2023-10-20

Confirmation Report: Modelling Interlocutor Confusion in Situated Human Robot Interaction

Na Li
Technological University Dublin, d19125334@mytudublin.ie

Funder: Science Foundation Ireland

Confirmation Report:

Modelling Interlocutor Confusion in Situated Human Robot Interaction

Na Li

Technological University Dublin

Supervisors:

Dr. Robert J. Ross

October 2023

Abstract

Human-Robot Interaction (HRI) is an important but challenging field focused on improving the interaction between humans and robots such to make the interaction more intelligent and effective. However, building a natural conversational HRI is an interdisciplinary challenge for scholars, engineers, and designers. It is generally assumed that the pinnacle of human-robot interaction will be having fluid naturalistic conversational interaction that in important ways mimics that of how humans interact with each other. This of course is challenging at a number of levels, and in particular there are considerable difficulties when it comes to naturally monitoring and responding to the user's mental state.

On the topic of mental states, one field that has received little attention to date is monitoring the user for possible confusion states. Confusion is a non-trivial mental state which can be seen as having at least two substates. There two confusion states can be thought of as being associated with either negative or positive emotions. In the former, when people are productively confused, they have a passion to solve any current difficulties. Meanwhile, people who are in unproductive confusion may lose their engagement and motivation to overcome those difficulties, which in turn may even lead them to drop the current conversation. While there has been some research on confusion monitoring and detection, it has been limited with the most focused on evaluating confusion states in online learning tasks.

The central hypothesis of this research is that the monitoring and detection of confusion states in users is essential to fluid task-centric HRI and that it should be possible to detect such confusion and adjust policies to mitigate the confusion in users. In this report, I expand

on this hypothesis and set out several research questions. I also provide a comprehensive literature review before outlining work done to date towards my research hypothesis, I also set out plans for future experimental work.

Table of contents

Chapter 1

Introduction

Human-Robot Interaction (HRI) is an interdisciplinary field. It is related to various disciplines, including human-computer interaction (HCI), robotics, artificial intelligence, design, and philosophy. HRI is also considered its own unique discipline concerned with building concepts, methods, and HRI frameworks (Bartneck *et al.*, 2020; Sharkawy, 2021; Cantrell *et al.*, 2010). Therefore, it is beneficial to study HRI when scholars, engineers, and designers work together in different task environments such as digital learning environments (Pachman *et al.*, 2016), domestic environments (Kontogiorgos *et al.*, 2020), laboratory environments (Morales *et al.*, 2019), and even noisy and unpredictable environments (Kontogiorgos *et al.*, 2020), *etc.*.

HRI has been the subject of great interest since the 1940s. Isaac Asimov (2 January 1920–6 April 1992) coined the term "robotics". Asimov also raised three questions within his stories to emphasise the relationship between humans and robots (Bartneck *et al.*, 2020): "How much will people trust robots?", "What kind of relationship can a person have with a robot?" and "How do our ideas of what is human change when I have machines doing human-like things in our midst?". These questions are still relevant to our research today.

In 1978, the term "social robot" was first mentioned in the context of robotics in an article called the "social robot" in the *Interface Age* magazine. It was mentioned that the

"social robot" has the social skills to handle human conversations in a domestic setting. Since then, some social robots have been created to play different roles in the human world. For example, the Kismet robot played the role of a "infant" in that it was assisted and taught by a human caretaker (Breazeal and Velásquez, 1999). Subsequently, there was, for example, the playful Keepon robot that was designed to interact with children (Kozima *et al.*, 2009b); In the last decade, robots such as Nao and Pepper, humanoid robots from the Softbank Robotics company [1], have commonly been seen in customer assistance roles such as greeting robots in banking, healthcare [2], education, and retail (Abbas *et al.*, 2020; Ikeuchi *et al.*, 2018; Lehmann and Svarny, 2021; Song and Kim, 2022).

HRI is distinguished from the field of pure robotics, wherein physical robots are designed to manipulate physical tasks – also called physical interaction (Bartneck *et al.*, 2020). For instance, food-serving robots were used in the Beijing Winter Olympics 2022, these robots can cook food, make coffee, serve cocktails and deliver these items to the customers who need them. Thanks to these physical interventions, people were able to effectively reduce the chance of social contact to avoid the possibility of COVID-19 transmission.

In contrast, social robots interact with humans across diverse domains. In general, the trend, driven by technology, has moved researchers from the case of the embedded robotics system to the robotics enabled through a spoken-dialogue system. Examples of this include tutors for children's study and online learning, tour guides in a museum, and health assistants (Gordon *et al.*, 2016; Doherty and Doherty, 2018a; Duchetto *et al.*, 2019; Esterwood and Robert, 2020). Thus, HRI is an activity by which social robots interact with people in a human world through natural mechanisms. These robots should be designed such that they follow social rules in physical environments to make people feel safe and comfortable, and even help users develop a strong enthusiasm for interaction with the robot.

[1] https://www.softbankrobotics.com/emea/en
[2] https://www.softbankrobotics.com/emea/en/pepper-healthcare-ga

Conversation, including verbal and nonverbal interaction, is a straightforward and essential way of communicating in our daily social lives. Susan Brennan defined the term conversation as "A joint activity in which two or more participants use linguistic forms and nonverbal signals to communicate interactively" (Brennan, 2010). Appropriately, then, in human-to-social robot interaction, the conversational capabilities of social robots are often considered the principal functionalities for designing effective natural interaction.

To ensure that users continue to engage in the interaction process, a smooth and fluid conversation is necessary in HRI. Conversational HRI should be designed to include appropriate responses to the words, mental states, and related emotions of the interlocutor. People's emotions, undoubtedly, can stimulate and modulate their behaviour during ongoing experiences (Bartneck *et al.*, 2020). Due to the fact that emotions play such an important role in human social cognition, it is useful to design emotional communication in situated HRI environments. However, existing research in this area still faces many challenges.

First, some research has only considered facial expression for the recognition of emotions (Barrett *et al.*, 2019; Roy and Etemad, 2021). However, from a single facial expression, it is hard to recognise the exact emotion without also seeing a body's behaviours. Aviezer *et al.* (2012) argued that different body behaviours with the same facial expression can convey different information and emotions in real life. Second, it frequently happens that training data for modelling emotion recognition have been collected from actors; therefore, it loses the natural behaviour condition (*e.g.*, Busso *et al.* (2008); Celiktutan *et al.* (2017)). Third, only a small range of emotion categories have been used for emotion recognition. For example, the six most popular emotions classes: anger, disgust, fear, happiness, sadness, and surprise have been modelled (Barrett *et al.*, 2019), as have two additional classes of neutral and contempt (Greco *et al.*, 2019a), but yet there are many other nuanced emotions and mental states that have not been modelled. Finally, there is a lack of research on user engagement estimation in situated conversational HRI. Some studies have attempted to detect people's behaviours in

spontaneous conversation under a specific situation (*e.g.* Ben Youssef *et al.* (2017)), or to detect a certain group of people's behaviours (*e.g.*, Tapus *et al.* (2012)), or instead focus on online learning systems that are a kind of human-computer interaction that is quite different to human-robot interaction (Doherty and Doherty, 2018a).

Moving beyond emotion in general, confusion is a unique mental state that can either precede a high degree of positive engagement in a task, or can also be correlated with negative states such as boredom and subsequent disengagement from a conversation (D'Mello *et al.*, 2014). Estimating confusion states of a user can hence be a very important step in improving the pragmatics modelling properties of an interactive system. By checking for confusion, or indeed precursors of confusion, we can in principle adjust the dialogue policy or information being presented to the user in order to assist them in the specific task being undertaken. Such monitoring can be seen as a specific form of engagement detection (Sidner *et al.*, 2004; Dewan *et al.*, 2018). In mainstream Human-Computer Interaction (HCI) studies, there have to this point been a number of studies that have investigated the modelling and detection of confusion (Kumar *et al.*, 2019; Grafsgaard *et al.*, 2011; Zhou *et al.*, 2019). However, the majority of studies in this area have concerned online learning. Little work has focused on general engagement or task-oriented dialogue in HRI.

To enhance conversational HRI in situated spoken-dialogue interaction, we propose that detecting an interlocutor's different states of confusion in a situated interaction may be an impactful strategy for enhancing long-term user-system engagement. In general terms, confusion can occur at any time in social interaction when people want to express information and thoughts to each other. Similarly, when people interact with a computer, whether it is an online agent or a real robot, people may be confused whenever they encounter some obstacles to communication or task completion, *e.g.*, complex information, inconsistent information, contradictory information, or incongruous feedback (Lehman *et al.*, 2012) *etc.*. Therefore, through observing and recording people's natural behaviours when they are confused or not

confused in our situated HRI, we can model the different states of confusion of the user, and hence attempt to mitigate this confusion.

This report aims to comprehensively present my overall research goal, research questions, and experiments that I have designed to help address these. A review of the state-of-the-art in conversational HRI is first presented in Chapter 2, Chapter 3 then outlines my main research question along with 6 sub-research questions with explanations around these studies. The work I have done to date is presented in Chapter 4, and what follows is a discussion of ongoing studies and future work in Chapter 5. Then, a discussion including the current research progress and my achievements are presented in conclusion in Chapter 6.

Chapter 2

Literature Review

To understand conversational HRI and related studies of confusion, in the review, I jointly present the different aspects that are relevant to conversational HRI and confusion modelling. Figure 2.1 shows an overview of the connections between the areas of my work. The first of these is a high-level overview of conversational HRI in terms of verbal and nonverbal interaction in HRI, and also the novel applications of HRI. Next, I focus on affective computing in HRI. In this part, I examine the general concept of affective HRI computing. Engagement detection and emotion estimation are two foundational areas in HRI, which are closely associated with confusion detection studies in HRI; this is thus the next area that I examine. Given its importance to my work here, the following section looks at the topic of confusion in HRI across four aspects: definitions of confusion in specific study environments, states of confusion, confusion induction, and confusion detection models. Finally, given the fact that my work aims to verify situated HRI, I then examine some of the similarities and contrasts between two of the most well-known embodiments of HCI, *i.e.*, human avatar interaction, and human-robot interaction in the final section.

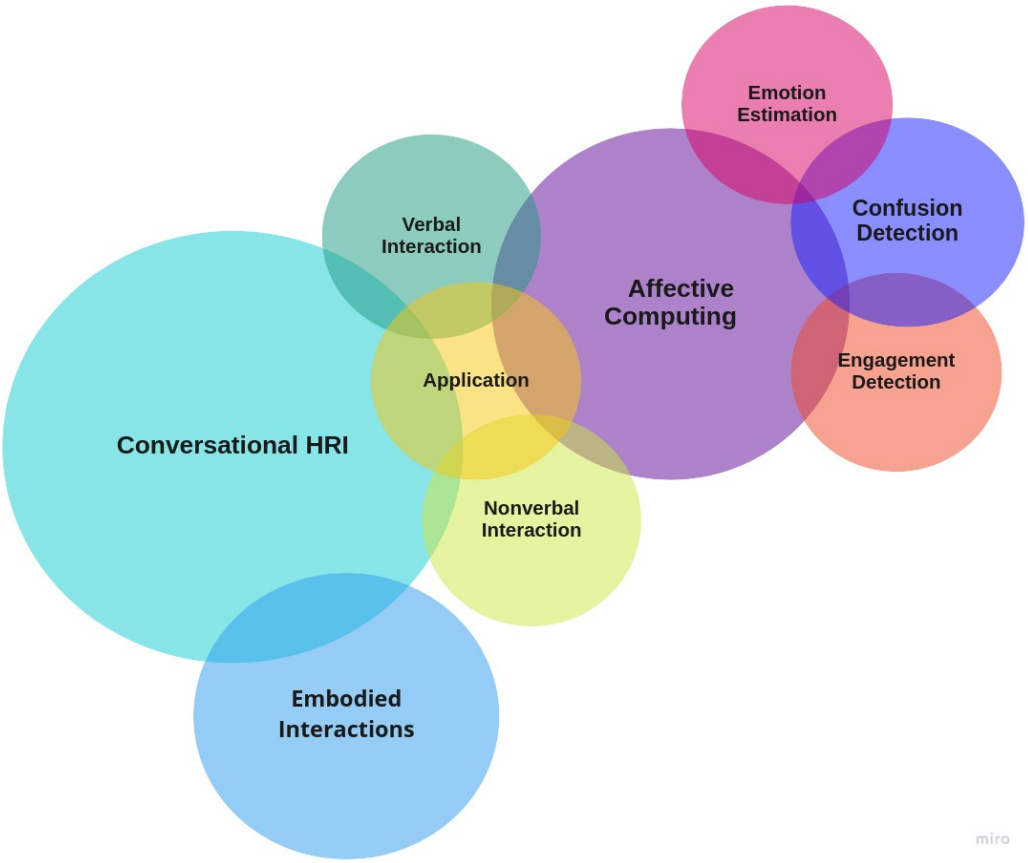


Fig. 2.1 Relationships between the areas that are related to my work

2.1 Conversational HRI

Increasing numbers of scientists, engineers, and designers are interested in developing multiple capabilities for human-like conversational interaction on social robots (Bartneck *et al.*, 2020; Breazeal, 2004). In social communication, a conversation is a significant but a complex way of conveying information. In particular, a face-to-face conversation between two humans is composed of multiple modalities of communication that include human voice, speech, facial expression, articulated gestures, and body posture (Breazeal, 2004; Cassell, 2001; Mavridis, 2015). Embodied conversational agents are one of the important mediums in Human-Computer Interaction (HCI), such that the agent can interpret the social behaviours of a human from a set of verbal and nonverbal behaviours by using multiple

devices (*e.g.*, cameras, microphones, sensors). Therefore, when a social robot has a face-to-face conversation with a human, it needs to have the same social signals as a human, such as time control to provide feedback or appropriate responds to the interlocutor (Hoffman *et al.*, 2014; Skantze, 2021). Meanwhile, a social robot should have the ability to recognise affect, emotion, or involvement by observing the behaviour of the interlocutor, such as tracking the gaze of the eye, the pose of the head, facial expressions, deictic gestures, or biological behaviours (*e.g.*, heartbeat, electroencephalography activity (EEG), and body temperature) (Admoni and Scassellati, 2017; Ginevra Castellano and W.Schuller, 2004; Busso *et al.*, 2008; Fischer *et al.*, 2019).

