. (2014). Selection of suitable maximum-heart-rate formulas for use with Karvonen formula to calculate exercise intensity. *International Journal of Automation and Computing*, *12*(1), 62–69.

Shin, D., & Choo, H. (2011). Modeling the acceptance of socially interactive robotics: Social presence in human-robot interaction. *Interaction Studies*, *12*, 430–460.

Shiomi, M., Sakamoto, D., Kanda, T., Ishi, C. T., Ishiguro, H., & Hagita, N. (2008). A semi-autonomous communication robot - a field trial at a train station. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (pp. 303–310).

Shu, L., Xu, H., & Liu, B. (2017). Doc: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (pp. 2911–2916). Association for Computational Linguistics.

Shum, H.-Y., He, X., & Li, D. (2018). From Eliza to Xiaoice: Challenges and opportunities with social chatbots. `https://arxiv.org/abs/1801.01957`.

Siegert, R. J., & Taylor, W. J. (2004). Theoretical aspects of goal-setting and motivation in rehabilitation. *Disabil Rehabil.*, *26*(1), 1–8.

Simms, K., Myers, C., Adams, J., Hartman, J., Lindsey, C., Doler, M., & Suhr, J. (2007). Exercise tolerance testing in a cardiac rehabilitation setting: an exploratory study of its safety and practicality for exercise prescription and outcome data collection. *Baylor University. Medical Center*, *20*(4), 344–347.

Simpson, A., & Eraser, N. M. (1993). Black box and glass box evaluation of the sundial system. In *Third European Conference on Speech Communication and Technology*.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., & Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 196–205). Denver, Colorado: Association for Computational Linguistics.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, HRI '06, (pp. 33–40). New York, NY, USA: Association for Computing Machinery.

Stent, A., Dowding, J., Gawron, J. M., Bratt, E. O., & Moore, R. (1999). The commandtalk spoken dialogue system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, (pp. 183–190). USA: Association for Computational Linguistics.

Suaya, J., Stason, W., Ades, P., Normand, S., & Shepard, D. (2009). Cardiac rehabilitation and survival in older coronary patients. *J Am Coll Cardiol*, *54*(1), 25–33.

Suguna, N., & Thanushkodi, K. (2010). An improved k-nearest neighbor classification using genetic algorithm. *International Journal of Computer Science*, *7*, 18–21.

Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.) *Advances in Neural Information Processing Systems 28*, (pp. 2440–2448). Curran Associates, Inc. URL `http://papers.nips.cc/paper/5846-end-to-end-memory-networks.pdf`

Sun, K., Chen, L., Zhu, S., & Yu, K. (2014). A generalized rule based tracker for dialogue state tracking. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, (pp. 330–335).

Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1891–1898). Columbus, OH, USA: IEEE.

Sung, J., Christensen, H. I., & Grinter, R. E. (2009). Robots in the wild: Understanding long-term use. In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (pp. 45–52).

Süssenbach, L., Riether, N., Schneider, S., Berger, I., Kummert, F., Lütkebohle, I., & Pitsch, K. (2014). A robot as fitness companion: Towards an interactive action-based motivation model. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, (pp. 286–293).

Sutskever, I., Vinyals, O., & V. Le, Q. (2014). Sequence to sequence learning with neural networks. `https://arxiv.org/abs/1409.3215`.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, (pp. 1701–1708). Washington, DC, USA: IEEE Computer Society.

Tanaka, F., Movellan, J. R., Fortenberry, B., & Aisaka, K. (2006). Daily hri evaluation at a classroom environment: Reports from dance interaction experiments. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, HRI '06, (pp. 3–9). New York, NY, USA: Association for Computing Machinery.

Tapus, A. (2009). Improving the quality of life of people with dementia through the use of socially assistive robots. In *2009 Advanced Technologies for Enhanced Quality of Life*, (pp. 81–86).

Tapus, A., Matarić, M. J., & Scassellati, B. (2007). Socially assistive robotics [grand challenges of robotics]. *IEEE Robotics Automation Magazine*, *14*(1), 35–42.

Tapus, A., Ţăpuş, C., & Matarić, M. J. (2008). User-robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intelligent Service Robotics*, *1*(2).

Thill, S., Pop, C. A., Belpaeme, T., Ziemke, T., & Vanderborght, B. (2012). Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook. *Paladyn*, *3*, 209–217.

Thomas, R. J., King, M., Lui, K., Oldridge, N., Piña, I. L., Spertus, J., & ACC/AHA Task Force Members (2007). Aacvpr/acc/aha 2007 performance measures on cardiac rehabilitation for referral to and delivery of cardiac rehabilitation/secondary prevention services. *J Cardiopulm Rehabil Prev*, *27*(5), 260–290.

Thompson, C. A., Göker, M. H., & Langley, P. (2004a). A personalized system for conversational recommendations. *J. Artif. Int. Res.*, *21*(1), 393–428.
URL http://dl.acm.org/citation.cfm?id=1622467.1622479

Thompson, C. A., Göker, M. H., & Langley, P. (2004b). A personalized system for conversational recommendations. *J. Artif. Int. Res.*, *21*(1), 393–428.

Togneri, R., & Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits and Systems Magazine*, *11*(2), 23–61.

Tokunaga, S., Tamamizu, K., Saiki, S., Nakamura, M., & Yasuda, K. (2017). Virtualcaregiver: Personalized smart elderly care. *IJSI*, *5*, 30–43.

Traum, D. R. (1999). *Foundations of Rational Agency*, chap. Speech acts for dialogue agents, (pp. 169–201). Springer.

Triantafillou, E., Zemel, R., & Urtasun, R. (2017). Few-shot learning through an information retrieval lens. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.) *Advances in Neural Information Processing Systems 30*, (pp. 2255–2265). Curran Associates, Inc.
URL http://papers.nips.cc/paper/6820-few-shot-learning-through-an-information-retri
pdf

Tur, G. (2011). *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley and Sons.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Turk-Adawi, K. I., Oldridge, N. B., Tarima, S. S., Stason, W. B., & Shepard, D. S. (2013). Cardiac rehabilitation patient and organizational factors: What keeps patients in programs? *Journal of the American Heart Association*, *2*(5), e000418.

UK, A. (2017). Briefing: Health and care of older people in england 2017.
URL https://www.ageuk.org.uk/Documents/EN-GB/For-professionals/Research/The_Health_and_Care_of_Older_People_in_England_2016.pdf?dtrk=true

Vasco, V., Willems, C., Chevalier, P., De Tommaso, D., Gower, V., Gramatica, F., Tikhanoff, V., Pattacini, U., Metta, G., & Wykowska, A. (2019). International conference on social robotics (icsr 2019). In *Train with Me: A Study Comparing a Socially Assistive Robot and a Virtual Agent for a Rehabilitation Task*. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.) *Advances in Neural Information Processing Systems 30*, (pp. 5998–6008). Curran Associates, Inc.
URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478.

Verlinde, P., Druyts, P., Cholet, G., & Acheroy., M. (1999). Applying Bayes based classifiers for decision fusion in a multi-modal identity verification system. In *Intl. Symposium. on Pattern Recognition*.

Vinyals, O., & Le, Q. V. (2015). A neural conversational model. *CoRR*, *abs/1506.05869*.

Wada, K., & Shibata, T. (2007). Living with seal robots-its sociopsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics*, *23*(5), 972–980.

Wainer, J., Feil-seifer, D. J., Shell, D. A., & Mataric, M. J. (2006). The role of physical embodiment in human-robot interaction. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, (pp. 117–122).

Wallbridge, C. D., Lemaignan, S., & Belpaeme, T. (2017). Qualitative review of object recognition techniques for tabletop manipulation. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, (pp. 359–363). New York, NY, USA: Association for Computing Machinery.

Wang, L., Lei, B., Li, Q., Su, H., Zhu, J., & Zhong, Y. (2020). Triple memory networks: a brain-inspired method for continual learning. `https://arxiv.org/abs/2003.03143`.

Wang, Y., & Wang, Z. (2007). A fast knn algorithm for text categorization. In *2007 International Conference on Machine Learning and Cybernetics*, vol. 6, (pp. 3436–3441).

Werry, I., Dautenhahn, K., Ogden, B., & Harwin, W. (2001). Can social interaction skills be taught by a social agent? the role of a robotic mediator in autism therapy. In *Cognitive Technology: Instruments of Mind. CT 2001. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer.

Weston, J., Chopra, S., & Bordes, A. (2014). Memory networks. `https://arxiv.org/abs/1410.3916`.

Wilcox, R. (2017). *Introduction to Robust Estimation and Hypothesis Testing*. Elsevier, 4 ed.

Williams, J. (2013). Multi-domain learning and generalization in dialog state tracking. In *Proceedings of the SIGDIAL 2013 Conference*, (pp. 433–441). Metz, France: Association for Computational Linguistics.
URL `https://www.aclweb.org/anthology/W13-4068`

Williams, J. D. (2014). Web-style ranking and SLU combination for dialog state tracking. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, (pp. 282–291). Philadelphia, PA, U.S.A.: Association for Computational Linguistics.

Williams, T., Thames, D., Novakoff, J., & Scheutz, M. (2018). "Thank you for sharing that interesting fact!": Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, (pp. 298–306). New York, NY, USA: Association for Computing Machinery.

Winkle, K., Caleb-Solly, P., Turton, A., & Bremner, P. (2018). Social robots for engagement in rehabilitative therapies: Design implications from a study with therapists. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, (pp. 289–297). New York, NY, USA: Association for Computing Machinery.

Winograd, T., & Flores, F. (1986). *Understanding Computers and Cognition: A New Foundation for Design*. Northwood, NJ: Ablex Publishing.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell. (Translated by Anscombe, G.E.M.).

Wójcik, W., Gromaszek, K., & Junisbekov, M. (2016). Face recognition: Issues, methods and alternative applications. In S. Ramakrishnan (Ed.) *Face Recognition - Semisupervised Classification, Subspace Projection and Evaluation Methods*, chap. 02. InTech.

Wolf, T., Sanh, V., Chaumond, J., & Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, *abs/1901.08149*. URL `http://arxiv.org/abs/1901.08149`

Wood, R., Baxter, P., & Belpaeme, T. (2011). A review of long-term memory in natural and synthetic systems. *Adaptive Behavior*, *20*(2), 81–103.

Woodworth, B., Ferrari, F., Zosa, T. E., & Riek, L. D. (2018). Preference learning in assistive robotics: Observational repeated inverse reinforcement learning. In *Machine Learning Research*, vol. 85, (pp. 1–19).

Wu, L., Fisch, A., Chopra, S., Adams, K., Bordes, A., & Weston, J. (2017). Starspace: Embed all the things! `https://arxiv.org/abs/1709.03856`.

Xiong, C., Merity, S., & Socher, R. (2016). Dynamic memory networks for visual and textual question answering. In M. F. Balcan, & K. Q. Weinberger (Eds.) *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48 of *Proceedings of Machine Learning Research*, (pp. 2397–2406). New York, New York, USA: PMLR.

Yan, R. (2018). "chitty-chitty-chat bot": Deep learning for conversational ai. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, (pp. 5520–5526). International Joint Conferences on Artificial Intelligence Organization.

Young, S., Gašić, M., Thomson, B., & Williams, J. D. (2013). Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, *101*(5), 1160–1179.

Young, V., & Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: a literature review. *Assist Technol.*, *22*(2), 99–112.

Yuan, X. (2015). An approach to integrating emotion in dialogue management. In Y. Tan, Y. Shi, F. Buarque, A. Gelbukh, S. Das, & A. Engelbrecht (Eds.) *Advances in Swarm and Computational Intelligence*, (pp. 297–308). Cham: Springer International Publishing.

Yusupov, I., & Kuratov, Y. (2018). NIPS conversational intelligence challenge 2017 winner system: Skill-based conversational agent with supervised dialog manager. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.) *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, (pp. 3681–3692). Association for Computational Linguistics.