2.1.1 Verbal Interaction

Talking/Speech is the most common form of communication between humans, as it is explicit and straightforward to share information. It should thus also be a basic functionality of the robot to interact with people. But, of course, understanding language or speech is more complex than producing robot speech. Bartneck *et al.* (2020), for example, emphasised that a robot should be capable of transcribing speech into texts and understand words to generate speech by presenting appropriate responses. In light of its importance, over the last number of decades, many language and speech-based technologies have been developed in both the context of human-computer interaction (*e.g.*, avatar, chatbot) (Crovari *et al.*, 2021) and HRI. The development of smooth and natural communication is a crucial technology in, of course, HRI and HCI (Mubin *et al.*, 2014; Forsberg, 2003).

Automated speech recognition (ASR) is a well-known technology in which a transcription process converts a digital recording of human speech to words/texts (Novoa *et al.*, 2021, 2018). Technically speaking, recorded speech is in the time domain, making it difficult to transcribe the speech into words, so the speech must first be converted into the frequency domain. Thereafter, a classical speech recognition system would use Gaussian Mixture

Models (GMMS) based on hidden Markov models (HMM) to extract features (phonemes, words, and sentences) from these data that have different strong phonemes (Bartneck *et al.*, 2020; Pranto *et al.*, 2021). In the traditional architecture, these raw inputs would be processed by probabilistic models to string phonemes and words together into words and sentences. In recent years, researchers have instead been using deep neural networks (DNN) instead of those probabilistic models for speech processing, *e.g.*, feature transformation and dimensional reduction (Singh *et al.*, 2018; Nassif *et al.*, 2019) as well as for speech recognition itself. As a result, not only has speech recognition performance improved dramatically, leading to a higher rate of correct recognition, but speech recognition systems can also manage background noise, incorrect speech, *etc.*. Many software companies *e.g.*, Google, IBM, and Microsoft provide cloud-based speech recognition and services to integrate speech recognition functionality into varied applications, including robotics (Nassif *et al.*, 2019; Bartneck *et al.*, 2020).

Voice-activity detection (VAD) is a related field to ASR that is beneficial for all speech and audio processing applications, as it can distinguish effective speech from non-speech, *e.g.*, background noise, echo speech, or silence, in order to robustly constrain the capacity and coverage of communication bandwidth (Chang *et al.*, 2006). When contrasted with traditional VAD systems, advanced VAD systems have been developed using deep learning to enhance VAD performance, *e.g.*, the NAS-VAD framework which uses neural architecture search (NAS) to optimise the VAD task (Rho *et al.*, 2022). Indeed, the VAD needs not to be only audio based; Visual Voice Activity Detection provides visual input from a robot's camera, and can detect whether an interlocutor is speaking (Lubitz *et al.*, 2021). Thus, VAD systems are advantageous in implementing robotic dialogue systems (*e.g.*, Robotics Dialogue System (RDS)) as the verbal capacities of robots, such as speaking language, turn-taking without understanding the interlocutor's speech, localising and identifying users and users' gender, or recognising voice emotion (Alonso-Martin *et al.*, 2013) can be enhanced.

Moving beyond speech, *Language understanding* is a subfield of Natural Language Processing (NLP) (Bates, 1995) that focuses on the understanding of the content itself. Semantic analysis is a substudy field of NLP closely related to national language understanding that aims to extract context from spoken language or text data (Bartneck *et al.*, 2020; Salloum *et al.*, 2020). It is used to detect specific topics and related features from the sentiment of people (Hussen Maulud *et al.*, 2021). In conversational HRI, NLP and semantic analysis can also be involved in classifying the emotional state of speakers. Incremental NLU and semantic analysis is an advanced framework in conversational systems and HRI, in that it allows robots to respond quickly when they do not understand the speaker's expression, and quickly detect and react to the main syntactic, semantic or pragmatic ambiguities (Cantrell *et al.*, 2010; Braun *et al.*, 2017).

National language understanding (NLU) in the true sense is a fundamental task in the construction of task-oriented dialogue systems (Tseng *et al.*, 2020). It can extract keywords (*e.g.*, commands, location, person, event, and date) from a message (Hirschman and Gaizauskas, 2001) to properly respond to a person's contribution. In spoken language, audio signals from a conversation also need to be transcribed into text to be explained at different linguistic levels in a robot system (Bastianelli *et al.*, 2014). Indeed, many methods of NLP for robots focus on sequential signals/messages, but may ignore context, task knowledge, and the physical aspects of spoken users. To establish robust spoken language understanding, Cantrell *et al.* (2010) presented an integrated architecture that assembles speech recognition, incremental parsing, incremental semantic analysis, disfluency analysis, and situated reference resolution. Bastianelli *et al.* (2014), meanwhile, combined contextual knowledge (*e.g.*, the human's position and gaze) and NLU to transfer the human's message and intents to a robot system.

Dialogue management (DM) is the primary process for controlling a dialogue between a user and a computer. Dialogues state tracking (DST) and dialogue policy learning (DPL)

are two main functions of dialogue management (Zhao *et al.*, 2019). DST, as a function, determines how the dialogue state is based on the diversity of conversations between the system and the user (Zhao *et al.*, 2019; Brabra *et al.*, 2021). The core of dialogue management is a framework for managing several rounds of interactions, such as booking a flight or ordering a pizza, and is not commonly studied when applied to simple tasks such as playing music or opening the washing machine (Williams and Young, 2007; Bartneck *et al.*, 2020). The dialogue system explores users' requirements through the way of the DM to keep track of the state of conversations, particularly unknown states from users' spoken utterances, and the dialogue system will in turn ask specific questions for those unknown states before moving forward to the user's next utterance. There are generally three main approaches to implementing DM models (Brabra *et al.*, 2021): First, handcrafted approaches are specific and implemented by developers who develop programmes or models to track the state of conversations and define their policy. For example, QiChat [1] is programmable dialogue management software for Nao and Pepper robots developed by the SoftBank Robotics company. Robots have the ability to detect sentences from users and then give a specific response from the programme. Second, data-driven approaches learn the dialogue state and policy from data mainly through the application of supervised machine learning and data collection from a corpus. The third type of DM model is commonly based on a variant of reinforcement learning, in particular, deep reinforcement learning. The idea is that agents use their self-learning experience and environmental feedback to interact with users and the environment (Zhao *et al.*, 2019); Hybrid approaches meanwhile to combine multiple methods (handcrafted or data-driven) to take advantage of the benefits of each approach.

[1] http://doc.aldebaran.com/2-5/naoqi/interaction/dialog/dialog.html

2.1.2 Nonverbal Interaction

Although verbal interaction is a central means of communication in HRI, it cannot be isolated from nonverbal cues (Mavridis, 2015). The role of nonverbal communication in the interaction between people cannot be ignored, as the interaction is enhanced by nonverbal communication (*e.g.*, eye gaze, facial expressions, gestures, posture, tone of voice, *etc.*), and also it improves people's understanding of the interaction. Not only is it important to design robots' nonverbal behaviours (*e.g.*, sounds, lights, eye colours, physical gestures with body parts) to communicate effectively with humans, but robots should also have capabilities to observe people's behaviours, interpret those behaviours, and respond to these nonverbal cues from humans appropriately. It can be said that the nonverbal channels are more meaningful for robots than for avatars/computers, as robots have more nonverbal communication potential such as through touch and spatial relations.

We also need to consider the specific context of the application that may be related to social and cultural norms. Bartneck *et al.* (2020), for example, notes that nodding heads mean "yes" or "agree" in western societies, while an "agree" expression of the same meaning in India is expressed by head shaking; similarly, for example greeting customs between western culture and Japanese culture are very different. Therefore, in recent HRI research, social and cultural differences have been studied in the context of culturally sensitive interactions and cross-cultural development in nursing robots (Bruno *et al.*, 2017).

Social eye-gazing, as a specific example, is a remarkable nonverbal signal. As Emery (2000) explained, the eyes are unique "hard-wired" pathways in the brain that focus on their interpretation; cognitively, the eyes are also special stimuli from their brain. The human-centred approach is that a robot or an agent is designed to understand the characteristics of human behaviours in the situated spoken dialogue of HRI. Therefore, the HRI design observes the features and limits of human behaviours and perception, whereas the robot is to stimulate different situations to provoke a measurable response. Design-focused approaches

to social robot creation focus on designing a robot, particularly the humanoid robot, with multiple behaviours, *e.g.*, head movements, and gaze to align with human attention. As an example, a minimal social robot, Keepon, can express its attention and affect using gaze and reactive motion (Kozima *et al.*, 2009a). Unfortunately, building computational tools for robot's eye gaze generation in HRI tends to be a technology-focused approach, which focuses on mathematical or technical contributions, but does not tend to focus on the measurement of the interaction effects in a robotics system.

Gesturing is an expressive way to convey information in an interaction; it can also emphasise moments during speech or to persuade people. Again, taking the Pepper robot as an example, it has beat gestures such that its arms, body, and head can move following the rhythm of speech. Moreover, the Pepper robot has vivid gestures automatically and flexibly so that Pepper can do a certain number of human-like body language actions, such as waving a hand when he says "Bye-bye", shaking a hand when he says "Hello"/"Nice to meet you", or different gesture when he says "No"/"Yes" or "You"/"I" *etc.*. Pepper can also be developed to demonstrate advanced body language, such as hugging or dancing, by programming using the animation library for both gesture and animated speech, which includes a list of predefined animations that can be used in an application.

There are of course many classifications of gesture in the literature. As an example, Nehaniv *et al.* (2005) presented five classifications of gestures: "irrelevant/manipulative gesture","side effect of expressive behaviour", "symbolic gesture","international gestures" and "referential/pointing gesture". They explained that robots can achieve limited recognition of situated human gestural motion through the five classifications of gestures in order to the robots will be able to respond appropriately.

Posture and Facial expression are a central research theme in nonverbal affect recognition (Kleinsmith and Bianchi-Berthouze, 2013). Body posture with facial expression can reflect and interpret mental states. Developing postures can enable robots to provide more expressive

information, as body expressions can be a primary way of expressing emotions if there is a lack of facial features (Bartneck *et al.*, 2020). Design-focused approaches focus on designing robots' body language and facial expressions to express robots' emotions in HRI, so that interlocutors can interpret the robots' emotions and even adopt their emotions from the robots' body language (Xu *et al.*, 2014). In contrast, human-centred approaches focus on designing an affect recognition system or model for a robot that is designed to observe user behaviour. Sanghvi *et al.* (2011) pointed out that affect-sensitive robots are more able to engage with people, in order to maintain interaction with people and even to extend interaction time.

2.1.3 Application of Conversational HRI

HRI applications are widely applied in different industries, such as service industries, agriculture, education for assistance, collaboration, or entertainment (Sharkawy, 2021). As a starting point, a novel robot always attracts people's attention in public spaces, *e.g.*, a shopping centre. As a service robot, the robot can also be a tour guide in a museum (Duchetto *et al.*, 2019). In such an application, when a tour is requested, the robot can lead visitors on the tour and give a brief explanation of each part of the exhibition. Reflecting on such applications, some HRI researchers mention that the design of different emotions in the robot can enrich the educational experience and that visitors have a great experience in HRI in such as a case (Nourbakhsh *et al.*, 1999). Meanwhile, receptionist robots are applied in various areas, for example, the Pepper robot has been set up at HSBC in the United States to enhance customer engagement and to educate customers about product information *etc.*. When robots are used for learning, Omar Mubin *et al.* (2016) noted that social robots have shown particular benefits for learning and education. When designed as a peer, such that the robot has a level of knowledge similar to that of a learner, the learner and the robot spend their time acquiring new knowledge together (Bartneck *et al.*, 2020). Meanwhile, as a tutor for online learning (Doherty and Doherty, 2018a), a robot can take over specific tasks from

human teachers, for example, the tutor robot can offer personalised and one-to-one tutoring experiences (VanLEHN, 2011). At the same time, when robots are used for entertainment, as a "toy", robots can be used to model joint attention with autistic children in autism therapy and education (Robins *et al.*, 2004). Finally, turning to the case where robots are also used in healthcare and therapy, a healthcare assistant robot can help workers obtain health results on their screens promptly (Esterwood and Robert, 2020); More recently, as a COVID-19 test robot, robots have been shown to be useful in delivering medicines to patients (Esterwood and Robert, 2020) while robots also help to ensure pipetting safety before COVID-19 laboratory staff load samples taken from COVID-19 patients tested on a tray (Sharkawy, 2021).

2.2 Affective Computing in HRI

To study affective computing in HRI, the differences between mood, emotion, and affect in a human should first be clarified. Affect consists of emotion and mood, as a comprehensive term, representing the entire spectrum of emotionally charged responses from rapid and subconscious responses to external events and complex emotions (Bartneck *et al.*, 2020); while the term emotion is related to changes in internal feelings from external incentives and thoughts (Picard, 2003). Moreover, emotions can motivate and regulate people's behaviour and are an integral part of human social cognition and behaviours (Jarvela, 2011). Moods, however, are diffuse and internal, and are usually not attributable to a specific stimuli, and are typically low intensity, although moods can continue for a longer duration.

In human-human interaction, we rely on multimodal information more than unimodal to get a better understanding of each speaker's intention from their facial expressions or speaking *etc.* In general, this shows that combining aural and visual mediums conveys more information than they convey individually (Poria *et al.*, 2017). Thus, to build a natural and smooth conversation in either HRI or HCI, the collection of multimodal data on human social behaviours is a vital first step for researchers. For example, Ben Youssef *et al.* (2017) studied

HRI with a fully autonomous robot, the Pepper robot, to recognise user engagement in spontaneous conversations between users and the robot. However, in traditional HCI, virtual agents such as avatars or chatbots for affective computing study have been used for online learning on special platforms such as AutoTutor, ITS (Intelligent Tutoring Systems), MOOCs (Massive Open Online Courses), or other forms of serious games. However, unfortunately, these studies (HCI) usually lack multiple modalities, *e.g.*, touch, posture, or spatial interaction *etc.*, which are essential interactive parts of interaction between a physical robot and a human.

While capturing data from multiple modalities is very important, in practise, the realities of data collection with real operationalised systems can be very challenging. For this reason, the Wizard of Oz (WoZ) methodology in HRI is a frequently used as it is a flexible mechanism for researchers to easily manipulate the robot for their specific research purpose and data collection, which also has the advantage of being safer for participants (Riek, 2012). Nowadays, many researchers have designed WoZ HRI experiments to collect multimodal datasets for their specific studies, *e.g.*, emotion estimation (Jo *et al.*, 2020) or engagement recognition (Tapus *et al.*, 2012; Ben Youssef *et al.*, 2017).

2.2.1 Emotion Estimation

Emotion is a fundamental factor in human-human interaction (HHI), which affects people's attitudes and influences their decisions, actions, learning, communication, and situation awareness in human-centre environments (Poria *et al.*, 2017). If a person has a strong ability to observe others' emotions and manage their own emotions, they are likely to contribute more successfully to interaction with others (Poria *et al.*, 2017). Similarly, a social robot is arguably expected to have human-like capabilities for observing and subsequently predicting human emotions. Building on this idea, Spezialetti *et al.* (2020) identified three broad sets of tasks that are required to equip robots with emotional capabilities: (a) designing emotional states of robots in existing cognitive architectures or emotional models; (b) formulating

rich emotional expressions for robots through facial expression, gesture, voice, *etc.*; and (c) detecting and predicting human emotions. The first two areas are in robots-oriented research, while the last area is in human-centred research that is related to this work. However, Cohn (2007) indicated that human emotions cannot be observable because emotion is a cognition, a feeling, or a physiological and neuromuscular change. Therefore, he points out that emotion must be explained through an interaction context, a user survey, behaviours or physiological indicators (Tran *et al.*, 2020), such as EEG, heat flux, near-infrared spectroscopy (fNIRS), and facial electromyography, *etc.*.

Thus, to build emotion models, it is important to design a special experiment from a multimodal learning perspective. To do this, the facial expression is often the most natural emotional expression for a human being. The Facial Action Coding System (FACS) with AUs (facial action units) (Cohn, 2007; Menne and Lugrin, 2017) is a part-based method that is well-known in facial behaviours research for the analysis of facial expressions. Convolutional neural networks (CNNs) have also been shown to provide highly accurate results for analysing images in emotion recognition (Refat and Azlan, 2019). Similarly, various recurrent and ensemble network architectures have been built to analyse multimodal datasets including speech (audio) data, text-based data and video data and to estimate emotional states (Tripathi and Beigi, 2018; Hazarika *et al.*, 2018). Furthermore, to improve emotion recognition performance, the "Dialogue Emotion Correction Network (DECN)" (Lian *et al.*, 2021) was proposed as a new correction model for emotion recognition in a conversation. This network has two modules: an utterance-based emotion recognition engine and a conversation-based correction model. DECN applies the grated graph neural network (GGNN) as a structure for modelling self-influences and inter-speaker influences. The DECN model was trained on two popular benchmark datasets: IEMOCAP (Tripathi and Beigi, 2018) and the MELD datasets (Poria *et al.*, 2019). Additionally, head pose, eye tracking, and eye-gazing are typical nonverbal signals for people's emotional expression. In particular, Emery (2000) explained

that the eye gaze is a component of facial expressions that is used as a cue to demonstrate people's attention to another individual, an event, or an object. They have shown that eye gaze as a special cognitive stimuli is a "head-wired" approach in the human brain.

From a slightly different perspective, in advanced driving assistance systems (ADAS) research (Khan and Lee, 2019), tracking drivers' eyes and gazes is one of the most interesting topics in the road safety research community. Researchers set up cameras or devices (*e.g.*, wearable eye-tracking systems, google glasses) (Mavridis, 2015) in a specific environment, and then the devices are used to track the participants' eye gaze when participants perform the designed tasks (Zhang *et al.*, 2020). Head-pose estimation is also a form of face-related visual research and is inherently related to visual eye-gaze estimation. Murphy-Chutorian and Trivedi (2009) provided the example of the Wollaston illusion, where while eyes are in the same pose, the direction of eye gaze is decided by two differently orientated heads (see Figure 2.2). They also indicated that people with different head poses can reflect more emotional information such as dissent, confusion, consideration, and agreement. Meanwhile, methods for training models of eye-gaze and head-pose estimation are generally consistent with facial expression analysis.

Fig. 2.2 Wollaston illusion: different eye-gazing with head poses (Murphy-Chutorian and Trivedi, 2009)

2.2.2 Engagement Detection

Although challenging in itself, the complexity of engagement detection is small but more controlled in comparison to the more general challenges of the emotion estimation in social interactions. How the robot can explore the interacting context and environment so that users are motivated to continuously engage and communicate with the robot is a crucial issue in HRI study. The definition of engagement in social and cognitive psychology can be expressed through three aspects: a social connection, a mental state, and a motivated and captivated phenomena (Doherty and Doherty, 2018a; Sidner *et al.*, 2004; Jaimes *et al.*, 2011). For the social connection, engagement is a process in which participants start establishing a connection, try to maintain this connection, and eventually finish their connection (Doherty and Doherty, 2018a; Sidner *et al.*, 2004). With respect to the mental state, engagement can be behavioural while the engaged interaction is persistent; engagement can be emotional when the engaged interaction could be interesting, valuable, and valent; and engagement can also be cognitive when the interaction is motivated, effective, and strategic (Sidner *et al.*, 2004). Concerning the motivated and captivated phenomena, engagement may not apply to a single interaction but may instead measure a long-term relationship; this is particularly true where engagement is with a social platform, although interacting with robotic platforms across time is certainly the desire in the long run (Jaimes *et al.*, 2011). Various other more concrete definitions of engagement have been proposed in the literature. O'Brien and Toms (2008) for example defined *"Engagement is a category of user experience characterised by attributes of challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control."*, with Sidner *et al.* (2005) also defined engagement as *"the process by which individuals in an interaction start, maintain and end their perceived connection to one another"*.

The engagement detection methods in HRI and HCI have three basic methods, which are manual, automatic, and semi-automatic categories (See Figure 2.3) (Dewan *et al.*, 2018).

The manual category refers to self-tasks for participants, including self-reporting and observational checklists. Self-reporting entails a real-time feedback of the participant at each task during the engagement detection process, which is useful for researchers to administer and collect more information directly for engagement detection. The observational checklist is to collect participants' questionnaires at the end of processing of the experiment. However, the observational checklist has some limitations, as participants may not provide exact correct answers for the experiment, and it can take a long time for participants to complete this questionnaire. Semi-automatic methods utilise the timing and accuracy of responses, such as the reaction time at each interaction, the judgement of user responses, and the tracking of this experiment to estimate the engagement. The automatic category meanwhile indicates the computational observation of participants' behaviours and the extraction of features from computer vision-based analysis images, videos, and audios to observe facial expressions, gestures or postures, eye movement, and voice pitch tracking. In these automated methods, sensor data may also be extracted from physiological and neurological sensors such as heart rate, electrocardiogram (ECG), EEG, and blood pressure monitors. Moving away from the case of truly situated interaction to online learning systems, the engagement can also be automatically estimated by taking advantage of online activity metrics such as time spent on pages or in a learning process.

Studying and building engagement detection models requires careful attention to design. As an example study, Ben Youssef *et al.* (2017) studied the spontaneous conversation in HRI through the use of the Pepper robot in a public place for 54 days. All participants were students, teachers, researchers, or visitors who did not know the purpose of the study. Firstly, a participant is detected by the robot when the participant was oriented toward Pepper and the straight line distance from her/his was 1.5 metres. Following this an initial verification of criteria, questionnaire, and introductory dialogue took place. This was followed by a spontaneous dialogue session, which in turn was followed by a session in which the robot

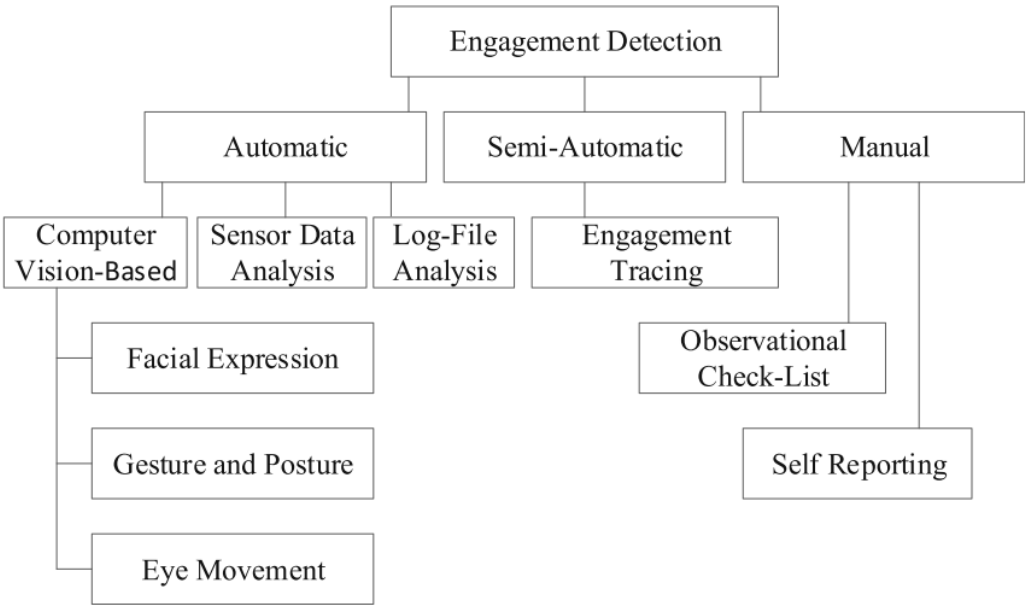


Fig. 2.3 Learners' Engagement Detection Methods (Dewan *et al.*, 2018)

outlined the difference between a cucumber and a human. While this study was highly designed, the full autonomous robot did interact with participants in public. Unfortunately, the interaction between these participants and the Pepper robot lacked social conversations and was more akin to a question-and-answer dialogue.