Zewail, R., Elsafi, A., Saeb, M., & Hamdy, N. (2004). Soft and hard biometrics fusion for improved identity verification. In *Circuits and Systems, 2004. MWSCAS '04. The 2004 47th Midwest Symposium on*, vol. 1, (pp. I–225). Hiroshima, Japan: IEEE.

Zhang, B., Xu, X., Li, X., Ye, Y., Chen, X., & L., S. (2019). Learning personalized end-to-end task-oriented dialogue generation. In J. Tang, M. Y. Kan, D. Zhao, S. Li, & H. Zan (Eds.) *Natural Language Processing and Chinese Computing (NLPCC 2019)*, vol. 11838 of *Lecture Notes in Computer Science*. Springer, Cham.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? `https://arxiv.org/abs/1801.07243`.

Zheng, X., Glas, D. F., Minato, T., & Ishiguro, H. (2019). Four memory categories to support socially- appropriate conversations in long-term hri. In *2019 ACM/IEEE International Conference on Human-Robot Interaction Personalization in Long-Term Human-Robot Interaction (PLOT-HRI19) workshop*. Daegu, South Korea.

Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2018). The design and implementation of xiaoice, an empathetic social chatbot. `https://arxiv.org/abs/1812.08989`.

Zhou, Y., & Huang, T. S. (2006). Weighted Bayesian network for visual tracking. In *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1, (pp. 523–526). Hong Kong, China: IEEE.

Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C., & Manning, J. R. (2018). Is automatic speech-to-text transcription ready for use in psychological experiments? *Behavior Research Methods*, *50*, 2597–2605.

Appendices

Appendix A

Information Form for User Recognition Study

This appendix section presents the self consent and information form used in the long-term user study in recognition, presented in Chapter 4.

PLYMOUTH UNIVERSITY

FACULTY OF SCIENCE AND ENGINEERING

CONSENT TO PARTICIPATE IN RESEARCH PROJECT

Name of Principal Investigator
Tony Belpaeme

Title of Research
Social Learning through Contingent Interaction

Aim of research
The aim of the research is to test the recognition rate of the multi-modal person recognition system.

Description of procedure
You will initially register into the system with your name, gender, birth year, height through the tablet interface of Pepper robot, and it will take a picture at the end of the registration. The robot will use this information to guess who you are in the following encounters and it will ask for confirmation of the guessed name through the tablet interface. The recognition will happen at least every half an hour when the robot sees you, unless the head of the robot is touched. The study will last four weeks. You will be asked to complete a questionnaire at the end of the study about your opinions of the recognition system. Consent form includes an option to opt-in for sharing of your images in academic publications and presentations.

Description of risks
Risks are minimal – there will be no physical contact with any other people, and the only physical contact with the robot is through the tablet interface on the robot.

Right to withdraw
You may withdraw from the study at any time – simply inform the experimenter and you will be withdrawn from the study. Any data collected at this stage will be destroyed if requested.

The objectives of this research have been explained to me.

I understand that I am free to withdraw from the research at any stage, and ask for my data to be destroyed if I wish.

I understand that the personal information will be anonymized and will be securely stored on university servers.

I understand that the Principal Investigator of this work will have attempted, as far as possible, to avoid any risks, and that safety and health risks will have been separately assessed by appropriate authorities (e.g. under COSHH regulations)

If you are dissatisfied with the way the research is conducted, please contact the experimenter, Bahar Irfan, in the first instance through email on bahar.irfan@plymouth.ac.uk . If you feel the problem has not been resolved please contact the secretary to the Faculty of Science and Environment Human Ethics Committee: Mrs Paula Simson 01752 584503.

We might be asked to share images when we present or publish our work. Please **tick one of the following boxes regarding your consent**:

I consent to the sharing of my images. ☐

I do not wish the researchers to share the images. ☐

Under these circumstances, I agree to participate in the research.

Name: ……………………………………….

Signature: ....……………..   Date: ………………

Appendix B

Additional Evaluations on Multi-modal Long-Term User Recognition Dataset

This section presents the additional findings on multi-modal long-term user recognition dataset, presented in Chapter 5. In addition, we describe our modifications on the Extreme Value Machine (EVM) algorithm to accept sequential increments, along with the hyperparameters we used in the comparisons.

B.1 Evaluation of Normalisation Methods in Ten Samples Dataset

Comparison of the long-term recognition performance loss between the normalisation methods for the ten samples dataset (D-Ten) is shown in Figure B.1. Similar to the results for the all samples dataset (D-All) shown in Chapter 5, hybrid normalisation achieves significantly lower loss than the other normalisation methods within all conditions. Further analysis can be found in Section 5.5.1.3.

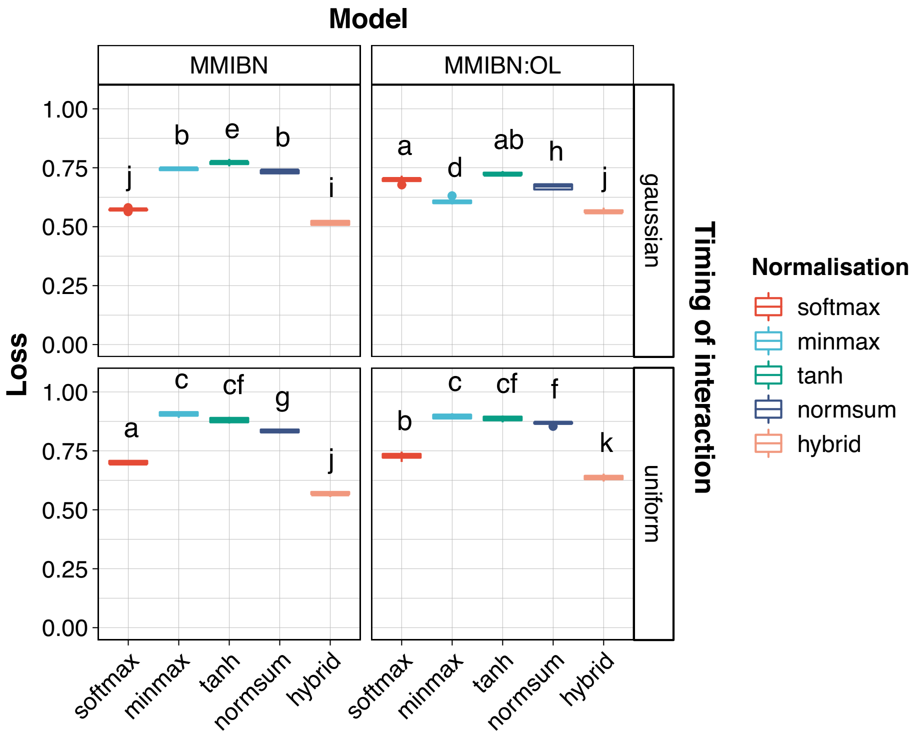


Figure B.1: Results of Tukey's Honestly Significant Differences (HSD) test of loss in the open-set for normalisation methods with optimised weights for the ten samples dataset (D-Ten) dataset: softmax, minmax, tanh, normsum, and hybrid. Lower loss is better.

B.2 Time Plot for Open-Set Recognition

The time plot for open-set recognition in Figure B.2 shows the change in long-term recognition loss with the increasing number of recognitions. The results are consistent with the results for the training set, presented in Section 5.5.2.1. Multi-modal Incremental Bayesian Network (MMIBN) and Multi-modal Incremental Bayesian Network with Online Learning (MMIBN:OL) have higher loss in the open-set compared to the training, due to the higher number of users to recognise. Extreme Value Machine trained with face recognition data (EVM:FR) has a lower loss during the enrolment period due to lower False Alarm Rate (FAR) compared to MMIBN models, and a higher Detection and Identification Rate (DIR) compared to Extreme Value Machine trained with multi-modal data (EVM:MM), but the MMIBN models significantly outperform it overall and in the closed-set, as shown in Figure 5.6.

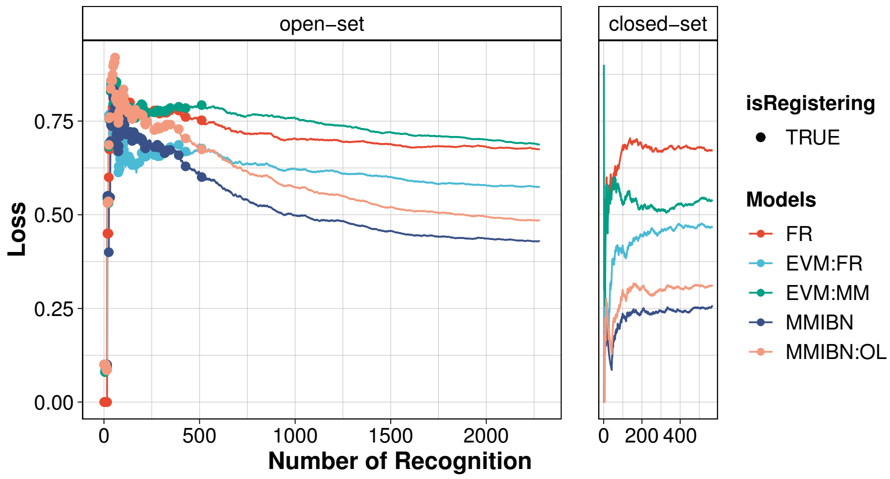


Figure B.2: The change of loss with increasing number of recognitions for the all samples dataset with Gaussian times (D-All$_{\text{Gaussian}}$) for open-set and closed-set (open). The loss decreases with increasing number of recognitions.

B.3 Adapting Extreme Value Machine for Sequential Learning

EVM[1] (Rudd et al. 2018) is a state-of-the-art open world classifier based on the Extreme Value Theory, as described in Section 2.2. However, it was only evaluated using batch learning, which is not suitable in real-world human-robot interaction, because the users will be encountered sequentially. Hence, we transformed the method for accepting sequential data and incremental online learning to compare its performance to MMIBN.

The hyperparameters of EVM are tail size (τ, the number of points that constitute extrema for Extreme Value Theory), number of models to average (k), coverage threshold (σ, probabilistic threshold to designate redundancy between points), and open set threshold (δ, if maximum probability is below this threshold, the identity is estimated as unknown). The ranges considered for these hyperparameters in (Rudd et al. 2018) are as follows: 100-32000 for τ (can be minimum 2), 1-10 for k, $[0.008, 0.186, 0.492, 1.0]$ for σ, and $[0.05, 0.1, ..., 0.3]$ for δ. Moreover, Euclidean distance or cosine similarity can be used as the distance function to compute margins for the EVM.

As described in Section 3.2.4, we set MMIBN to declare the user as unknown in the first 5 recognitions, in order to allow the network to make meaningful estimations. This was achieved for EVM by setting $\tau = 3$. After the initial training, sequential learning is achieved with updating the model with a single data point (i.e., a single recognition) at each recognition, by setting $k = 1$. We optimised σ and δ over the ranges given, and found that $\sigma = 1.0$ and $\delta = 0.05$ resulted in the lowest long-term performance loss. Cosine similarity is used as the distance function, as it is stated in (Rudd et al. 2018) that Euclidean distance led to poor performance for EVM. It is important to note that in (Rudd et al. 2018), $\tau = 33998$, $k = 6$, and $\sigma = 0.5$. However, the authors stated that σ and k had a slight impact on performance (2% increase in accuracy and F1 score), whereas, the vast majority of performance variation was attributed to τ.

We use the same data with the structure described in Section 5.4.1 for evaluating MMIBN and EVM models. Note that for EVM models, we only normalise the soft biometrics using norm-sum (dividing by the total sum). The reason is that hybrid normalisation is optimised for MMIBN and results in poor performance in EVM.

[1]https://github.com/EMRResearch/ExtremeValueMachine

Appendix C

Information Form and Questionnaires for Barista Robot Study

This appendix section presents the self consent and information form and the evaluation questionnaires used in the barista robot study, presented in Chapter 7. The questionnaire composed of three parts and varied according to the conditions: (A) perception of the robot based on Robotic Social Attributes Scale (RoSAS) (Carpinella et al. 2017), (B) task-specific questions to evaluate the capabilities of the robot, (C) additional questions for obtaining user attributes, user experience and an additional comments. The italicised questions in section (B) are asked only in the personalised condition, whereas the ones in section (C) are only asked after the first interaction of the user (enrolment condition).