As another experimental example consider the four single-subject experiments about the social engagement between children with autism and the Nao robot which was presented by Tapus *et al.* (2012). The four children were boys with a range of age from 2 years and 8 months to 6 years. The experimental design was such that the Nao robot imitated each child's arms movements and each child's behaviours were collected by video data during an interaction with the robot or a human charged with the same imitation task. Employing the WoZ technique (Riek, 2012), two rooms were set up, one for the child and the robot, and one for the operator who controlled the robot's movements.

In the context of the experiment design, the data measurement is a key component for deciding and collecting those types of data that can be used for the engagement detection. Ben Youssef *et al.* (2017) presented self-reports, monitoring participants' responses, tracking

postures and facial expressions during the interaction, and recording behavioural responses as measurements for their user engagement experiment. Tran *et al.* (2020), meanwhile, illustrated the accuracy, the reaction time, perceived mental workload, and perceived communicative effectiveness as measures. The accuracy recorded the correct rate for whether participants clicked on the objects correctly from the robot's request. The reaction time was the time recorded when the participant interacted with virtual objects. The free initiation in children and the robot experiment (Tapus *et al.*, 2012) concerned the gross motor actions that the child performed without prompt while looking at the robot or human interaction. Gaze shifting then referred to the occasions spent moving gaze between the robot and the human. Finally, the smile or laughter was a measurement that reflects how children engaged in HRI or HHI (Human-Human Interaction).

Data processing and statistical analysis are necessary to understand the data sets, *e.g.*, video data, sensor data including physiological and neurological sensors (Doherty and Doherty, 2018b), or log files that includes self-reporting, questionnaires, or basic user information (Ben Youssef *et al.*, 2017). Tapus *et al.* (2012) analysed the frequency of different target behaviours (*e.g.*, gaze of the eye, children's smiling/laughter, etc.) following the timeline of the experiment. They calculated a statistical test using Mann-Whitney for data variables to compare the results of all phases (e.g. the Nao-interaction, the human-interaction) (Tapus *et al.*, 2012). Ben Youssef *et al.* (2017) also analysed the average distance between the user and the robot, the average variance in head angles, or an average direction of the eye gaze as a statistic to detect the breakdown of the engagement.

In the context of engagement detection modelling, Ben Youssef *et al.* (2019) employed recurrent and deep neural networks to predict SED (Sign of Engagement Decrease) using their annotated multimodal dataset. Long-short-term memory (LSTM) and gated recurrent unit (GUR) were used for modelling temporal sequences, and in particular, re-establishing a baseline against a more basic logistic regression style analysis. Perhaps unsurprisingly, their

results showed that deep learning techniques are better than the traditional machine learning technique (*e.g.*, Logistic Regression) (see Figure 2.4).

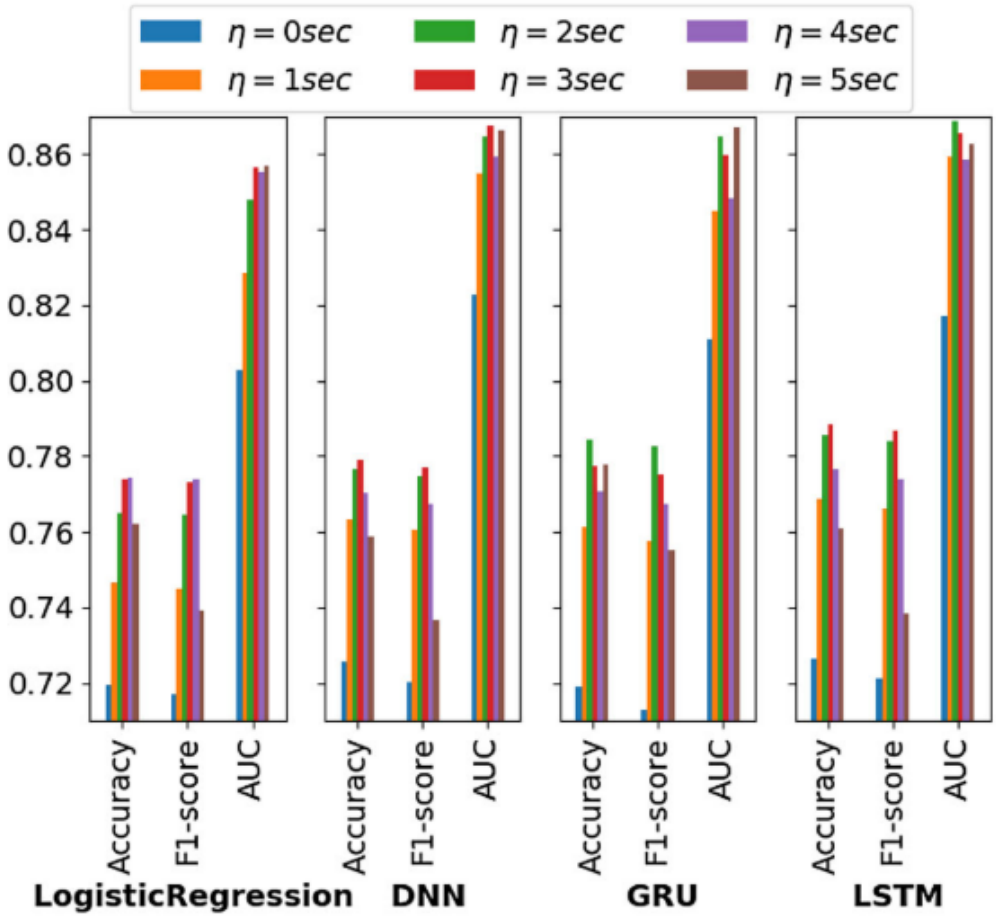


Fig. 2.4 Performance comparison between classifiers (Ben Youssef *et al.*, 2019)

2.2.3 Confusion Detection

Confusion unfortunately is not a term that is always simple to define. Confusion may be described as a bonafide emotion, a knowledge emotion, an epistemic emotion, an affective state, or a mere cognitive state (D'Mello and Graesser, 2014). When confusion is considered an effective response, confusion occurs in people who are enthusiastic to know or understand something. When confusion is defined as an epistemic emotion (Lodge *et al.*, 2018), it is associated with blockages or impasses in the learning process while trying to learn

something new or trying to clarify problems or issues. Confusion is also triggered by cognitive disequilibrium. Yang *et al.* (2015) explained that cognitive disequilibrium is defined as the state in which a participant learns when obstacles to the normal flow of the learning process are encountered, such that the participant may feel confused when encountering contradictory information leading to uncertainties and results in cognitive disequilibrium.

There have been several partial or full formalisation of confusion states in the literature. Arguel and Lane (2015) presented two thresholds (T_a and T_b) of levels of confusion in learning (see Figure 2.5). When the level of confusion is over T_b, it is said that confusion is persistent. At this stage, students might be frustrated or even bored. If the level of confusion is less than T_a, then the learners should be fully engaged (or have the potential to fully engage) in their learning. Between these two thresholds (T_a and T_b) is the confusion stage. In this range, the confusion state is such that learning may actually be encouraged.

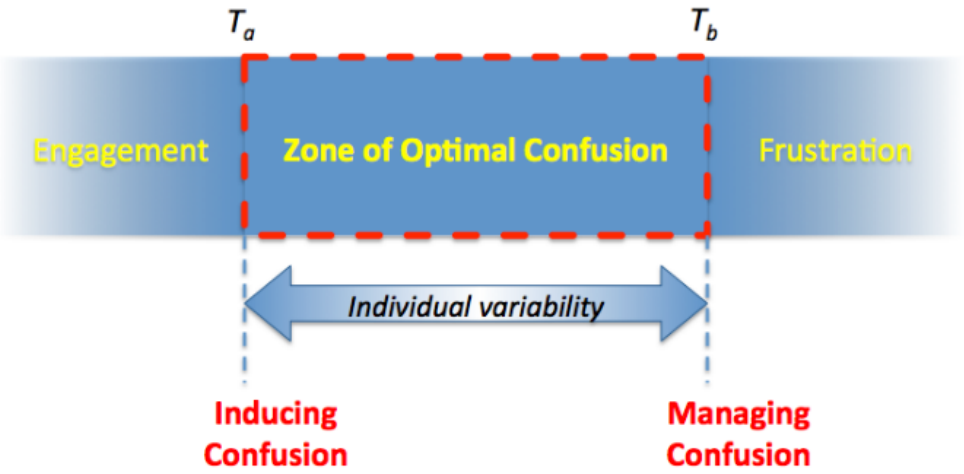


Fig. 2.5 The boundaries of the zone of optimal confusion (Arguel and Lane, 2015)

Lodge *et al.* (2018) meanwhile designed a learning event in which the learner was in cognitive disequilibrium, the disequilibrium being created by an impasse in the learning process (see Figure 2.6). in this model, when learners are in the zone of optimal confusion

(ZOC) which is productive confusion, they are engaged to overcome a confusing state; and as a result, the disequilibrium may be effectively resolved. However, if the confusion is persistent, such that learners cannot resolve the disequilibrium, then they may be in a zone of sub-optimal confusion (ZOSOC). At this stage, confusion becomes unproductive, leading to possible frustration or boredom. Finally, learners may lose engagement altogether and end the learning process. The state of confusion can also be said to be a part of emotional transitions between three emotions (engagement/flow, frustration, and boredom), and the concept is similar to ZOC.

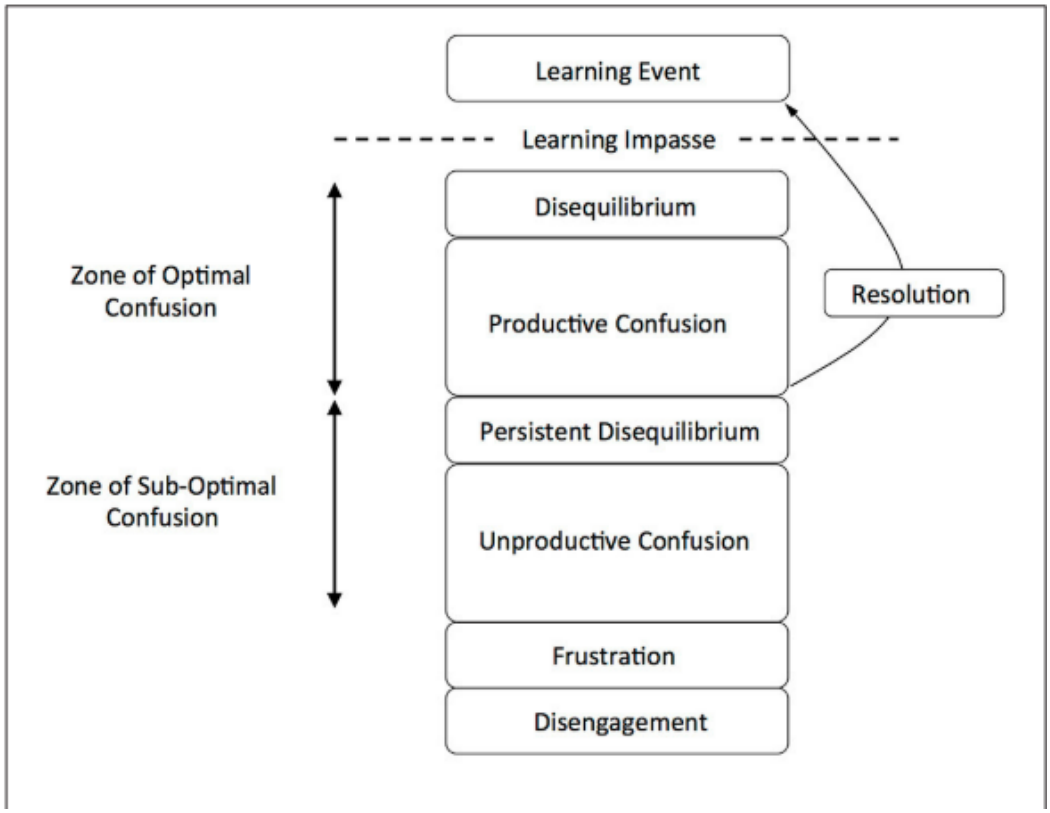


Fig. 2.6 Conceptual framework of ZOC and sub-optimal confusion (Lodge *et al.*, 2018)

Moreover, D'Mello *et al.* (2014) presented three bi-directional transitions to model confusion dynamics, *i.e.*, confusion-engagement, confusion-frustration and frustration-boredom transitions (see Figure 2.7). The confusion-engagement transition indicates an impasse that has been detected, then the user's state from engagement change to confusion and if the

user successfully resolved the confusion, then the user state can transition from confusion to engagement; The confusion-frustration transition likely occurs when a learner cannot resolve an impasse after trying different solutions, thus the learner's state transitions to frustration from confusion. Meanwhile, if the learner continually experiences another impasse(s), the learner's state would transition back to confusion. If the failure continues to persist, the transition involves frustration and boredom and can lead to disengaging (frustration to boredom). However, if the learner has to be forced to persist in their tasks, then the transition will change in this case from boredom to frustration. D'Mello *et al.* (2014) also noted that it is possible to transition to anxiety and hopelessness.

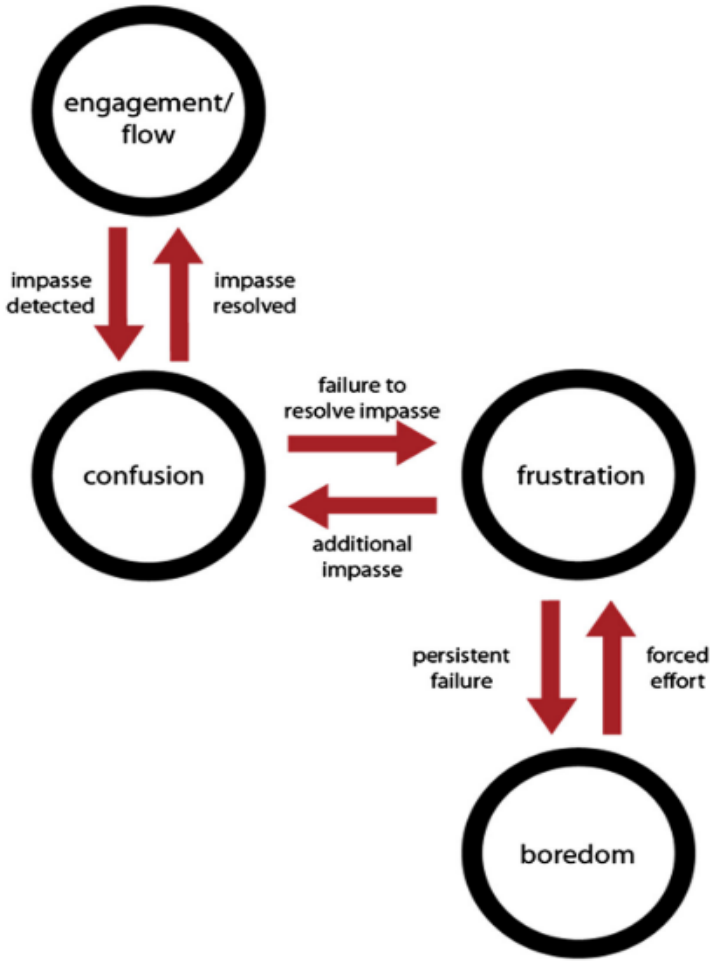


Fig. 2.7 Observed Emotion transition (D'Mello *et al.*, 2014)

Four patterns of confusion induction and non-confusion induction as strategies for the confusion stimuli have been proposed (Lehman *et al.*, 2012; Silvia, 2010). The first such pattern is complex information and simple information. Lehman *et al.* (2012) explained that complex learning is an experience full of emotions that occurs when learners are exposed to complex material, difficult issues, or indecisive decisions, such that their confusion may be triggered between positive and negative emotions (Arguel *et al.*, 2017). The second pattern is contradictory information and consistent information; here people may enter into a state of uncertainty and confusion when they are exposed to contradictory information (Lehman *et al.*, 2013). The third pattern of confusion is insufficient and sufficient information; here people do not receive enough information to respond to an interlocutor, as a result, they may get confused (Silvia, 2010). The final pattern of confusion is based on feedback. Lehman *et al.* (2012) designed a feedback matrix of feedback states to investigate feedback types and the confusion. This matrix essentially distinguishes between correct feedback which comprises correct-positive conditions and incorrect-negative conditions, and false feedback including correct-negative and incorrect-positive conditions. From their experiment, it was witnessed that the presentation of correct-negative feedback, *i.e.*, when learners responded correctly but got inaccurate or negative feedback, was an effective manipulation to stimulate confusion.

Exploring the best algorithm for confusion detection is a challenging task. Within the area of online learning, there have been a number of studies to this end, but there have been far fewer in HRI or related interaction fields. Ibrahim *et al.* (2021) proposed that the detection of confusion requires an artificial intelligence methodology. Traditional machine learning classification algorithms, *e.g.*, Naïve Bayes, Multi-Layer Perceptron (MLP), Feedforward neural network Decision Trees, or Random Forest Algorithm (Samani and Goyal, 2021a; Kavita Kelkar, 2021) can be trained on multiple e-learning datasets to detect the levels of confusion in real-time or offline. A comparison of performance in predicting levels of

confusion has shown that the MLP and Decision Tree algorithms perform better than Naïve Bayes for the estimation of the levels of confusion for quiz data submitted by students online on a MOOC (Samani and Goyal, 2021a). Meanwhile, the result of training the confusion detection model with the Random Forest algorithm achieved above 90% accuracy by capturing students' interacting behaviours (Kavita Kelkar, 2021).

In a similar style, Geller *et al.* (2021) compared the performance results of six classifiers training on data from online course forums. A pre-trained language model based on the bidirectional encoder representation from transformers (BERT) outperformed traditional machine learning for classifying confusion. Interesting, in this work, students posted hashtags that include confusion that reflected their affective states as a new label, which was better than manual labelling that is time-consuming and costly. Finally, it should be noted that fuzzy logic is a method that has also been used to detect the levels of confusion in clickstream data when learners answer online quiz types assessments (Samani and Goyal, 2021b).

Furthermore, EEG data as input and the facial expression data as output that was collected from users who played 3D games, and used to train Support Vector Machine (SVM), K-Nearest Neighbors algorithm (KNN), or Long Short-Term Memory (LSTM) based models for four levels of confusion detection (Benlamine and Frasson, 2021). In this study, the KNN and LSTM algorithms achieved the best accuracy. Turing instead to driver detection in a safe driving study, sensor data was collected and used for training a driver confusion states prediction model using typical neural network methods (*i.e.*, the logistic regression, feedforward neural networks, and recurrent neural networks (RNNs and LSTM RNN). In the case of this study, LSTM also outperformed the other models (Hori *et al.*, 2016). Moreover, we must also mention that the deep reinforcement learning (RL) has been also applied in HCI and HRI (see details Appendix C)

2.3 Embodied Interactions: From HCI to HRI

Due to the relative ubiquity of computer-mediated communication across different application domains, *e.g.*, online learning system (Doherty and Doherty, 2018a), healthcare assistants (Esterwood and Robert, 2020), virtual reality (VR) games (Lukosch *et al.*, 2019), and social VR (Baker *et al.*, 2021)), the expectations for multimodal interactive systems have grown in diversity and sophistication in the last decade. This is true for virtual online agents, but also extends to expectations for interaction with physical, or more precisely social, robots (Doherty and Doherty, 2018a; Pustejovsky and Krishnaswamy, 2021). A key trend in the development of communicative systems has been an assumption of multi-modality, *i.e.*, that our artificial interlocutors should have access to multiple modalities. However, the research community is well aware that the multimodal communicative skills of even state-of-the-art systems are still very limited.

Whether our interaction partner is a social robot, a 3D avatar, or even just a chat window, it is assumed that our interaction partners share similar conversational skills and abilities across these embodiments. This has benefits in terms of acclimatisation of technology across interaction partner types, but can also lead to frustration and disappointment when such alignment is not present in practise. This though is not just true in terms of users expecting systems to behave in uniform ways across device types, but may also be present in the expectations that systems – and their designers – make in terms of the behaviour and reaction of users to systems across different embodiment types.

This potential for mismatched expectations is, in some cases, exacerbated by the needs of researchers and industrial developers. Collecting real-world data in HRI studies for the investigation of particular phenomena is extremely challenging. Experimental hardware systems suffer malfunctions, recruiting participants or users is challenging and often expensive, and even finding appropriate real-world spaces to perform tasks can be difficult. These

problems were significantly heightened during the COVID-19 pandemic when it became in many ways unfeasible to perform human-robot interaction studies. For reasons such as these, numerous researchers have over the last four decades frequently turned to human-computer interaction studies, and particularly the use of avatars and chat systems, to conduct studies in the hope of bootstrapping studies of HRI. Invariably, these efforts have been based on the assumption that such data is as ecologically valid in one embodiment type as in another so long as the same basic interaction modalities are being used, *e.g.*, speech and cameras. While this assumption may have been true at one point, due to the relative novelty of all interactive systems interfaces, the ubiquitous nature of avatars and basic conversational systems in contrast with everyday social robotics has arguably laid waste to this assumption.

HCI has a strong relationship with HRI, as they are both studies of user interaction with computing systems although applied in different manifestations (Wei, 2016). As a field of study, HCI with computer-based technologies makes more contributions and insights into understanding context communication in order to improve interactions with users. On the other hand, HCI necessities can be part of a robotic system, and inspire HRI techniques such as an artificial conversation engine. Ultimately, combining physical characteristics, including autonomy, physical proximity, and decision-making capabilities from HCI makes HRI a distinct area of study (Singh *et al.*, 2021).

While HCI covers a vast number of physical system types as well as different goals of interaction, we are particularly interested in situated interaction where a user communicates with an embodied agent, which is typically a physical robot, but can also be embodied virtually in our study. In recent years, avatars have begun to become a prominent mechanism in virtual intelligent environments (Pan and Steed, 2016). Compared to other means of interaction, the avatar is presumed to be a more natural communication mechanism that can evoke strong agent-as-partner-style interactions through the use of human-like facial features and expression (Heyselaar *et al.*, 2017), vivid body language, and even specific personalities.

Moreover, the avatar has remarkable benefits over a speech-to-text-only interaction (Heyselaar *et al.*, 2017).

Multiple studies have commented on the relative properties of communication with physical robots versus other types of agents. Generally, it has been observed that people have more interactions with physical robots than with virtual agents or telecommunication agents in a number of different application areas (Wainer *et al.*, 2006; Lee *et al.*, 2006). Meanwhile, McNamara and Kirakowski (2006) revealed that customers or users can be affected to varying degrees in their overall user experience, due to the perception of different levels of social presence across both HCI and HRI. In the study of social presence by Herath *et al.* (2020)'s, the authors approached both HCI and HRI experiments with the same conversational engine but with a keyboard and monitor used for the HCI studies, and the Nao robot used for the HRI studies. In a post-questionnaire, in particular, the "UTAUT" (Unified Theory of Acceptance and Use of Technology Questionnaire (Heerink *et al.*, 2010)), it was shown that HRI trended more strongly with measures of animacy and likeability than with HCI, while on the measure of usefulness and trust, the experience of HCI was rated higher than in the HRI case. The authors believed that the HCI performance is better than the HRI performance in specific tasks or domains, but that the HRI performance was better than the HCI study for the exploratory and open-ended conversation domains. Also of note, in a cooperative HRI or HCI task, participants were found to be more engaged and enjoyed playing with a physically embodied robot in comparison to playing with a virtual embodied animated avatar as the physical robot was viewed as being informative and credible (Kidd and Breazeal, 2004; Hoffmann and Krämer, 2013). In contrast, Kidd and Breazeal (2004) also found that, such as verbal and role-playing tasks, there was no significant difference in attitudes between users who interact directly with a robot and those who play with the robot via video-displayed remotely in different rooms.

2.4 Summary

In this literature, I have systematically introduced the three mian components that are closely relevant to my study: conversational HRI, affective computing in HRI, and the different embodied interactions between HCI and HRI in terms of the diversity of aspects and applications. While there has been useful research in the field, none of the above research focuses on detecting interlocutor confusion in the multimodal situated dialogue of both HRI and HAI. While it is clear that smooth communication and task performance between a user and a system will require a level of situated awareness of confusion, the achievements to date in this field have been limited; this brings us to the main research question to be addressed in this work.

Chapter 3

Research Questions

As described in the previous chapter, to date a small level of research on confusion detection and modelling has been conducted; however, most of it is in the field of online learning with different human-computer interaction platforms. Particularly in terms of task-oriented dialogue between human and robot interaction, very little research concerning confusion detection or modelling has been published.

Given the need for systematic modelling and mitigation of confusion in task-oriented interactions, I consider the lack of systematic research on this topic in the HRI field to date to be a serious limitation. With this in mind, in this research work, the research question that I am aiming to address is as follows:

> How can we detect, model and mitigate user confusion states in situated Human-Robot Interaction?

This, of course, is a brief description and addressing it in its total form would, unfortunately, be beyond the scope of a single PhD dissertation. However, it should be possible to make some significant progress on this question. As such, the following subquestions have been identified to allow a systematic study of the issues.

- **RQ1**: How can we define what interlocutor confusion is in the context of situated spoken Human-Robot Interaction?

  There are many definitions of confusion in different domains (D'Mello and Graesser, 2014; Lodge *et al.*, 2018; Yang *et al.*, 2015) in the field of online learning, but none in the HRI area. Thus, it is necessary to clearly define what I mean by "confusion" in conversational HRI. This will be a guide that helps me frame my study scope and design experiments.

- **RQ2**: Are participants aware that they are confused if we give them a specific confusing situation?

  As I will outline later, it is difficult to clearly know specifically when participants are confused or when other emotions might instead manifest during a conversation. Therefore, looking at the relationship between assumed confusion states and participants' self-estimation of their confused state is very useful. Addressing this question also helps me to learn different nonverbal and verbal human behaviours in confusion states and to verify whether participants are successfully stimulated during more subtle interactions.