UNIVERSITY OF PLYMOUTH

FACULTY OF SCIENCE AND ENGINEERING

CONSENT TO PARTICIPATE IN RESEARCH PROJECT

Name of Principal Investigator
Tony Belpaeme

Title of Research
Social Learning through Contingent Interaction

Aim of research
The aim of the research is to evaluate repeated interactions with a robot in a coffee-shop scenario.

Description of procedure
You will greet the robot to start the interaction, with a phrase such as “Hi” or “Hello”. Upon greeting, the robot will ask you for your drink choice. You should respond accordingly to the questions of the robot to have a smooth interaction. The interaction will follow the steps of order taking in a coffee-shop scenario. The interaction will be autonomous and the experimenter will not be present during the interaction. Please make sure to speak clear and loud (without shouting) with a normal speaking pace for achieving the optimal speech recognition. The robot might ask you to repeat or rephrase your responses if there are problems. You will receive a free drink and a snack of your choice at each interaction with the robot as a compensation for your participation. If you change your mind about your order and would like to change it, you can do so after the robot asks for the confirmation of the order. The robot will update your order accordingly.

Please do not speak when the robot is speaking, as the robot cannot listen to you while it speaks. When the robot asks, please enter your full name (first and last name with a space between them) on the tablet.

The robot may be interacted at most two times a day. The study will last 5 days. You will be asked to complete a questionnaire after the first, third and the last day of the study about your opinions of the robot. The interaction will be video recorded. Consent form includes an option to opt-in for **sharing** of your images and videos in academic publications and presentations.

Please keep in mind that taking photographs or videos is strictly forbidden for the participants to protect the anonymity of the participants and the intellectual property of this research.

Description of risks
Risks are minimal – there will be no physical contact with any other people, and the only physical contact with the robot is through the tablet interface on the robot.

Right to withdraw
You may withdraw from the study at any time – simply inform the experimenter and you will be withdrawn from the study. Any data collected at this stage will be destroyed if requested.

The objectives of this research have been explained to me.

I understand that I am free to withdraw from the research at any stage, and ask for my data to be destroyed if I wish.

I understand that the interaction will be video recorded. These videos will not be distributed and will primarily be used for analysis of the interaction. I understand that the personal information will be anonymized and will be securely stored on university servers.

I understand that the Principal Investigator of this work will have attempted, as far as possible, to avoid any risks, and that safety and health risks will have been separately assessed by appropriate authorities (e.g. under COSHH regulations)

If you are dissatisfied with the way the research is conducted, please contact the experimenter, Bahar Irfan, in the first instance through email on bahar.irfan@plymouth.ac.uk . If you feel the problem has not been resolved please contact the secretary to the Faculty of Science and Environment Human Ethics Committee: Mrs Paula Simson +441752 584503.

– –

We might be asked to share images and short clips of videos, when we present or publish our work. Please **tick one of the following boxes regarding your consent**:

I consent to the sharing of my images and video. ☐

I do not wish the researchers to share the images and video. ☐

Under these circumstances, I agree to participate in the research.

Name: ……………………………………….

Signature: ……………………………………………. Date: ………………

Barista Robot Evaluation Form

Is this your first, third, or last day of interaction with the robot? Please circle the correct one.

Please give ratings based on your overall experience to the questions below.

A. Using the scale from 1 = "definitely not associated" to 9 = "definitely associated", how closely are the words below associated with the category for the robot?

Dangerous	1-2-3-4-5-6-7-8-9
Compassionate	1-2-3-4-5-6-7-8-9
Capable	1-2-3-4-5-6-7-8-9
Scary	1-2-3-4-5-6-7-8-9
Social	1-2-3-4-5-6-7-8-9
Aggressive	1-2-3-4-5-6-7-8-9
Knowledgeable	1-2-3-4-5-6-7-8-9
Awful	1-2-3-4-5-6-7-8-9
Reliable	1-2-3-4-5-6-7-8-9
Happy	1-2-3-4-5-6-7-8-9
Responsive	1-2-3-4-5-6-7-8-9
Organic	1-2-3-4-5-6-7-8-9
Feeling	1-2-3-4-5-6-7-8-9
Awkward	1-2-3-4-5-6-7-8-9
Interactive	1-2-3-4-5-6-7-8-9
Competent	1-2-3-4-5-6-7-8-9
Emotional	1-2-3-4-5-6-7-8-9
Machine-like	1-2-3-4-5-6-7-8-9
Strange	1-2-3-4-5-6-7-8-9

B. Please rate the following questions, according to the scale from 1 = "strongly disagree" to 9 = "strongly agree".

1. The robot was able to take my orders correctly.	1-2-3-4-5-6-7-8-9
2. The robot was able to change my orders correctly.	1-2-3-4-5-6-7-8-9
3. The robot was able to respond promptly to my requests.	1-2-3-4-5-6-7-8-9
4. The utterances said by the robot during the interaction(s) were clear.	1-2-3-4-5-6-7-8-9
5. The robot responded correctly to my requests.	1-2-3-4-5-6-7-8-9
6. The robot made many mistakes during the interaction(s).	1-2-3-4-5-6-7-8-9
7. The robot was able to recover from its mistakes.	1-2-3-4-5-6-7-8-9
8. The robot delivered the right order(s).	1-2-3-4-5-6-7-8-9
9. The interaction(s) could not be completed.	1-2-3-4-5-6-7-8-9
10. The interaction(s) ended abruptly.	1-2-3-4-5-6-7-8-9
11. I would like to have another interaction with the robot.	1-2-3-4-5-6-7-8-9
12. The interaction(s) flowed smoothly.	1-2-3-4-5-6-7-8-9
13. I was very dissatisfied with the conversation(s).	1-2-3-4-5-6-7-8-9
14. The interaction(s) was too long.	1-2-3-4-5-6-7-8-9
15. I enjoyed interacting with the robot.	1-2-3-4-5-6-7-8-9
16. The robot repeated itself many times.	1-2-3-4-5-6-7-8-9
17. I felt uncomfortable while interacting with the robot.	1-2-3-4-5-6-7-8-9
18. It was easy for me to communicate with the robot.	1-2-3-4-5-6-7-8-9
19. After the interaction, I was looking forward to meeting the robot the next day.	1-2-3-4-5-6-7-8-9
20. The vocabulary and the speech of the robot became more fluid and rich.	1-2-3-4-5-6-7-8-9
21. I feel attached to the robot.	1-2-3-4-5-6-7-8-9
22. I felt that the robot was autonomous.	1-2-3-4-5-6-7-8-9
24. I would be willing to interact with the robot if it was a barista in a coffee shop.	1-2-3-4-5-6-7-8-9
25. I felt that the robot personalised its interaction.	1-2-3-4-5-6-7-8-9
26. The robot recognised me correctly.	1-2-3-4-5-6-7-8-9
27. I feel that the robot knows my preferences well.	1-2-3-4-5-6-7-8-9
28. The robot remembered my previous order(s) correctly.	1-2-3-4-5-6-7-8-9
29. I was pleased that the robot was able to remember my previous orders.	1-2-3-4-5-6-7-8-9

C. Additional information:

1. How old are you?

2. What is your gender? (If you do not wish disclose it, you can leave this question empty.)

3. What is your approximate height in centimeters?

4. For a scale, 1 = "not at all" to 9 = "very familiar", how familiar were you with robots before the experiment?	1-2-3-4-5-6-7-8-9
5. For a scale, 1 = "none" to 9 = "substantial", what is your level of previous experience with robots?	1-2-3-4-5-6-7-8-9
6. For a scale, 1 = "not at all" to 9 = "every day", how frequently do you go to a coffee shop?	1-2-3-4-5-6-7-8-9
7. For a scale, 1 = "very dissatisfactory" to 9 = "very satisfactory", how would you rate your overall experience?	1-2-3-4-5-6-7-8-9

8. Any additional comments or suggestions about the robot or interaction?

Appendix D

Additional Information on the Barista Datasets

This section presents the additional information for the Barista and Personalised Barista Datasets, presented in Chapter 6.

D.1 Barista Dataset

This section contains the additional tables for the proportion of personal(ised), order details and other (remaining) phrase types in the *training*, *validation* and *out-of-vocabulary (OOV)* sets of the Barista Dataset.

Table D.1: The percentage of personal(ised), order details and other (remaining) phrase types in the tasks of 1,000 and 10,000 dialogue Barista *training* set.

Dataset Size	Phrases	B1	B2	B3	B4	B5	B6	B7
	Personal	25	0	0	0	0	14.29	12.88
1,000	Order	0	25	37.07	25	36.63	14.29	22.72
	Other	75	75	62.93	75	63.37	71.43	64.4
	Personal	25	0	0	0	0	14.29	12.9
10,000	Order	0	25	36.71	25	36.68	14.29	22.57
	Other	75	75	63.29	75	63.32	71.43	64.52

Table D.2: The percentage of personal(ised), order details and other (remaining) phrase types in the tasks of 1,000 and 10,000 dialogue Barista *validation* set.

Dataset Size	Phrases	B1	B2	B3	B4	B5	B6	B7
	Personal	25	0	0	0	0	14.29	12.92
1,000	Order	0	25	36.87	25	36.49	14.29	22.48
	Other	75	75	63.13	75	63.51	71.43	64.6
	Personal	25	0	0	0	0	14.29	12.89
10,000	Order	0	25	36.83	25	36.82	14.29	22.66
	Other	75	75	63.17	75	63.18	71.43	64.45

Table D.3: The percentage of personal(ised), order details and other (remaining) phrase types in the tasks of 1,000 and 10,000 dialogue Barista *out-of-vocabulary (OOV)* set.

Dataset Size	Phrases	B1	B2	B3	B4	B5	B6	B7
	Personal	25	0	0	0	0	14.29	12.91
1,000	Order	0	25	37.13	25	36.41	14.29	22.54
	Other	75	75	62.87	75	63.59	71.43	64.55
	Personal	25	0	0	0	0	14.29	12.9
10,000	Order	0	25	36.9	25	36.89	14.29	22.62
	Other	75	75	63.1	75	63.11	71.43	64.48

D.2 Personalised Barista Dataset

This section contains the additional tables for the proportion of personal(ised), order details and other (remaining) phrase types in the *training*, *validation* and *OOV* sets of the Personalised Barista Dataset.

Table D.4: The percentage of personal(ised), order details, other (remaining) and Barista Task 7 (B7) phrase types in the tasks of Second-Interaction, 1,000 and 10,000 dialogue Personalised Barista *training* set.

Dataset Size	Phrases	B0	B1	B2	B3	B4	B5	B6	B7	B8
200 - 400	Personal	27.91	36.31	38.91	33.16	28.94	35.78	30.6	29.36	30.99
	Order	25.58	27.38	28.16	29.56	31.6	29.36	31.43	31.18	30.99
	Other	55.81	50.84	48.71	48.9	51.04	47.94	50.36	49.87	49.38
	B7	90.7	85.48	79.32	84.26	88.43	79.82	84.53	86.26	83.35
1,000	Personal	54.57	54.6	56.25	45.58	37.16	46.89	39.17	36.52	37.92
	Order	31.07	31.03	31.19	32.51	35.77	32.61	35.83	35.27	35.71
	Other	40.21	40.23	39.21	40.8	44.67	40.1	43.51	43.38	42.31
	B7	74.15	74.14	67.28	75.4	82.4	70.9	77.82	80.33	76.11
10,000	Personal	65.28	65.29	65.31	51.58	40.75	52.33	42.14	39.76	41.12
	Order	33.08	33.07	33.05	34.21	38.19	34.19	37.94	37.1	36.88
	Other	34.12	34.12	34.11	36.62	41.34	36.44	40.87	40.34	40.01
	B7	67.52	67.52	61.71	71.05	79.73	66.79	75.21	77.61	73.83

Table D.5: The percentage of personal(ised), order details, other (remaining) and Barista Task 7 (B7) phrase types in the tasks of Second-Interaction, 1,000 and 10,000 dialogue Personalised Barista *validation* set.