- **RQ3**: Do participants express different physical or verbal/nonverbal behaviours when they are confused that we can detect?

  Although it is quite likely that subtleties in physical expression that could manifest confusion in HRI are possible, it is equally likely that different people might have slightly different behavioural or physical responses when confused. Therefore, these differences between participants across confused and nonconfused states should be monitored and modelled wherever possible.

- **RQ4**: Are there differences between embodiments when it comes to confusion expression and detection?

HRI studies are expensive and difficult to control when compared to HCI studies, and researchers sometimes have to turn to HCI studies, which are easier to conduct and collect data faster than is the case for true robotics studies. Thus, if I rely on human-avatar studies as a proxy for human-robot studies, I must ask whether the avatar interaction can substitute for robot interaction studies or whether the two forms of interaction are, in practise, too distinct to be useful.

- **RQ5**: How can we effectively detect the different states of confusion including non-confusion, productive confusion, and unproductive confusion using multimodal data?

  Ultimately, the goal of studying confusion in HRI should be to allow us to effectively detect confusion and mitigate that confusion in interaction. In HRI studies, we typically have access to a range of data types, and some of these are already well tuned to detect confusion. However, in practise, this detection is likely to be non-trivial when taking into account real-world circumstances and individual differences. Nevertheless, detecting different confusion states can help us to learn the exact confusion state in HRI dialogues in order to adjust the accuracy of policies to mitigate factors of confusion.

- **RQ6**: How can we use confusion state detection to help the participant overcome their confusion in conversational HRI?

  Confusion can happen any time whenever we communicate with a robot in a dynamic environment in real-time. However, how the system might deal with a confused state is not a trivial question. We must consider what policies would be most useful in addressing the confused state. Ideally, these policy decisions should be suited for integration into an Artificial Intelligence (AI) planning policy for the automated HRI.

To answer these research questions, I first present a definition of confusion and specifically three states of confusion. Following this, WoZ experiments are applied for HAI and HRI studies. The first study is an online situated conversation between an avatar and a participant

using a developed online chat web application. This study involved a series of situated conversations that aim to trigger confused states in participants. Meanwhile, participants' performances including verbal or nonverbal responses and facial expressions are analysed. The second study applies similar stimuli methods from the first study but uses the Pepper robot, where participants must physically attend a lab and have a situated face-to-face conversation with a robot.

Furthermore, these studies allow for a comparison of HRI and HAI to be considered to learn from different perspectives on my HAI and HRI studies. The confusion states detection model can be trained using deep learning techniques on multimodal interactive data from my HRI experiments. Finally, the pre-trained model will be generalised and evaluated in real-time conversational HRI to examine whether my human-robot interaction system can detect exactly when a participant is confused.

Conversational HRI is a broad topic of research. There are many methods for the design of HRI experiments that I have found. Then, to frame the scope of my study, there are limitations to my research study that need to be clarified. First, the dialogues that I design are semi-automatic and task-oriented spoken dialogues. These are targeted at my research purpose so I hope to precisely stimulate different states of confusion through these experiments. Second, it can happen that technical issues on the devices (avatars and robots) lead the participant to confusion or other states and there are outside the scope of my study. Third, one-to-one and face-to-face conversations between a robot and a participant will only be designed for this study. Thus, there will be no multi-party interactions considered. Fourth, for the experimental systems, these are not fully autonomous. Instead, a researcher controls those systems (*i.e.*, avatar and robot) for all interaction with participants. Finally, biological data, such as EEG, body temperatures *etc.*, will not be considered, but it is notable that they are likely to be useful indicators of mental state.

Chapter 4

Work Done To Date

In this section, I set out the work that has been done to date to address the research questions raised in the proceeding section. I begin with a brief summary of the perspective I take on the definition of confusion as this directly influenced subsequent study design.

4.1 Defining Confusion

Generally speaking, confusion as a psychological state has been defined in different studies; mostly to date within the context of pedagogy and related applied fields of learning. In terms of bonafide emotion to an epistemological state, confusion can be considered an effective response that occurs in people who are enthusiastic to know or understand something (D'Mello and Graesser, 2014). In contrast, confusion can be defined as an epistemic emotion, that is, learners have impasses or blockages during the learning process. In addition, many studies have shown that confusion can transition between the engagement state and the frustration state under certain conditions. I noticed that there was no well-established definition of confusion that can assist my further studies in modelling and mitigating confusion in interaction. In light of this and for use in the context of dialogue-centric human-machine interaction, I offer the following working definition of confusion:

Confusion is a mental state where under certain circumstances, a human experiences obstacles in the flow of interaction. A series of behaviour responses (which may be nonverbal, verbal, and, or non-linguistic vocal expression) may be triggered, and the human who is confused will typically want to solve the state of cognitive disequilibrium in a reasonable duration. However, if the confusion state is maintained for longer periods, the interlocutor may become frustrated or even drop out of the ongoing interaction.

This definition of confusion explains, in general interaction scenarios, that confusion can be inducted from obstacles, as a mental state, confusion can be kept for a certain duration, and transition to other states depending on whether this confusion state can be solved. This definition also mentions that people's behaviours including nonverbal/verbal behaviours in the state of confusion may be stimulated to represent their confusion states. This definition of confusion, as a guideline, drives the rest of my studies on detecting confusion in situated conversational interaction, starting with confusion detection in human-avatar interaction.

4.2 HAI Study: Detecting Interlocutor Confusion in Situated Human-Avatar Interaction

To validate my study designs for research purposes, I conducted a series of initial pilot studies with small groups; it allowed us to improve the design of my scenario and its applicability to my ultimate conversational HRI goals. The first two pilot studies have been completed to cover the first three research questions, namely: RQ1, RQ2, and RQ3, which were mentioned in Chapter 3.

According to the definition of confusion introduced in the previous section, I designed a WoZ study to explore: the effectiveness of confusion induction methods in HAI interactions; as well as the relative performance of a series of manual; semi-automatic and automatic methods to estimate confusion from the data collected.

4.2.1 Study Design

This pilot human-avatar study was designed to account for some challenges introduced by the COVID-19 pandemic of 2020-2021. Also, while it was a pilot study and the avatar did not provide a fully situated experiment, the avatar style interaction has been shown to have remarkable benefits in speech- and text-based interaction (Heyselaar *et al.*, 2017).

This study was based on a semi-spontaneous one-to-one conversation between a wizard-controlled avatar and a human participant. Those participants were recruited from universities and study programs around the world. They remained in their own locations, and the wizard was also located in their own work environment. Meanwhile, all participants were requested to use their own laptop with a connected camera and audio with a stable internet connection. Figure 4.1 shows the detailed experiment process for this study. The experiment time for each participant was less than 15 minutes in total, including 5 minutes for the central conversational task. At the beginning of the experiment, participants received instructions on this experiment, such as how to use a real-time HAI chat web application (described later), how to register, sign up and enter the chat room to meet the avatar, and for the purpose of the study, as well as for user consent. After the conversation task between the avatar and a participant, the participant was required to complete a survey. At the end of the experiment, a 3-minute interview was used to collect feedback from the participants.

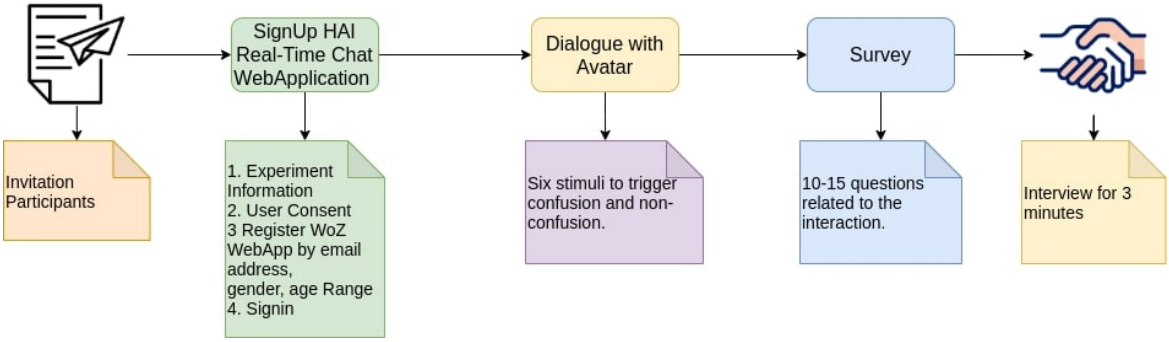


Fig. 4.1 HAI Experiment Process for Confusion Detection

The web application framework was developed and built with two main components: a real-time chat application [1] and an avatar application [2] that were embedded in a real-time communication application. The avatar application was based on the framework developed by Sloan *et al.* (2020), which provided a sandbox with modules of an e-learning platform with an animated avatar. This avatar integrates animation, speech recognition and synthesis, along with full control of: (a) the avatar's facial expressions to express happiness, sadness, surprise, *etc.*; (b) voices including pitch, speed, emotions, emphasis, and accents; and (c) body pose including leans, rotation, tilt, and blink The real-time chat application [3] meanwhile was a web application for online interaction between an agent/avatar and a participant that I developed to execute the entire experiment process, including user enrolment, communication, all survey presentations, as well as the user consent steps. Furthermore, this application was implemented such to enable full data recording of both the avatar and the participant's audio, text, and camera stream. Figure 4.2 depicts the complete framework. As for the technical architecture of this web application, the front-end was developed with ReactJS which is an open-source JavaScript library for UI (User Interface) components; NodeJS and Socket.IO which are back-end API services for a web server that allow data transmission between the front-end and database; WebRTC (Web Real-Time Communication) which is a technology to capture video media without requiring an intermediary, and which presented a simple mechanism to implement Real-Time online communication. Finally, I chose MongoDB, which is a document-oriented database, to store user registration information. To publish the web application, the back-end was deployed on the Heroku service platform, and the front-end was deployed on a cloud platform.

There were 23 participants from six countries who participated in this study. Three of the participants were unable to complete this study due to Internet connectivity or equipment problems. All participants were over 18 years of age and were at least capable of having

[1] https://rt-webchatapp-v5.netlify.app
[2] https://avatarv2.herokuapp.com
[3] https://github.com/lindalibjchn/WoZ_WebChat

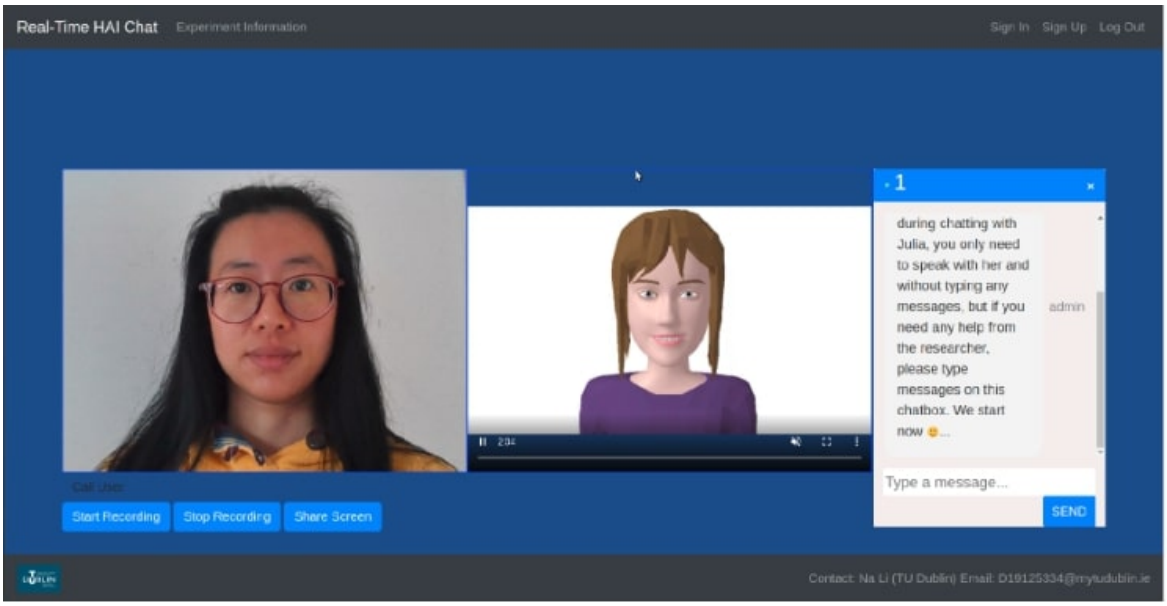


Fig. 4.2 HAI Real-Time Chat Web Application

a simple conversation in English. Ultimately, there were video data, user surveys, and demographic information from 19 participants (8 males, 11 females), and I acquired their permission to use their data for this study.