Dataset Size	Phrases	B0	B1	B2	B3	B4	B5	B6	B7	B8
200 - 400	Personal	27.78	38.95	41.88	33.66	29.35	37.29	30.95	29.87	32.2
	Order	25.93	27.13	28.3	29.3	32.29	29.36	32.27	32.12	31.9
	Other	55.56	50	47.22	48.84	50.38	47.27	49.75	48.91	48.02
	B7	90.74	83.92	77.76	83.78	87.98	78.74	84.17	85.77	81.44
1,000	Personal	54.69	54.82	55.66	45.47	36.62	47.22	38.58	36.79	38.44
	Order	30.92	30.76	31.38	32.73	35.9	32.67	35.85	35.38	35.15
	Other	40.3	40.39	39.27	40.65	44.83	40.06	43.88	43.21	42.63
	B7	74.09	74.03	68.68	75.48	82.66	70.68	77.77	80.31	76.08
10,000	Personal	65.33	65.29	65.41	51.4	40.82	52.27	42	39.74	41.05
	Order	33.03	33.07	33.09	34.19	38.17	34.19	37.99	37.1	36.8
	Other	34.14	34.12	34.04	36.68	41.31	36.45	40.9	40.39	40.06
	B7	67.5	67.52	61.46	71.09	79.69	66.58	75.4	77.66	73.85

Table D.6: The percentage of personal(ised), order details, other (remaining) and Barista Task 7 (B7) phrase types in the tasks of Second-Interaction, 1,000 and 10,000 dialogue Personalised Barista *out-of-vocabulary (OOV)* set.

Dataset Size	Phrases	B0	B1	B2	B3	B4	B5	B6	B7	B8
200 - 400	Personal	28.2	28.14	30.29	27.27	24.12	29.63	27.27	24.53	26.86
	Order	24.81	24.95	25.77	26.45	27.33	27.37	29.23	28.98	28.96
	Other	56.39	56.29	54.25	54.55	56.59	52.26	52.95	54.04	52.45
	B7	90.6	90.62	86.53	89.26	91.96	84.37	85.95	90.57	85.9
1,000	Personal	54.87	45.75	47.75	40.28	32.54	41.99	35.67	33.55	35.21
	Order	30.7	28.83	29.67	30.45	33.91	31.04	33.93	33.03	33.62
	Other	40.43	45.75	43.73	45.05	48.01	43.87	46.3	46.45	45.23
	B7	74.01	79.66	74.16	79.45	85.54	75.68	79.81	82.74	78.9
10,000	Personal	65.31	64.01	64.3	50.51	40.29	51.78	41.89	39.34	40.9
	Order	33.05	32.76	32.84	34.09	37.89	34.03	37.53	36.97	36.53
	Other	34.13	34.91	34.7	37.13	41.76	36.82	41.31	40.58	40.37
	B7	67.51	68.32	62.21	71.67	80.06	66.98	75.18	77.9	73.95

Appendix E

Hyperparameters and Additional Results on Data-Driven Approaches

This section presents the additional findings on Barista and Personalised Barista Datasets, presented in Chapter 8. In addition, we present the hyperparameters for the baselines used in the evaluations, and dialogue examples with attention weights.

E.1 Hyperparameters

This section presents the hyperparameters of the models used in the experiments for the Barista Datasets. These hyperparameters correspond to the parameters from the original implementations (Joshi et al. 2017; Zhang et al. 2018), unless otherwise noted in Chapter 8. For the graphics processing unit (GPU), we used GeForce GTX 1080 Ti or Tesla V100-SXM3-32GB, depending on the availability.

Table E.1: Hyperparameters of the models used in the experiments for the Barista Datasets. These correspond to the parameters from the original implementations (Joshi et al. 2017; Zhang et al. 2018), unless otherwise noted in Chapter 8.

Hyper-parameter	**MemN2N**	**Split Memory**	**Key-Value**	**Profile**	**Seq2Seq**	**Supervised**
Learning Rate	0.001	0.001	0.1	0.001	3	0.01
Embedding Size	20	20	1000	300	256	32
Negative Candidates	100	100	10	All	All	100
Optimiser	Adam	Adam	SGD	Adam	SGD	Adam
User Profile	Dialogue turn	Separate	Separate	Separate	Prepended	Dialogue turn
Hops	1-3	1-3	1	1	-	-
Batch Size	32	32	1	128/ 1 (for test)	64/ 1 (for test)	128
Training Epochs	100	100	25	100	100	25/ 15 (for 10,000 dialogues)
Resource	1 GPU+ 1 CPU	1 GPU+ 1 CPU	16-18 CPUs	1 GPU+ 1 CPU	1 GPU+ 1 CPU	1 GPU+ 1 CPU
Vocabulary	all sets	all sets	*training, validation, test*	*training, validation, test*	*training, validation, test*	all sets
Context	Previous user-bot labels (may be cut off)	Previous user-bot labels (may be cut off)	Last bot-user labels	Previous user-bot labels (training)/ Previous user-predicted bot responses (validation and test)	Previous user-bot labels	Previous user-bot labels

E.2 Attention Weights for End-to-End Memory and Split Memory Networks

Table E.2: A dialogue example from the *recognition error* task (2) of the Personalised Barista Dataset (PB2) shows the attention weights in the End-to-End Memory Networks (MemN2N) model for varying hops. The *training* set includes dialogue examples of Anne Hathaway, which enable the model to correctly predict the response without available preferences information. Hops direct the focus to the most relevant dialogue history input, which contains the name of the customer. Zero attention weight signifies a very small value ($< 10^{-5}$).

Time	**Speaker**	**Conversation Context**	**Hop1**	**Hop2**	**Hop3**
1		True , 182 , Tom Welling	0.221	0.0004	0
2	Customer	Hey there !	0.032	0	0
3	Barista	Hello , Tom Welling ! Would you like to have a large latte again ?	0.027	0	0
4	Customer	That is not my name .	0.029	0	0
5	Barista	I am terribly sorry ! May I take your name ? api_call getCustomerName	0.009	0	0
6	Customer	It is Anne Hathaway .	0.105	0.0007	0
7	Barista	Let me see if I have any previous records of you , Anne .	0.24	0.913	1
8		True , 23 , Anne Hathaway	0.337	0.086	0
Customer Input		Okay .			
Correct Response		I thought you looked familiar , **Anne** ! Would you like a **small mocha** and a **blueberry muffin** again ?			
Predicted Response		I thought you looked familiar , **Anne** ! Would you like a **small mocha** and a **blueberry muffin** again ?			

Table E.3: A dialogue example from the *recognition error* task (2) of the Personalised Barista Dataset (PB2) shows the attention weights in the Split Memory model for varying hops. The profile memory in the Split Memory architecture focuses mostly on the customer's ID (23) instead of the customer's name, which might be the reason behind mixing the customer and their preference in contrast to the End-to-End Memory Networks (Table E.2), especially in hops 2 and 3. Note that both customers appear in the *training* set. Zero attention weight signifies a very small value ($< 10^{-5}$).

		Profile	**Hop1**	**Hop2**	**Hop3**
		True	0.323	0.079	0.002
		23	0.551	0.892	0.998
		Anne Hathaway	0.122	0.03	0.0002
Time	**Speaker**	**Conversation Context**			
1	Customer	Hey there !	0.098	0.00009	0
2	Barista	Hello , Tom Welling ! Would you like to have a large latte again ?	0.045	0.0005	0
3	Customer	That is not my name .	0.218	0.006	0
4	Barista	I am terribly sorry ! May I take your name ? api_call getCustomerName	0.059	0.79	1
5	Customer	It is Anne Hathaway .	0.126	0.002	0
6	Barista	Let me see if I have any previous records of you , Anne .	0.454	0.201	0
Customer Input		Okay .			
Correct Response		I thought you looked familiar , **Anne** ! Would you like a **small mocha** and a **blueberry muffin** again ?			
Predicted Response		I thought you looked familiar , **Roseanne** ! Would you like a **large cappuccino** and a **slice of banana bread** again ?			

Table E.4: A dialogue example with a change of order in the *changes to preference* task (4) of the Personalised Barista Dataset (PB4) shows the attention weights in the End-to-End Memory Networks (MemN2N) model for varying hops. In contrast to Table E.5, the model focuses only on the last bot response, which contains the previous order that needs to be changed, in order to update correctly the remaining parts of the order. Zero attention weight signifies a very small value ($< 10^{-5}$).

Time	**Speaker**	**Conversation Context**	**Hop1**	**Hop2**	**Hop3**
1		True , 90 , David Boreanaz	0	0	0
2	Customer	Hello !	0	0	0
3	Barista	Hello , David Boreanaz ! Would you like to have a large peppermint hot chocolate again ?	0	0	0
4	Customer	No . Can I get an espresso macchiato ?	0.00007	0	0
5	Barista	Which size ?	0	0	0
6	Customer	Medium would do .	0.00001	0	0
7	Barista	Would you like anything else with that ?	0	0	0
8	Customer	A breakfast tortilla wrap .	0.00001	0	0
9	Barista	So that would be a medium espresso macchiato and a breakfast tortilla wrap . Is that right ?	1	1	1
Customer Input		Can I get small instead ?			
Correct Response		Sure ! So that would be a **small espresso macchiato** and a **breakfast tortilla wrap** . Is that right ?			
Predicted Response		Sure ! So that would be a **small espresso macchiato** and a **breakfast tortilla wrap** . Is that right ?			

Table E.5: A dialogue example from the *changes to preference* task (4) of the Personalised Barista Dataset (PB4) shows the attention weights in the End-to-End Memory Networks (MemN2N) model for varying hops. The model should ideally focus on both the new drink order and the requested size. In all hops, the model primarily and correctly focuses on the new drink item. However, the attention weight on the size is not high enough, which caused an incorrect response in hop 1 (*large* instead of *small*), while it was unexpectedly correct in the other hops. Zero attention weight signifies a very small value ($< 10^{-5}$).

Time	**Speaker**	**Conversation Context**	**Hop1**	**Hop2**	**Hop3**
1		True , 58 , Shane West	0	0	0
2	Customer	Hey !	0.00002	0	0
3	Barista	Hello , Shane West ! Would you like to have a medium Earl Grey and a classic oatmeal again ?	0.001	0	0
4	Customer	Nope .	0.00007	0	0
5	Barista	Alright ! Which drink can I get for you , Shane ?	0.0008	0	0
6	Customer	Can I get a green tea ?	0.998	1	1
7	Barista	Which size ?	0.00001	0	0
8	Customer	Small would do .	0.00003	0	0
9	Barista	Would you like anything else with that ?	0	0	0
Customer Input		Could I get a pain aux raisins ?			
Correct Response		So that would be a **small green tea** and a **pain aux raisins** . Is that right ?			
Predicted Response		So that would be a **small green tea** and a **pain aux raisins** . Is that right ?			

E.3 Out-of-Vocabulary Sets

E.3.1 1,000 Dialogues

Table E.6: The *out-of-vocabulary (OOV)* set results of the Personalised Barista Dataset with 1,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average and for task 8 (containing all tasks), Supervised Embeddings is the best performing model.

Task	**MemN2N**			**Split Memory**			**Key-**	**Profile**	**Seq2-**	**Super-**
	Hop1	Hop2	Hop3	Hop1	Hop2	Hop3	**Value**		**Seq**	**vised**
0	41.73	42.88	41	43.52	43.63	41.96	20.88	40.69	39.94	**55.16**
1	47.53	46.52	43.01	46.82	47.33	44.89	20.84	47.23	46.01	**53.14**
2	43.59	45.02	43.83	44.11	44.93	45.02	20.29	47.8	42.73	**52.76**
3	43.08	42.01	38.38	43.63	44.38	41.78	19.8	45.64	43.83	**55.73**
4	44.29	43.46	42.52	42.95	44.54	41.4	16.45	47.51	49.24	**57.72**
5	42.64	42.84	40.99	39.61	44.37	40.07	16.52	43.76	43.53	**53.26**
6	40.73	44.32	45.82	43.79	42.75	39.71	32.46	46.97	42.82	**54.6**
7	41.21	43.91	41.99	39.38	40.03	41.04	14.53	47.88	46.74	**55.72**
8	42.47	40.56	39.31	42.79	39.06	43.88	13.87	39.91	46.26	**53.82**

Table E.7: The *out-of-vocabulary (OOV)* set results of the Personalised Barista with Preferences Information Dataset with 1,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average and for task 8 (containing all tasks), Supervised Embeddings is the best performing model, similar to the Personalised Barista Dataset.

Task	**MemN2N**			**Split Memory**			**Key-**	**Profile**	**Seq2-**	**Super-**
	Hop1	Hop2	Hop3	Hop1	Hop2	Hop3	**Value**		**Seq**	**vised**
0	41.41	41.52	40.11	40.83	39.56	40.77	17.35	40.54	39.39	**53.71**
1	46.57	46.77	44.18	46.52	43.21	42.65	17.79	46.31	43.42	**51.11**
2	44.35	44.21	40.96	40.96	43.64	41.72	16.27	45.84	43.64	**56.59**
3	40.87	45.88	44.02	40.43	39.68	38.3	14.99	44.42	45.25	**54.9**
4	45.73	43.89	42.77	41.94	46.89	42.34	10.99	47.29	45.55	**57.93**
5	40.99	42.37	38.61	37.99	38.65	37.92	11.33	44.41	43.3	**54.12**
6	44.74	45.92	42.02	44.32	44.21	45.4	11.3	46.55	42.96	**55.22**
7	41.66	39.8	40.98	41.92	38.4	40.91	11.76	47.46	44.07	**55.64**
8	41.63	40.13	40.66	40.38	37.09	43.19	9.92	43.38	40.41	**53.19**

E.3.2 Second Interaction

Table E.8: The *out-of-vocabulary (OOV)* set results of the Personalised Barista Dataset for Second Interaction set (*few-shot learning*). The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average and for task 8 (containing all tasks), Supervised Embeddings is the best performing model, similar to the *out-of-vocabulary (OOV)* set of the Personalised Barista Dataset with 1,000 dialogues.

Task	**Memory Networks**			**Split Memory**			**Key-**	**Profile**	**Seq2-**	**Super-**
	Hop1	Hop2	Hop3	Hop1	Hop2	Hop3	**Value**		**Seq**	**vised**
0	**56.39**	56.02	54.42	54.51	51.69	55.17	18.98	56.02	55.45	41.35
1	56.38	53.38	55.25	53.75	**57.5**	51.31	23.83	50.47	56	50.66
2	52.62	54.52	46.65	48.73	54.16	53.44	20.98	**55.79**	55.33	52.87
3	54.38	50.58	51.24	53.31	50.74	54.05	21.49	55.12	55.54	**61.59**
4	53.54	56.03	55.71	55.14	52.25	55.47	16.4	56.51	53.78	**56.8**
5	46.58	48.21	46.42	48.68	47.2	46.5	15.47	49.92	54.43	**55.01**
6	47.66	50.91	51.81	48.11	47.96	51.13	18.5	53.25	53.47	**57.45**
7	51.47	51.17	52	53.43	53.43	49.43	16.3	53.66	50.42	**61.42**
8	45.51	50.56	48.18	50.42	50.63	50.28	15.64	44.04	52.24	**58.27**

Table E.9: The *out-of-vocabulary (OOV)* set results of the Personalised Barista with Preferences Information Dataset for Second Interaction set (*few-shot learning*). The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average, Sequence-to-Sequence is the best performing model, whereas Supervised Embeddings model performs best in task 8 (containing all tasks).

Task	**Memory Networks**			**Split Memory**			**Key-**	**Profile**	**Seq2-**	**Super-**
	Hop1	Hop2	Hop3	Hop1	Hop2	Hop3	**Value**		**Seq**	**vised**
0	53.38	55.73	53.95	**56.67**	55.36	55.55	26.5	56.3	55.64	25.95
1	**56.85**	**56.94**	54.03	53.66	56.29	55.25	22.61	51.22	54.97	42.97
2	51.9	54.79	50.81	49.01	53.35	53.35	23.24	**54.52**	**54.52**	47.98
3	53.39	51.98	50.66	48.02	54.38	47.52	21.65	50.58	53.06	**59.74**
4	53.46	54.82	53.7	53.22	**57.4**	50.4	22.03	54.66	56.59	57.25
5	49.22	49.38	47.9	44.09	49.61	43.23	16.41	47.59	**52.72**	51.19
6	51.06	50	48.19	46.9	46.07	50.68	15.03	51.89	**53.32**	44.93
7	51.17	53.58	50.64	50.79	48.91	50.26	19.09	52.53	55.47	**59.2**
8	48.53	45.72	48.95	49.72	46.21	48.25	18.02	52.17	52.66	**53.31**

E.3.3 10,000 Dialogues

Table E.10: The *out-of-vocabulary (OOV)* set results of the Barista Dataset with 10,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average and for task 8 (containing all tasks), Sequence-to-Sequence (Seq2Seq) is the best performing model.

Task	**Memory Networks**			**Key-Value**	**Seq2Seq**	**Supervised**
	Hop1	Hop2	Hop3			
1	**80.02**	79.73	76.14	75.26	78.53	78.46
2	72.39	50.5	67.58	18.65	**75**	64.73
3	52.14	49.65	53.7	7.97	61.99	**62.56**
4	49.64	71.11	66.79	10.99	**75**	73.81
5	61.13	53.01	41.9	5.94	**63.11**	46.68
6	57.64	65.69	58.29	28.21	**70.13**	65.65
7	56.62	49.45	59.2	31.25	**65.3**	65.04

Table E.11: The *out-of-vocabulary (OOV)* set results of the Personalised Barista Dataset with 10,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that in all tasks, Supervised Embeddings is the best performing model.

Task	**Memory Networks**			**Split Memory**			**Key-Value**	**Profile**	**Seq2-Seq**	**Super-vised**
	Hop1	Hop2	Hop3	Hop1	Hop2	Hop3				
0	36.42	36.2	36.4	35.13	38.59	39.08	30.61	35	35.38	**50.85**
1	38.13	37.25	37.84	41.58	41.34	38.74	29.41	35.42	35.99	**51.21**
2	40.73	40.9	39.71	33.48	39.04	40.08	28.3	38.28	36.38	**49.2**
3	31.27	32.58	30.72	38.9	29.66	31.88	18.6	39.02	37.76	**50.07**
4	32.4	35.1	37.27	41.63	32.19	32.52	15.47	43.33	42.76	**51.19**
5	33.54	33.14	32.08	32.2	34.91	34.34	19.58	39.91	38.79	**46.4**
6	35.55	40.63	32.71	29.48	28.28	39.27	18.48	18.75	43.13	**49.39**
7	35.65	26.58	30.47	29.54	35.97	37.46	15.5	39.35	38.27	**45.89**
8	32.3	32.56	35.97	34.36	31.81	34.0	15.48	37.28	41.64	**47.96**

Table E.12: The *out-of-vocabulary (OOV)* set results of the Personalised Barista with Preferences Information Dataset with 10,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that in all tasks, Supervised Embeddings is the best performing model.

Task	**Memory Networks**			**Split Memory**			**Key-Value**	**Profile**	**Seq2-Seq**	**Super-vised**
	Hop1	Hop2	Hop3	Hop1	Hop2	Hop3				
0	36.4	35.39	34.6	36.75	36.74	36.23	15.1	34.12	34.9	**50.45**
1	36.58	35.24	35.37	38.7	39.19	39.32	13.23	33.79	34.29	**51.28**
2	39.28	39.48	39.42	38.98	40.31	38.82	13.58	36.35	34.5	**50.24**
3	31.61	31.23	32.13	29.86	35.01	34.69	6.74	36.81	35.61	**50.4**
4	37.05	41.23	38.47	30.48	34.32	35.2	4.86	42.43	41.45	**51.29**
5	31.89	33.86	36.2	32.06	31.49	35.25	8.15	38.52	38.78	**48.28**
6	35.4	35.88	37.97	32.17	33.81	33.97	6.05	41.6	42.16	**51.54**
7	30.13	33.07	28.87	28.8	33.11	31.06	6.07	37.59	40.46	**41.31**
8	31.18	31.81	33.79	27.41	30.1	32.92	6.85	40.24	40.5	**48.23**

E.4 Test Set Results for the Second Interaction and 10,000 Dialogue Datasets

E.4.1 Second Interaction

Table E.13: The *test* set results of the Personalised Barista with Preferences Information Dataset with Second Interaction set (*few-shot learning*). The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average and for task 8 (containing all tasks), Sequence-to-Sequence (Seq2Seq) is the best performing model.

Task	**MemN2N**			**Split Memory**			**Key-Value**	**Profile**	**Seq2-Seq**	**Super-vised**
	Hop1	Hop2	Hop3	Hop1	Hop2	Hop3				
0	**59.87**	**59.96**	59.77	59.68	59.3	58.65	54.61	56.67	56.86	33.87
1	73.65	73.34	74.19	70.03	**74.5**	73.96	54.7	70.34	60.94	56.95
2	69.99	68.54	68.9	64.86	68.11	67.32	50.14	**72.01**	64.5	56.55
3	65.25	65.7	63.92	61.58	62.59	62.22	45.44	44.94	**71.33**	60.59
4	67.06	68.8	64.78	65.25	65.66	63.32	42.27	67.52	**78.66**	57.82
5	60.39	59.65	59.3	56.22	56.22	56.79	38.36	45.49	**74.43**	50.12
6	63.28	60.07	59.68	58.68	60.62	59.01	32.45	62.34	**76.82**	51.47
7	**64.14**	61.94	61.34	61.13	58.76	59.78	40.16	60.97	60.86	59.99
8	58.52	59.46	55.95	53.52	55.8	55.45	35.73	59.32	**74.43**	50.91

Table E.14: Percentage of errors in dialogue state tracking (DST), *personal(ised)*, *order details*, other and Barista Task 7 (B7) phrase types for Second Interaction *test* sets. The best performing methods (or methods within 0.1%) are given in bold for the error in per-response accuracy metric, and the error percentages within the phrase types are given in parentheses.