4.2.2 Dialogue Design

I aimed to stimulate confusion and non-confusion during a short conversation. Here, I defined two conditions with appropriate stimuli. In condition A, the stimuli were designed to trigger confusion in the participants; in condition B, the stimuli were designed so that the participants should straightforwardly complete a similar task without entering confusion states. Three tasks were executed by each participant. Task 1 was a simple logical problem; task 2 was a word problem; while task 3 was a math question.

As for the designation of situated dialogues, there were three patterns of confusion for the two conditions (see Table 4.1): the first pattern was complex information and simple information, the second pattern was insufficient information and simple information, and the third pattern was of correct-negative feedback and correct-positive feedback. What follows

is a sample of a word problem question, for the insufficient information case in condition A: *"There are 66 people in the playground including 28 girls, boys and teachers. How many teachers were there in total?"*, while for the sufficient information case in condition B we have: *"There are 5 groups of 4 students, how many students are there in the class?"* (for all dialogue designs, see Table A.1). In practise, I prepared two questions (also as two dialogues) for each condition in each task. To balance the number of conditions between participants for data analysis, the sequence of the experiment with all conditions shown in Table 4.2 was used. For the first participant, the sequence of conditions was Task 1 with condition A, Task 2 with condition B, and Task 3 with condition A (called: ABA). Then, the second participant's sequence of conditions was BAB, *i.e.*, Task 1 with condition B, Task 2 with condition A, and Task 3 with condition B.

Table 4.1 A matrix of tasks and causes of confusion is divided by conditions.

Condition A	**Condition B**	**Tasks ***
Complex information	Simple information	Task 1, Task 2, Task 3
Contradictory information	Consistency information	Task 1, Task 2, Task 3
Insufficient information	Sufficient Information	Task 1, Task 2, Task 3
Correct-negative feedback	Correct-positive feedback	Task 1, Task 2, Task 3

* Task 1: Logic problem; Task 2: Word problem; Task 3 Math question

Table 4.2 Example of the experiment sequence for sample participants

Participant 1		
Stimulus	**Task**	**Condition**
1st	Task 1	A
2nd	Task 2	B
3rd	Task 3	A
Participant 2		
Stimulus	**Task**	**Condition**
1st	Task 1	B
2nd	Task 2	A
3rd	Task 3	B

Additionally, I also considered that the individual stimuli should include verbal and non-verbal features of interactions. It should be noted though that avatar responses needed to be mapped to visible behaviours (Cassell and Vilhjálmsson, 2004). Figure 4.3 shows an example of the mappings of avatar facial expressions and body gestures for the conversational responses and conversational behaviours that reflect positive reaction and negative reaction.

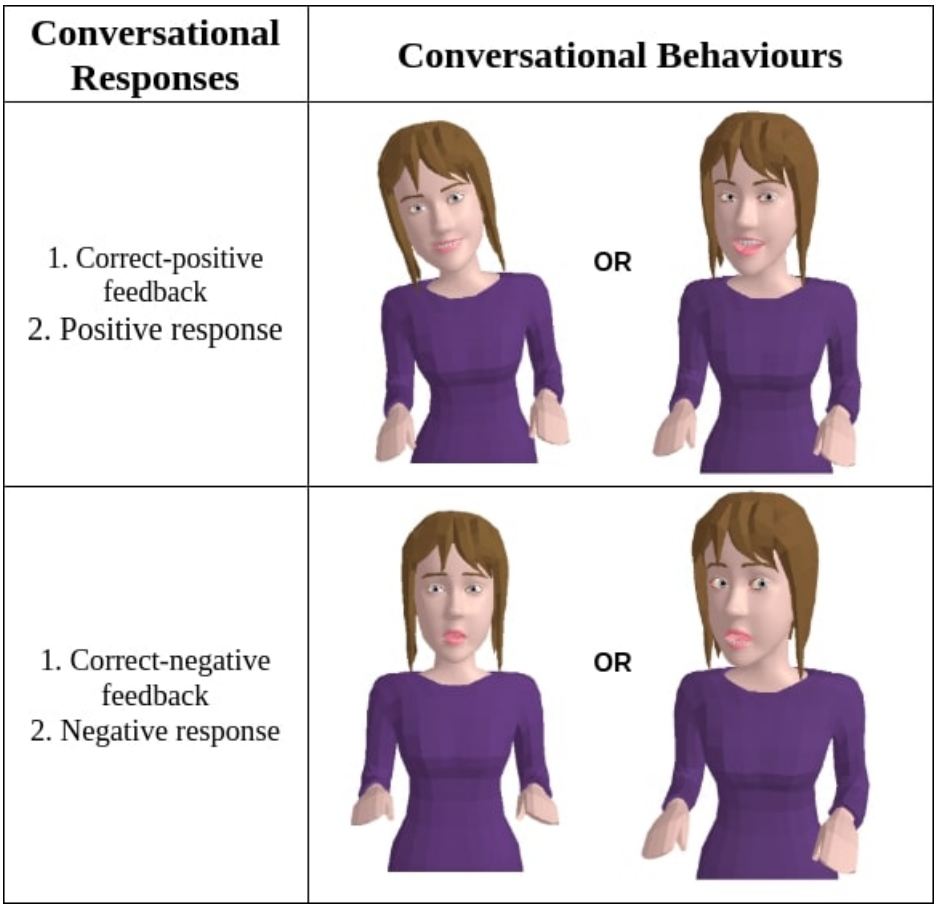


Fig. 4.3 The mapping of the reaction status and visible traits for the avatar

4.2.3 Data Preparation

Frame data was extracted from 19 participants' videos, and each video was labelled for one of the sequences of conditions (*e.g.*, ABA or BAB). Thus, each frame was labelled as either condition A or condition B. Ultimately, I collected 4084 frames for condition A as well as

3273 frames for condition B. As I focus on facial frames, facial recognition and alignment were necessary before analysis of the data was possible. I applied a Multitask Cascaded Convolutional Neural Networks (MTCNN)-based face detection algorithm to detect the face without frame margins (Savchenko, 2021) and centre cropped a region of 224×224 for all frames. For illustration, in Figure 4.4, a comparison is shown between the original frame on the left and the aligned face image on the right.

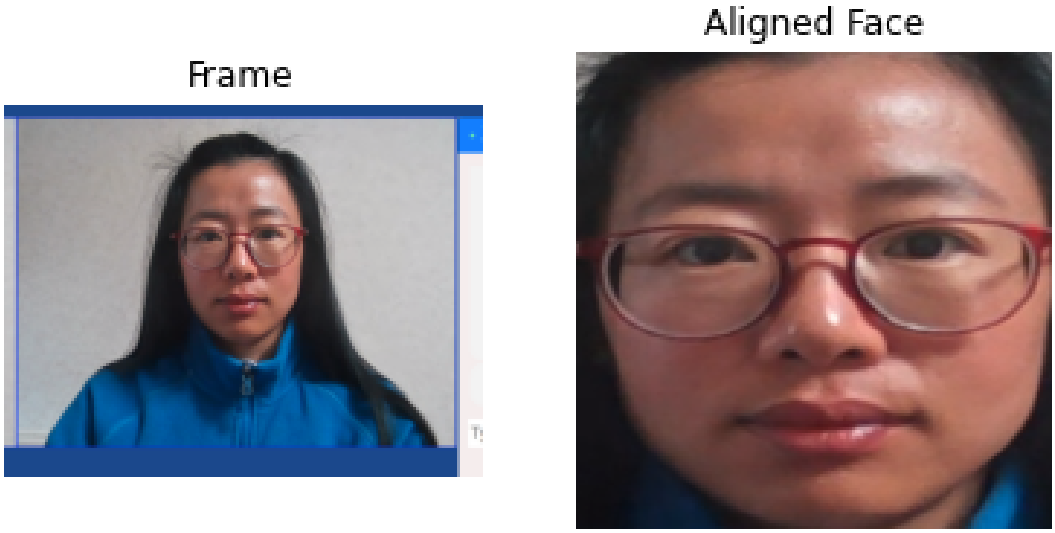


Fig. 4.4 The mapping of response status and behaviours of participant B

In addition, for the user study I designed 10 post-interaction questions using a 5-level Likert scale. There was one question to be passed to each task (logical questions, word problems, and math questions) (see Table B.1).

4.2.4 Data Analysis

Frame Data Measurement My primary data analysis focused on the automated processing of video data with three analysis aspects, which were emotion detection, head position estimation, and eye gaze estimation; based on this I then analysed whether there was a significant correlation with the confusion state. Firstly, for emotion detection, I used a visual emotion detection algorithm based on a Mobile Net architecture as a backbone face

recognition network (Savchenko, 2021; Howard *et al.*, 2017), which was trained on the AffectNet dataset (Mollahosseini *et al.*, 2017) for 8 target classes *i.e.*, 8 facial expressions, which are: the 7 primary emotions: neutral, happy, sad, surprise, fear, anger, disgust, and an 8th: contempt. The result shows the 7 primary emotion classes grouped by the two conditions (see Table 4.3). For condition A, the predicted results of the negative emotions classes (anger, disgust, fear, and sadness) were higher than the predicted results for the positive emotions classes (happiness and surprise). On the contrary, the predicted results for the positive emotions class in condition B were greater than for condition A. As for the neutral emotion, the results for condition A were more than for condition B. Meanwhile, Figure 4.5 illustrates a comparison of the results for the prediction of emotions for the three categories (negative, positive and neutral) grouped by conditions.

Furthermore, I investigated the correlation relationships between the three categories of emotions and the conditions with statistical analysis. The result of an independent-sample t-test was that there was a significant difference in the three emotion categories (negative, positive, and neutral) with the two conditions, ($M = 0.77, SD = 0.94$ for condition A, $M = 0.48, SD = 0.60$ for condition B), $t(715) = 5.05, \rho - value < 0.05$.

Table 4.3 Result of emotion estimation by condition A and condition B

Condition	**Anger**	**Disgust**	**Fear**	**Sadness**	**Happiness**	**Surprise**	**Neutral**	**Overall**
A	262	282	136	677	702	65	1799	3923
B	77	165	57	480	858	95	1502	3234

Regarding head-pose estimation, I applied a CNN model with dropout and adaptive gradient methods (Patacchiola and Cangelosi, 2017), which was trained on three datasets: Prima head-pose dataset (Gourier *et al.*, 2004), the Annotated Facial Landmarks in the Wild (AFLW) dataset (Köstinger *et al.*, 2011), and the Annotated Face in the Wild (AFW) dataset (Zhu and Ramanan, 2012). The predicted results combined the angles of pitch, yaw, and roll for each frame. I note that the values of angles were positive and negative numbers as a person has different angles of direction. To mitigate the case that the sum of these angles was

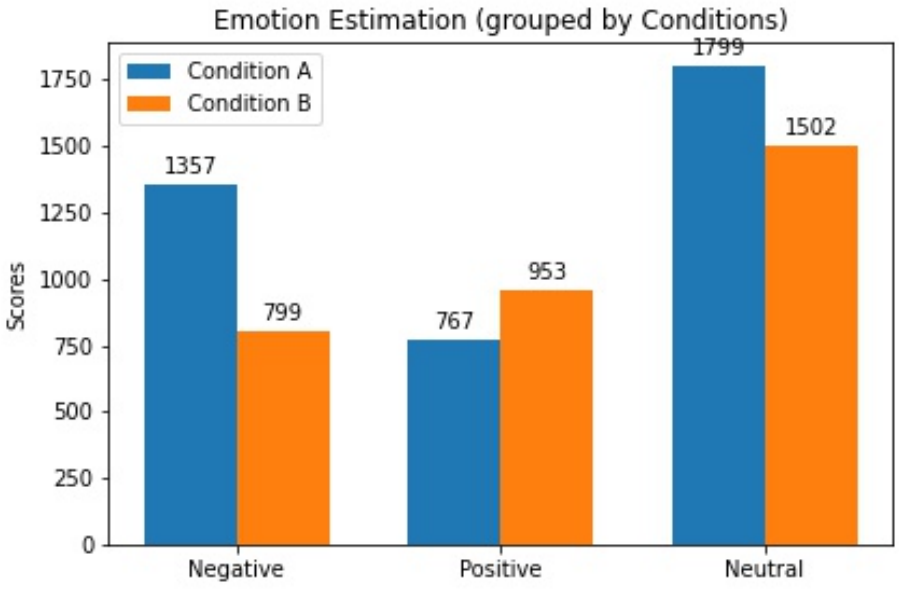


Fig. 4.5 Comparison of three emotional categories grouped by condition A and condition B

0, I calculated the sum of absolute values of the three angles as a new feature for analysis. The research question here was whether there is a statistically significant relationship between the sum of the absolute values of these three angles and the two conditions. The result of an independent sample t-test showed that there is a significant difference in the sum of the absolute values of these three angles and the two conditions ($M = 21.96, SD = 9.46$ for condition A, $M = 27.40, SD = 12.21$ for condition B), $t(703) = -6.61, \rho - value < 0.05$.