Task	**Error Type**	**MemN2N** Hop1	Hop2	Hop3	**Split Memory** Hop1	Hop2	Hop3	**Key-Value**	**Profile**	**Seq2-Seq**	**Super-vised**
PB0	DST	7.8	5.45	2.91	11.09	4.32	3.29	19.27	1.5	**0.38**	16.64
	Personal	27.07 (95.98)	27.54 (97.65)	27.07 (95.98)	26.41 (93.65)	26.41 (93.65)	27.07 (95.98)	**8.36** (29.66)	28.20 (99.98)	28.20 (99.98)	28.01 (99.32)
	Order	**22.46** (90.54)	**22.37** (90.16)	22.56 (90.92)	23.31 (93.95)	22.84 (92.05)	23.97 (96.60)	23.31 (93.95)	24.81 (100.00)	24.44 (98.49)	24.06 (96.98)
	Other	0.56 (1.00)	0.94 (1.67)	0.94 (1.67)	0.66 (1.17)	0.66 (1.17)	0.38 (0.67)	21.33 (37.83)	0.38 (0.67)	**0.00** (0.00)	29.14 (51.67)
	B7	**13.63** (15.04)	13.91 (15.35)	14.10 (15.56)	14.57 (16.08)	14.10 (15.56)	14.94 (16.49)	36.47 (40.25)	15.79 (17.43)	15.04 (16.60)	43.80 (48.34)
PB1	DST	6.24	2.47	3.16	6.01	3.24	3.39	28.51	**0.15**	0.39	5.39
	Personal	16.49 (41.63)	16.02 (40.47)	15.95 (40.27)	17.03 (43.00)	**14.79** (37.35)	15.64 (39.49)	15.87 (40.08)	16.26 (41.05)	29.04 (73.35)	30.35 (76.65)
	Order	**13.94** (51.14)	14.41 (52.83)	15.02 (55.09)	15.02 (55.09)	14.56 (53.40)	14.79 (54.24)	25.73 (94.36)	15.56 (57.07)	21.65 (79.39)	23.96 (87.86)
	Other	0.31 (0.62)	0.23 (0.47)	0.62 (1.24)	0.31 (0.62)	0.39 (0.78)	0.69 (1.40)	19.41 (39.13)	2.16 (4.35)	**0.00** (0.00)	9.94 (20.03)
	B7	9.32 (11.16)	9.71 (11.62)	10.71 (12.82)	9.71 (11.62)	9.94 (11.90)	10.55 (12.64	29.51 (35.33)	12.48 (14.95)	**8.40** (10.06)	19.57 (23.43)
PB8	DST	10.9	13.73	11.2	10.36	9.17	10.01	49.65	5.2	**1.73**	9.07
	Personal	25.82 (77.88)	27.06 (81.62)	26.36 (79.53)	24.78 (74.74)	25.02 (75.49)	24.98 (75.34)	**19.52** (58.90)	33.00 (99.56)	21.56 (65.03)	29.78 (89.84)
	Order	**26.26** (81.79)	26.91 (83.80)	25.62 (79.79)	28.64 (89.20)	28.74 (89.51)	27.45 (85.50)	30.97 (96.45)	32.06 (99.85)	27.21 (84.72)	30.43 (94.76)
	Other	2.58 (5.47)	4.31 (9.15)	2.53 (5.36)	3.22 (6.83)	2.18 (4.63)	2.33 (4.94)	24.98 (52.99)	1.78 (3.79)	**0.15** (0.32)	9.86 (20.92)
	B7	17.29 (21.65)	19.72 (24.69)	17.34 (21.71)	19.82 (24.81)	19.18 (24.01)	18.53 (23.20)	42.47 (53.16)	20.91 (26.18)	**14.92** (18.67)	27.55 (34.49)
PBPI8	DST	12.09	12.49	16.85	10.6	9.51	11.11	59.27	**0.79**	1.68	13.28
	Personal	23.44 (70.71)	23.49 (70.86)	25.97 (78.33)	26.36 (79.53)	26.16 (78.93)	25.22 (76.09)	**18.53** (55.91)	21.01 (63.38)	20.91 (63.08)	29.19 (88.05)
	Order	21.51 (66.98)	22.10 (68.83)	23.04 (71.76)	27.80 (86.58)	25.12 (78.24)	26.51 (82.56)	20.71 (64.51)	28.84 (89.82)	**12.49** (38.89)	29.58 (92.13)
	Other	3.77 (7.99)	3.07 (6.52)	3.72 (7.89)	2.82 (5.99)	2.97 (6.31)	2.38 (5.05)	26.16 (55.52)	**0.10** (0.21)	0.30 (0.63)	10.70 (22.71)
	B7	16.70 (20.91)	15.86 (19.85)	17.00 (21.28)	18.88 (23.64)	16.95 (21.22)	18.09 (22.64)	43.90 (54.96)	19.57 (24.50)	**4.41** (5.52)	28.74 (35.98)

E.4.2 10,000 Dialogues

Table E.15: The *test* set results of the Personalised Barista with Preferences Information Dataset with 10,000 dialogues. The best performing methods (or methods within 0.1% of best performing) are given in bold for the per-response accuracy metric. The results show that on average and for task 8 (containing all tasks), Split Memory is the best performing model.

Task	**MemN2N**			**Split Memory**			**Key-Value**	**Profile**	**Seq2-Seq**	**Super-vised**
	Hop1	Hop2	Hop3	Hop1	Hop2	Hop3				
0	34.39	34.56	35.59	37.94	36.69	38.63	**66.15**	33.78	34.37	50.43
1	67.66	67.46	67.73	69.37	67.84	69.99	69.26	62.24	63.82	**70.33**
2	68.06	69.24	68.76	68.31	69.69	68.89	**74.23**	64.44	63.13	67.54
3	75.15	75.05	75.39	75.75	75.38	**76.01**	54.47	58.88	72.05	66.3
4	79.58	79.75	79.21	77.3	**80.49**	79.65	61.12	62.28	76.51	66.66
5	75.45	**75.94**	75.61	75.38	74.91	75.18	55.24	57.51	70.72	63.12
6	**79.41**	**79.38**	79.25	74.53	72.03	75.84	56.67	61.29	74.9	61.58
7	80.73	79.56	80.34	72.27	80.79	**81.1**	49.79	55.63	71.8	66.39
8	**81.07**	79.39	80.85	72.82	63.38	**81.02**	46.64	58.45	75.93	62.92

Table E.16: Percentage of errors in DST, *personal(ised)*, *order details*, other and Barista Task 7 (B7) phrase types for 10,000 dialogue *test* sets. The best performing methods (or methods within 0.1%) are given in bold for the error in per-response accuracy metric, and the error percentages within the phrase types are given in parentheses.

Task	**Error Type**	**MemN2N** Hop1	Hop2	Hop3	**Split Memory** Hop1	Hop2	Hop3	**Key-Value**	**Profile**	**Seq2-Seq**	**Super-vised**
B7	DST	**0.02**	**0.01**	**0.02**	-	-	-	**0**	-	**0**	0.33
	Personal	**0.01**	**0.01**	**0.01**	-	-	-	**0.00**	-	**0.00**	1.06
		(0.09)	(0.07)	(0.09)				(0.00)		(0.00)	(8.20)
	Order	0.73	0.60	0.86	-	-	-	22.24	-	**0.02**	10.80
		(3.21)	(2.68)	(3.83)				(98.54)		(0.08)	(47.86)
	Other	**0.01**	**0.00**	**0.00**	-	-	-	2.92	-	**0.00**	0.77
		(0.01)	(0.01)	(0.00)				(4.53)		(0.00)	(1.19)
PB0	DST	21.99	3.22	0.79	27.97	26.36	23.58	33.16	1.36	0.78	**0.28**
	Personal	64.39	63.72	62.05	60.41	58.71	**58.00**	64.09	63.58	64.05	63.92
		(98.62)	(97.59)	(95.04)	(92.53)	(89.92)	(88.84)	(98.16)	(97.38)	(98.10)	(97.90)
	Order	32.89	32.97	33.05	32.95	32.96	32.43	**32.09**	33.07	33.07	33.05
		(99.46)	(99.71)	(99.94)	(99.63)	(99.66)	(98.05)	(97.03)	(100.00)	(100.00)	(99.95)
	Other	**0.01**	**0.00**	**0.00**	**0.01**	**0.01**	**0.01**	1.02	1.30	0.47	0.74
		(0.02)	(0.01)	(0.00)	(0.02)	(0.02)	(0.02)	(2.98)	(3.80)	(1.37)	(2.17)
	B7	**0.42**	**0.50**	0.57	**0.48**	**0.49**	0.56	1.60	1.89	1.06	1.32
		(0.62)	(0.74)	(0.85)	(0.71)	(0.72)	(0.84)	(2.38)	(2.80)	(1.56)	(1.95)
PB1	DST	13.43	10.19	**0.13**	15.41	13.29	13.68	10.62	1.25	0.24	**0.18**
	Personal	31.21	31.96	31.26	29.69	29.22	29.79	**25.27**	34.85	34.27	41.44
		(47.80)	(48.96)	(47.89)	(45.48)	(44.76)	(45.63)	(38.71)	(53.38)	(52.50)	(63.49)
	Order	16.53	16.66	16.61	16.62	16.62	16.58	**16.19**	20.01	19.85	25.78
		(49.96)	(50.37)	(50.20)	(50.23)	(50.24)	(50.12)	(48.95)	(60.50)	(60.02)	(77.94)
	Other	**0.02**	**0.02**	**0.02**	**0.03**	**0.03**	**0.00**	1.26	1.01	0.38	0.34
		(0.05)	(0.05)	(0.05)	(0.09)	(0.08)	(0.00)	(3.70)	(2.96)	(1.11)	(1.01)
	B7	**0.50**	**0.60**	**0.55**	**0.57**	**0.57**	**0.57**	1.86	1.61	0.98	0.93
		(0.73)	(0.89)	(0.81)	(0.85)	(0.85)	(0.85)	(2.76)	(2.39)	(1.45)	(1.38)
PB8	DST	6.89	5.64	2.19	**0.59**	8.79	1.01	50.18	**0.59**	**0.62**	3.66
	Personal	20.98	21.21	**20.55**	21.81	21.55	21.56	27.09	25.71	25.96	32.10
		(51.00)	(51.56)	(49.97)	(53.02)	(52.40)	(52.43)	(65.87)	(62.50)	(63.13)	(78.04)
	Order	**12.87**	12.99	13.10	**12.90**	**12.95**	12.99	34.06	33.40	17.59	32.74
		(34.95)	(35.30)	(35.58)	(35.04)	(35.17)	(35.28)	(92.54)	(90.73)	(47.79)	(88.93)
	Other	**0.02**	**0.02**	**0.03**	**0.03**	**0.04**	**0.08**	14.20	0.31	**0.03**	4.69
		(0.04)	(0.06)	(0.09)	(0.09)	(0.09)	(0.19)	(35.45)	(0.78)	(0.08)	(11.70)
	B7	**0.88**	0.98	1.21	**0.87**	0.98	1.09	31.32	18.94	2.78	19.24
		(1.20)	(1.33)	(1.64)	(1.19)	(1.33)	(1.47)	(42.52)	(25.71)	(3.78)	(26.12)
PBPI8	DST	**0.31**	0.48	0.58	5.84	4.67	4.67	43.87	**0.33**	**0.3**	3.46
	Personal	**18.12**	19.76	18.95	21.60	26.43	26.43	25.62	22.76	21.71	26.16
		(44.05)	(48.04)	(46.06)	(52.52)	(64.26)	(64.26)	(62.30)	(55.34)	(52.79)	(63.61)
	Order	**9.81**	**9.85**	**9.89**	16.17	25.30	25.30	28.18	29.97	12.46	27.39
		(26.66)	(26.76)	(26.86)	(43.94)	(68.73)	(68.73)	(76.56)	(81.42)	(33.85)	(74.41)
	Other	**0.01**	**0.03**	**0.02**	1.21	1.24	1.24	8.92	0.16	0.21	3.25
		(0.02)	(0.09)	(0.06)	(3.03)	(3.08)	(3.08)	(22.27)	(0.40)	(0.53)	(8.11)
	B7	**0.81**	**0.83**	**0.90**	5.58	10.18	10.18	26.06	18.79	2.36	17.62
		(1.10)	(1.13)	(1.23)	(7.57)	(13.82)	(13.82)	(35.37)	(25.51)	(3.20)	(23.91)

Appendix F

Information Forms and Questionnaires for Socially Assistive Robotics Study

This appendix section presents the self consent and information form and the evaluation questionnaires used in the socially assistive robotics study in cardiac rehabilitation, presented in Chapter 9. The first document is the consent form used for all of the conditions. The next two documents contain the information and additional questions for *control* (no robot) and *social robot* conditions, respectively. All conditions have been evaluated with a modified version of Unified Theory of Acceptance and the Use of Technology (UTAUT) (Venkatesh et al. 2003; Casas et al. 2019) questionnaire, as shown in Table F.1. The *personalised robot* condition have been evaluated additionally with a set of personalisation-specific questions (presented in Table 9.2 in Section 9.3.2) and the Working Alliance Inventory (WAI) (Horvath & Greenberg 1989) (Table F.2).

Información para el paciente y formulario de consentimiento informado

Este es un formulario de consentimiento informado diseñado para invitarlo a participar en forma voluntaria en la investigación descrita a continuación:

Nombre de la investigación*: Evaluación del impacto de la intervención de un robot social en las respuestas cardiovasculares de los pacientes del programa de Rehabilitación Cardíaca de la Fundación Cardio-Infantil Instituto de Cardiología.*

Investigador Principal: Carlos Andrés Cifuentes García, Mónica Rincón R
Coinvestigador: Luisa Fernanda Gutiérrez, Lorena Pinzón, Marcela Cristina Múnera R., Wilson Alexander Sierra, Luis Eduardo Rodríguez

Entidades participantes: Escuela Colombiana de Ingeniería Julio Garavito, Fundación Cardio Infantil Instituto de Cardiología

Este documento de Consentimiento Informado tiene dos partes:

- Información sobre el estudio
- Formulario de Consentimiento para firmar si está de acuerdo en participar

Parte I: Información

Introducción

Estas hojas de consentimiento informado pueden contener palabras que usted no entienda. Por favor pregunte al investigador principal o a cualquier persona del estudio para que le explique cualquier palabra o información que usted no entienda claramente. Se le dará una copia del documento completo de consentimiento informado.

Propósito

Las enfermedades cardiovasculares son la principal causa de mortalidad en países en vía de desarrollo; específicamente, en Colombia en el 2010, fue reportado que hubo un total de muertes prematuras por enfermedades cardiovasculares equivalentes a 14.589 hombres y 9.910 mujeres por cada 100.000 habitantes. El mayor porcentaje de dichas muertes están dadas por enfermedad isquémica del corazón (57 %), enfermedad cardiovascular (20 %), otras enfermedades cardiovasculares (12 %), enfermedad hipertensiva (7 %), insuficiencia cardíaca (4 %) y enfermedad reumática (0,5 %). Dentro de este contexto, la rehabilitación cardiovascular (RC) es una herramienta fundamental para la mejoría de la calidad de vida de pacientes con alguna patología cardiovascular, y del mismo modo para prevenir la reincidencia en dichas patologías. RC constituye actividades físicas las cuales buscan que los pacientes consigan niveles óptimos físicos y mentales para reintegrarse a la vida cotidiana.
En este estudio los pacientes serán escogidos al azar para que estén frente a un robot y otros no estarán frente al robot.

El propósito de esta investigación es hacer mediciones de su caminata sobre la banda sin fin (Caminadora o trotadora), con unos electrodos que se colocarán sobre la piel a nivel del tronco y el brazo y serán detectados por un robot que estará frente a su banda sin fin, dichos electrodos y el robot no tienen ningún efecto sobre su salud y no producen molestia ya que no son invasivos. Posteriormente se analizarán los datos medidos y servirán para analizar el efecto del robot en la motivación y el desempeño de los pacientes a hacer mejor el ejercicio. Con el fin de realizar una evaluación de su respuesta cardiovascular (del corazón y los vasos sanguíneos) se harán mediciones de frecuencia cardiaca y tensión arterial, antes, durante y después del ejercicio. Esto servirá para ver si los pacientes que están frente a un Robot tienen mejor desempeño en el ejercicio, están más motivados y mejoran su condición física Si usted queda en el grupo sin Robot tendrá las mismas mediciones que el otro grupo que tiene el robot.

Tipo de Intervención de Investigación

Esta investigación incluirá un protocolo de medidas no invasivas durante la sesión de rehabilitación cardíaca sobre banda sin fin durante 36 sesiones en la fase II o III. Se harán mediciones de como usted realiza la caminata sobre la banda sin fin como son: la velocidad, la amplitud del paso. De igual manera se le registrará la frecuencia cardiaca y tensión arterial, antes, durante y después del ejercicio. Su participación en esta investigación es totalmente voluntaria. Usted puede elegir participar o no hacerlo. Tanto si elige participar o no, continuarán todos los servicios que reciba en esta institución y nada cambiará. Usted puede cambiar de idea más tarde y dejar de participar aun cuando haya aceptado antes.

Riesgos

Al participar en esta investigación usted no se expone a un riesgo mayor que si no lo hiciera.

Molestias

Al participar en esta investigación puede existir una mínima molestia ocasionada por el contacto de electrodos con la piel al nivel del tronco y el brazo.

Incentivos

No se le dará ningún dinero, regalos o incentivos por tomar parte en esta investigación.

Confidencialidad

En este proyecto la información será vinculada, es decir, la información puede relacionarse o conectarse con la persona a quien se refiere. Sin embargo, esta información será registrada de forma anónima, en este caso no se puede vincular con la persona a quién se refiere excepto mediante un código u otros medios conocidos solo por el titular de la información. De esta forma se protege la información personal de los sujetos participantes. Su identidad nunca será revelada o publicada

Compartiendo los Resultados

Durante el estudio, los participantes podrán conocer en todo momento el estado del proyecto de investigación y los resultados preliminares, se buscará la divulgación de los resultados definitivos que se obtengan de esta investigación, para que otras personas interesadas puedan aprender. No se compartirá información confidencial.

Derecho a negarse o retirarse

Usted no tiene por qué tomar parte en esta investigación si no desea hacerlo. Puede dejar de participar en la investigación en cualquier momento que quiera. Es su elección y todos sus derechos serán respetados.

A Quién Contactar

Si tiene cualquier pregunta puede hacerlas ahora o más tarde, incluso después de haberse iniciado el estudio. Si desea hacer preguntas más tarde, puede contactar cualquiera de las siguientes personas Mónica Rincón R. Correo electrónico: mrinron@hotmail.com, Teléfono-Extensión: 6672727 ext 51401,51402, 51406

Parte II: Formulario de Consentimiento

Yo, ____, identificado con cedula de ciudadanía número ____________________________, declaro que he leído y comprendido el presente documento y que mis preguntas han sido respondidas satisfactoriamente; por lo tanto doy mi consentimiento informado para participar en la investigación llamada "Evaluación del impacto de la intervención de un robot social en las respuestas cardiovasculares de los pacientes del programa de Rehabilitación Cardiaca de la Fundación Cardio-Infantil Instituto de Cardiología.". Estoy de acuerdo en que mi nombre, edad y otros datos antropométricos sean almacenados. Sé que puedo retirarme del experimento en cualquier momento.

Sujeto Participante:
Nombre: ___________________________________
Dirección: ___________________________________
Teléfono: ___________________________________
Firma: ___________________________________ Cédula: ___________________________________

Declaración del investigador
Yo certifico que le he explicado a esta persona la naturaleza y el objetivo de la investigación, y que esta persona entiende en qué consiste su participación, los posibles riesgos y beneficios implicados. Todas las preguntas que esta persona ha hecho le han sido contestadas de forma adecuada. Así mismo, he leído y explicado adecuadamente las partes del consentimiento informado. Hago constar con mi firma.

Testigo:

Nombre: ___________________________________ Cédula: ___________________________________
Firma: ___________________________________

Investigador:
Nombre: ___________________________________ Cédula: ___________________________________
Firma Investigador: ___________________________________

Fecha (aaaa/mm/dd): ___________________________________

Colombian School of Engineering Julio Garavito

Fundación Cardioinfantil – Instituto de Cardiología

Project: Protocol 2 – Study in perception of a social robot for personalised cardiac rehabilitation

Date: ___________
Name: ____
ID:____
Age:____
Gender: M____ F____
Level of Education: Primary School__ High School__ Bachelors__ Postgraduate__

The following questionnaire attempts to evaluate five key concepts of the perception of social robotics focused on cardiac rehabilitation. This questionnaire is aimed at patients who have NOT had any therapy assisted by the robot.
It is very important to read the following introduction so that all patients have the same information about the robot, its operation and the purpose of this type of therapy. You must have a robot while we are giving this information in order to understand its functionality.

Physical activity has had multiple health benefits, such as reducing mortality rates caused by heart disease, making it one of the most prominent components of the cadiaca rehabilitation program. This program aims to evaluate the progress of the patient suffering from these diseases. On the other hand, there is evidence that robots can motivate patients during therapeutic procedures, why, this project is oriented to the evaluation of routines typical of cardiac rehabilitation through the integration of a robotic agent. The system has sensors that control the values of the variables used by therapists to monitor the progress of therapy:

- Cardiopulmonary parameters: Heart rate and its evolution through the sessions.
- Gait spatio-temporal parameters: cadence, step length, and gait velocity.
- Exertion parameters: Borg scale

Additional Questions:

1. Would you use the robot during the therapy? Why?

2. What expectations do you have about the therapy assisted by a robot?

Colombian School of Engineering Julio Garavito

Fundación Cardioinfantil – Instituto de Cardiología

Project: Protocol 1 – Study in perception of a social robot for personalised cardiac rehabilitation

Date: ___________
Name: ____
ID:____
Age:____
Gender: M____ F____
Level of Education: Primary School__ High School__ Bachelors__ Postgraduate__

The following questionnaire attempts to evaluate five key concepts of the perception of social robotics focused on cardiac rehabilitation.

Additional Questions:

1. Would you recommend the use of the robotic system to the patients that are starting the rehabilitation therapy?

2. According to your experience, what would you recommend to improve the robot-based therapy?

Table F.1: Adapted Unified Theory of Acceptance and the Use of Technology (UTAUT) questionnaire developed for all conditions, in terms of perceived usefulness (U), perceived utility (PU), perceived safety (S), ease of use (EU), perceived trust (PT), perceived sociability (PS) and social presence (SP).

Construct	No.	Question
U	1	I consider that a robot is a good tool to assist cardiac rehabilitation therapies.
	2	I consider that my interaction with the robot was comfortable.
	3	I enjoyed when the robot gave me verbal encouragement when I did a good job.
	4	I am satisfied with the work that the robot did.
	5	I consider that the robot adapts to my needs.
PU	1	I consider that the interaction with the robot was beneficial for my recovery.
	2	I consider that the role of the robot was important for the therapy development.
	3	I think that the use of the robot helps me to commit doing my task well.
S	1	I feel safe during the sessions when working with the robot.
	2	I consider it easy to give information to the robot.
EU	1	I consider that the robot is ease to use.
	2	I consider that using the robot did not affect the time of the therapy sessions.
	3	I consider that the robot's instructions were clear.
PT	1	The robot made me confident.
	2	I did instruction the robot told me because I trusted it.
	3	I like using the robot during the therapies.
	4	It gave me confidence that the robot guides my therapy.
PS	1	I consider the robot to be a pleasant conversational partner.
	2	I find the robot pleasant to interact with.
	3	I feel the robot understands me.
	4	I think the robot is nice.
SP	1	When interacting with the robot I felt like I am talking to a real person.
	2	It sometimes felt as if the robot was really looking at me.
	3	I can imagine the robot to be a living creature.
	4	I often think the robot is not a real person.
	5	Sometimes the robot seems to have real feelings.

Table F.2: Adapted Working Alliance Inventory (WAI) questionnaire that measures the long-term perception of the robot within the *personalised robot* condition, with *Bond*, *Task* and *Goal* constructs.

Construct	Formulation	Question
Bond	Positive	The robot and I understand each other. I believe the robot likes me. I believe the robot is genuinely concerned for my welfare. The robot and I respect each other. I am confident in the robot's ability to help me. I feel that the robot appreciates me. The robot and I trust one and other. My relationship with the robot is very important to me. I feel the robot cares about me even when I do things that the robot does not understand me.
	Negative	I feel uncomfortable with the robot. I feel the robot is not totally honest about its feelings toward me. I have the feeling that if I say or do the wrong things the robot will stop working with me.
Task	Positive	The robot and I agree about the things I will need to do in the therapy to help improve my situation. What I am doing in the therapy gives me new ways of looking at my problem. I am clear on what my responsibilities are in therapy. I feel that the things i do in therapy will help me to accomplish the changes that I want. I am clear as to what the robot wants me to do in these sessions. The robot and I are in agreement on what is important for me to work on. I believe the way that the robot and I are working in my problem is correct.
	Negative	I find what I am doing in therapy confusing. I believe the time robot and I are spending together is not spent efficiently. I find that the robot tasks during the therapy are unrelated to my concerns. I am frustrated by the things I am doing in therapy. The things that robot is requesting me do not make sense.
Goal	Positive	The robot perceives accurately what my goals are. I wish that the robot could configure the therapy according the purpose of our session. The goals of these session are important to me. The robot and I are working towards mutually agreed upon goals. As a result of these session I am clearer as to how I might be able to change. The robot and I collaborate on setting goals for my therapy. The robot and I established a good understanding of the kind of changes that would be good for me.
	Negative	I am worried about the outcome of these sessions. I disagree with the robot about what I ought to get out of therapy. The robot does not understand what I am trying to accomplish in therapy. The robot and I have different ideas on what my problems are. I do not know what to expect as the result of my therapy.

Glossary

D-**All** All samples dataset: Each user is encountered different amount of times (10-41) in the dataset. xxv, 91, 93–95, 97–108, 110, 111, 113, 298, 299

D-**Ten** Ten samples dataset: Each user is encountered exactly ten times in the dataset. xxv, 91, 93–95, 97–105, 111, 113, 298

attention Attention in this thesis refers to attention mechanisms, which focus on particular elements of a task to respond to queries, through non-uniformly weighting parts of the input to optimise the learning and recall processes. 49, 51, 54, 159, 161–164

Bayesian network Bayesian network is a probabilistic graphical model which represents conditional dependencies of a set of variables through a directed acyclic graph. 5, 40, 58

catastrophic forgetting Catastrophic forgetting refers to the drastic loss of performance on previously learned classes when a new class is introduced. 5, 42, 107, 126, 176, 246, 252, 259

closed-set recognition Closed-set recognition assumes that the probe should belong to a known user, that is, all users to be identified are previously enrolled in the system. 38, 81, 86

continual learning Continual (or lifelong) learning in machine learning refers to learning information continuously, incrementally and adaptively. In other words, it refers to incremental learning with adaptation, but it covers both batch and online (sequential) learning. 58, 128, 155, 157, 175, 176, 189, 246

data-driven Data-driven approaches rely on extracting the structures and values from the training data, instead of using a knowledge-base or rules. vii, 4, 8, 9, 12, 15, 17, 20, 32, 33, 48, 49, 51, 53, 115, 118–120, 122, 123, 125, 126, 128, 132, 133, 135, 136, 155, 157, 160, 169, 172–174, 176, 179, 193, 194, 244, 246, 247, 250, 255–260

DIR Detection and Identification Rate is the fraction of correctly classified probes (samples) within the probes of the enrolled users. See Equation 3.13. xv, 66, 73, 89, 128, 228, 253, 299

end-to-end End-to-end learning is training the system from input-to-output as a whole, thereby, removing the need of any intermediate steps of processing. 4, 163, 166

FAR False Alarm Rate is the fraction of incorrectly classified probes within the probes of unknown users. See Equation 3.14. xv, 66, 78, 89, 128, 227, 253, 299

few-shot learning Few-shot learning refers to the ability to learn from a few labeled samples in machine learning. 156, 157, 170, 189, 190, 194, 195, 246

FTC Failure to Capture error occurs when a biometric system cannot obtain meaningful data (e.g., a face cannot be detected in the image). 39

FTE Failure to Enroll error is fraction of users that cannot be successfully enrolled in a biometric system because of FTC error. 39, 78

generative model Generative models generate a response word-by-word based on the conversation history (context). 48, 50, 158, 174, 194, 195

hybrid normalisation Hybrid normalisation is an introduced normalisation technique that combines the normalisation methods that achieve the lowest loss for each modality. In other words, hybrid normalisation uses the best performing normalisation method for each modality. 68, 87, 90, 92, 97–100, 109, 113, 114, 245, 300

incremental learning Incremental learning refers to expanding a model for new users or attributes with input data. 4, 5, 32, 38, 42, 43, 157

knowledge-base Knowledge-base is a structured database with entities and their corresponding values and relations. 4, 47, 120, 140

long-term recognition performance loss The introduced performance metric that creates a balance between DIR and FAR for long-term interactions, based on the average number of observations per user, as presented in Equation 3.15. 23, 67, 68, 70, 89, 92, 114, 245, 253

online learning Online learning is updating the model sequentially with incoming information. 4, 43, 58, 71, 89

open world recognition Open world recognition refers to incremental learning of new ("unknown") classes, in addition to recognising previously learned classes. 5, 38, 41–43, 57, 70, 72, 84, 87, 89, 128, 175, 245

open-set recognition In open-set recognition, the probe might belong to a previously enrolled user or an "unknown" user, and it should be identified as such. 38, 41, 42, 44, 71, 81, 86

primary biometrics Primary biometrics are identifiers that can help uniquely identify a person, such as face recognition or fingerprints. 5, 41, 57, 60

quality of the estimation The introduced confidence measure that compares the highest posterior probability to the second highest that enables decreasing the number of incorrect recognitions. See Equation 3.5. 23, 62, 67, 68, 70, 72, 76, 83, 84, 86, 94, 98–100, 108, 252

retrieval-based model Retrieval-based (ranking) models select a dialogue response from a set of predefined responses (candidates). 48, 49, 158, 163, 165, 174, 178

service robot A service robot is a robot that performs useful tasks for humans or equipment excluding industrial automation applications, based on its current state and sensing without human intervention. 9, 35

socially assistive robotics Socially assistive robotics refers to the assistive and supportive robotics applications in social interactions, such as in healthcare and therapy. 9, 36, 201

soft biometrics Soft biometrics are ancillary physical or behavioural characteristics, such as age and gender, that cannot uniquely identify a person. 5, 41, 59, 89

Acronyms

AI Artificial Intelligence. 1, 7, 11, 25, 46

ANOVA analysis of variance. 95, 96, 100, 214, 230

API application program interface. 47, 120, 122, 178

APRIL Applications of Personal Robotics for Interaction and Learning. 21

ASR automatic speech recognition. vii, 6–8, 16, 22, 27, 47, 126, 137, 140, 142, 143, 148, 149, 151, 152, 193, 206, 246, 249, 254, 255, 258

BN Bayesian network. 5, 6, 40–42, 58, 59, 62–64, 69, *Glossary:* Bayesian network

BP blood pressure. 203, 204, 212, 216, 217

BS Borg scale. xxii–xxiv, 203, 204, 206–208, 212, 216–219, 222–225, 230, 233, 235

CR cardiac rehabilitation. vii, xxiv, 10, 12, 18, 20, 200, 202–205, 208–211, 214–216, 219–226, 228–234, 236–241, 243, 248, 250, 260–262

DIR Detection and Identification Rate. xv, 66, 67, 73, 78, 81–86, 89, 92, 94, 96, 100, 103–106, 111, 112, 128, 228, 253, 299, *Glossary:* DIR

DM dialogue manager. 47

DMN Dynamic Memory Networks. 49

DP dialogue policy. 47, 143

DST dialogue state tracking. xvi, xviii, 47, 143, 155, 156, 170, 180, 181, 183, 185–188, 190, 194, 195, 247, 321, 323

EM Expectation Maximization. 19, 44, 55, 64, 249

EVM Extreme Value Machine. 15, 42, 89, 99, 101, 106, 110, 112–114, 245, 246, 297, 300

EVM:FR Extreme Value Machine trained with face recognition data. 99, 101, 102, 104, 105, 107, 108, 110, 299

EVM:MM Extreme Value Machine trained with multi-modal data. 99, 101, 102, 104, 107, 108, 110, 112, 299

FAR False Alarm Rate. xv, 66, 67, 78, 79, 81–86, 89, 94–96, 100, 103, 105, 106, 111, 128, 227, 228, 253, 299, *Glossary:* FAR

FCI-IC Fundación Cardioinfantil-Instituto de Cardiología. 197, 201, 202, 206–212, 215, 247, 260, 261

FR face recognition. 4, 5, 38, 41, 44, 58, 60, 62, 64, 68, 69, 73, 74, 76, 79–81, 83–85, 89, 92, 94, 95, 99–114, 227, 228, 246, 252

FSM finite state machine. 47, 207

FTC Failure to Capture error. 39, *Glossary:* FTC

FTE Failure to Enroll error. 39, 78, 80, *Glossary:* FTE

GRU gated recurrent neural networks. 49

GUI graphical user interface. 238, 261

HR heart rate. xxii–xxiv, 203, 204, 206–208, 211, 212, 216–219, 221–225, 231–234, 238

HRI Human-Robot Interaction. vii, 1–3, 5–9, 11, 13, 15, 17, 19–21, 23–27, 29, 31, 32, 34, 36, 37, 41, 44, 45, 48, 51–54, 57, 58, 60, 63, 65, 69, 70, 72, 73, 76, 84, 89–91, 113, 117–119, 122, 123, 125, 132, 139, 143, 157, 162, 205, 211, 243–246, 250–252, 254, 255, 258

HSD Honestly Significant Differences. xxv, 97, 101, 103, 105, 106, 108, 298

KB knowledge-base. 4, 32, 33, 47, 50–53, 115, 120, 125, 126, 132, 134, 135, 140, 141, 143, 151, 155, 170, 179, 195, 247, *Glossary:* knowledge-base

LSTM long short-term memory. 48, 160, 162

MemN2N End-to-End Memory Networks. xvi, 50, 51, 158, 163–166, 171, 172, 174, 177–179, 182–184, 188, 189, 191–195, 247, 250, 256–259, 313, 315, 316

MMIBN Multi-modal Incremental Bayesian Network. vii, 18, 19, 22–24, 55, 56, 58, 59, 65, 67–74, 76, 79–81, 83–89, 92, 94, 98–108, 110–114, 137, 141, 151, 152, 227, 228, 244–249, 251, 252, 299, 300

MMIBN:OL Multi-modal Incremental Bayesian Network with Online Learning. 65, 73, 76, 78, 79, 81, 83, 84, 89, 92, 94, 99, 101–106, 108, 110, 112, 141, 198, 208, 222, 227, 228, 299

NLG natural language generation. 6, 27, 47, 143

NLP natural language processing. 6, 8

NLU natural language understanding. 6, 27, 47, 143

OL online learning. 4, 6, 11, 13, 15, 18, 19, 23–25, 28, 29, 32, 43, 44, 50, 53, 55, 58, 64–66, 70–74, 79, 81, 83–85, 89, 92–94, 96, 98, 99, 101, 102, 108, 110, 112, 114, 141, 208, 222, 227, 228, 244–249, 252, 253, 259, *Glossary:* online learning

OOV out-of-vocabulary. xvi–xviii, 122, 123, 126, 134, 169–172, 182, 185–189, 256, 257, 259, 260, 308–310, 317–320

POMDP Partially Observable Markov Decision Process. 32, 48, 255, 259

QA question answering. 49, 162, 166

R-HR recovery heart rate. xxiii, 204, 212, 216, 217, 224, 231–233

RBDMS rule-based dialogue management system. 7–9, 12, 15, 16, 47, 135, 138, 143, 149, 151, 169, 172, 179, 194, 246, 249

RNN recurrent neural networks. 162

ROC receiver operating characteristic. 78

RoSAS Robotic Social Attributes Scale. xxi, 146–148, 301

SAR socially assistive robotics. 9, 10, 36, 37, 201, 205, 206, 212, 236, 237, *Glossary:* socially assistive robotics

SB soft biometrics. 5, 25, 41, 45, 58, 59, 69, 89, 99, 101–104, 111–113, *Glossary:* soft biometrics

Seq2Seq Sequence-to-Sequence. 20, 48, 157, 158, 160, 161, 167–169, 171, 174, 178, 179, 183–186, 190, 191, 193–195, 247, 250, 256–259, 320

SGD stochastic gradient descent. 159, 160, 166

SVM Support Vector Machine. 39, 41, 42, 44, 162

TTS text-to-speech. 7, 47, 142, 143

UTAUT Unified Theory of Acceptance and the Use of Technology. xvi, xviii, 212–214, 216, 220, 226, 227, 235–237, 325, 329

WAI Working Alliance Inventory. xviii, xxiii, 212–214, 216, 219, 225, 226, 325, 330

WoZ Wizard-of-Oz. 4, 27, 35, 52, 118, 119, 124, 257