Moreover, to intuitively analyse the predicted results of the head-pose estimation, I compared two of these results with respect to condition A and condition B. Firstly, I plotted the sum of angles for conditions A and B. In Figure 4.6, the values of condition A (red bubble spots) form a less discrete distribution than condition B (green bubble spots). Second, I plotted the specific yaw, roll, and pitch angles for individuals on the timeline of my experiment. Figure 4.7 shows the fluctuations of the pitch angle, yaw angle and roll angle in the time series for the labelled time of condition A (red line) and the labelled time of condition B. From this, we can see that the angle of a head pose in condition A is generally smaller than the angle of the same participant's head pose angle in condition B.

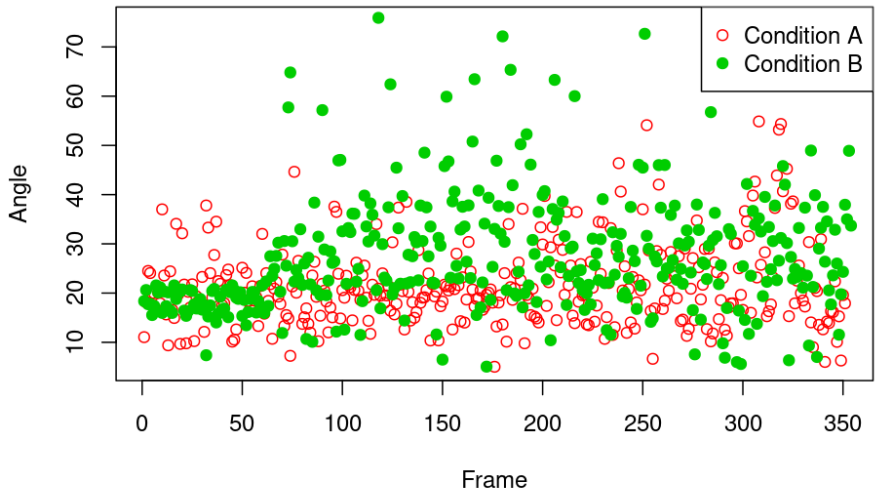


Fig. 4.6 Head-pose estimation: plot the sum of angles values for condition A and condition B

Regarding gaze estimation, I applied a state-of-the-art eye gaze estimation model that was trained on the ETH-XGaze dataset (Zhang *et al.*, 2020). This dataset collects more than one million high-resolution images of diverse human natural gazes in extreme head poses from 11 participants. The estimated values are angles of pitch and yaw that correspond to different directions of the eyes. Similarly to head-pose estimation, I calculated the sum of absolute angles of pitch and yaw and analysed whether there is a significant difference between the estimates and the conditions. Using an independent sample t-test, the result was that there is a significant difference in the sum of absolute values of pitch and yaw and the two conditions ($M = 0.44, SD = 0.26$ for condition A, $M = 0.49, SD = 0.22$ for condition B), $t(728) = -2.58, \rho - value < 0.05$.

I also applied the same methods as for head-pose estimation to plot the two groups of data features with labelled conditions for eye gaze estimation. Figure 4.8 indicates that the eye-gaze values of condition A form a more discrete distribution than these values of condition B. Figure 4.7 shows the fluctuations of the same individual participant's angles of pitch and yaw. Comparing the pitches and yaw angles in the timeline with the two conditions,

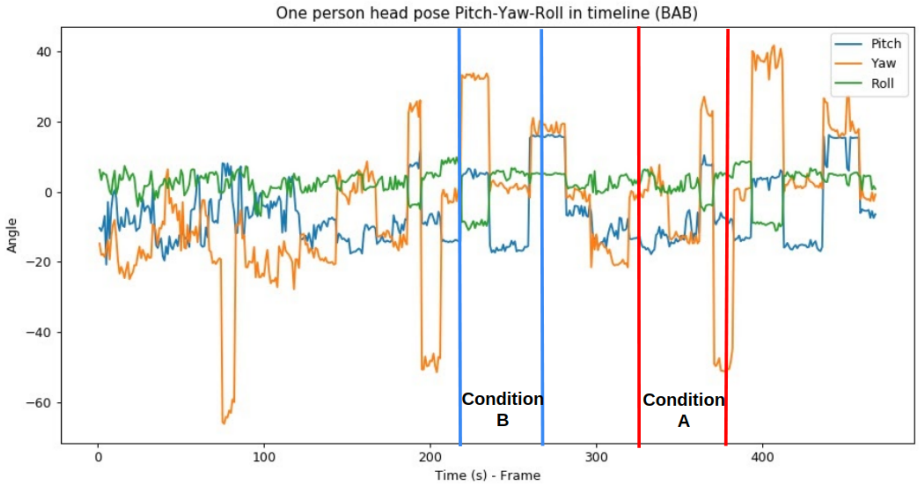


Fig. 4.7 Head-pose estimation: plot the change of one person's pitch, yaw and roll angles at an experiment

I can see that the eye-gazing angle of this participant in condition A is greater than that of this participant in condition B.

4.2.5 Subjective Measurement

I analysed the survey scores from user feedback with the two conditions, and proposed a number of sub-questions with respect to RQ2. The first sub-question is whether there is a statistically significant relationship between the average of self-reported confusion scores for each of the three executed tasks and the two conditions. The second three sub-questions are whether there is a statistically significant relationship between confusion scores for each of the three executed tasks and the two conditions. The statistical results are from an independent-samples t-test. I found that there is no significant difference between the average confusion scores of the three tasks and the two conditions ($M = 3.50, SD = 1.40$ for condition A, $M = 2.97, SD = 1.12$ for condition B), $t(36) = 1.28$, $\rho - value = 0.21$. For the second sub-questions with three pre-made tasks, first, there was no significant difference in the confusion scores for task 1 with the two conditions ($M = 3.00, SD = 1.07$ for condition A, $M = 2.44, SD = 1.33$ for condition B), $t(15) = 0.94, \rho - value = 0.36$;

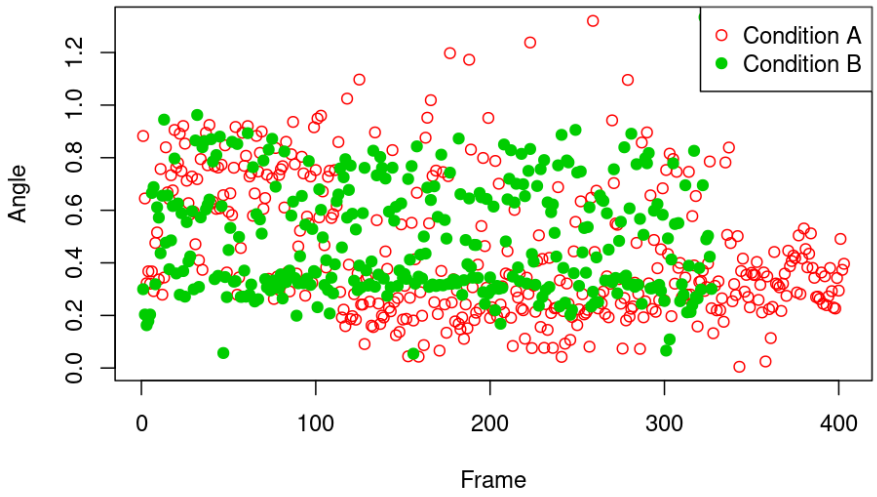


Fig. 4.8 Eye-gaze estimation: plot the sum of angles values for condition A and condition B

second, there was no significant difference in the confusion scores for task 2 with two conditions ($M = 3.09, SD = 1.22$ for condition A, $M = 3.10, SD = 1.29$ for condition B), $t(19) = -0.02, \rho - value = 0.99$; finally, the results did however indicated that there is a significant difference in the confusion scores for task 3 ($M = 4.38, SD = 0.74$ for condition A, $M = 3.00$, $SD = 1.12$ for condition B), $t(15) = 2.94$, $\rho - value < 0.05$.

4.2.6 Discussion

With respect to RQ2 and RQ3 identified earlier, there are a number of direct observations I can make on this data:

- Participants were not always aware they are confused if I gave them a specific confusing situation.
- When the participants answered the complex questions that seemed to trigger confusion states, the participants' emotion was more negative than when they answered straightforward questions that should avoid confusion stimuli.

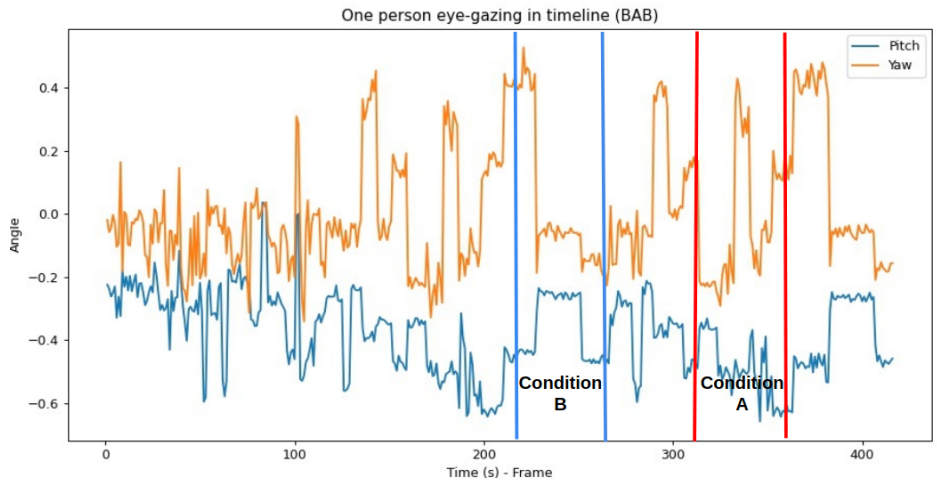


Fig. 4.9 Eye-gaze estimation: plot the change of one person's pitch and yaw angles at an experiment

- When the participants answered the complex questions that seemed to trigger confusion states, the range of angles of eye gaze was greater than when they answered the straightforward questions that should avoid confusion stimuli.
- Finally, when the participants answered the complex questions that seemed to trigger confusion states, the range of head-shaking angles was less than when they answered the straightforward questions that should avoid confusion stimuli.

While there are interesting observations, it should of course be noted that this was a pilot study of confusion induction and detection with a limited sample size and scopes. Moreover, there were also some limitations in the study execution that I identified. First, the quality of the videos of the participants was varied due to the different properties of the network, camera specification and camera position; the sample size and the range of participants' backgrounds were major limitations leading to a kerb on the conclusions of this study. Third, as noted earlier, confusion can transition to the productive confusion state or the unproductive confusion state; in my study, there are no clear time and dialogue boundaries to clarify whether participants are in a productive or unproductive confusion state. Finally, in the 3-minute post-interaction interview, many participants expressed that they expected more

natural and casual conversation with the avatar; however, this experiment lacked flexible and free conversations between participants and the avatar.

Despite these limitations, the study results reflect that the method for data collection and analysis was worthy and meaningful as an important initial step down the path to conducting further studies in human and real robot interaction. This work was presented in the following paper at Semdial 2021:

- Li, N., Kelleher, J.D. and Ross, R. (2021) Detecting interlocutor confusion in situated human-avatar dialogue: A pilot study in: 25th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2021) University of Potsdam, Germany

4.3 HRI Study 1: Detecting Interlocutor Confusion in Situated Human-Robot Interaction

In Study 1, I was limited by the challenge of having users interact remotely over uncontrolled hardware (*e.g.*, microphone and camera challenges on user laptops) and the more general challenge of managing interactions remotely. Nevertheless, the work did identify certain indicators of participant confusion, and hence in a second study, I wished to broaden the investigation to provide a complete interaction scenario with a dataset that can, subject to privacy concerns, be made available for general study in language-based HRI.

For this study, I made use of a Pepper robot. Of its many features, those that are relevant here are its onboard 2D cameras, ability to articulate arms and head for gesticulation, and on-chest touch screen. Pepper has speech recognition and dialogue available in 15 languages. For this study, the Pepper was configured for English. The Pepper backend is a fully open and programmable platform built on the Naoqi framework with comprehensive animated speech, motion, and vision modules, which were used to support my WoZ experiment.

4.3.1 Study Design

This study used a semi-spontaneous one-by-one physical face-to-face conversation between the Pepper robot and a participant. The Pepper robot was controlled by a wizard. All participants were required to be able to walk into our physical laboratory. Two rooms were set up (see Figure 4.10): the experiment room was set up for the participants with the Pepper and some additional recording equipment. Participants were asked to stay standing in Zone 1 which is around 80 cm in front of the robot to ensure that they were close enough to Pepper for practical interaction. A high-definition (HD) webcam (Webcam 1) was placed behind the Pepper robot and aimed toward the participants' faces to collect their facial expressions. A second HD webcam (Webcam 2) was placed on the right side of the Pepper to record the body gestures of the participants. Figure 4.10 shows the actual scene of the laboratory setting.

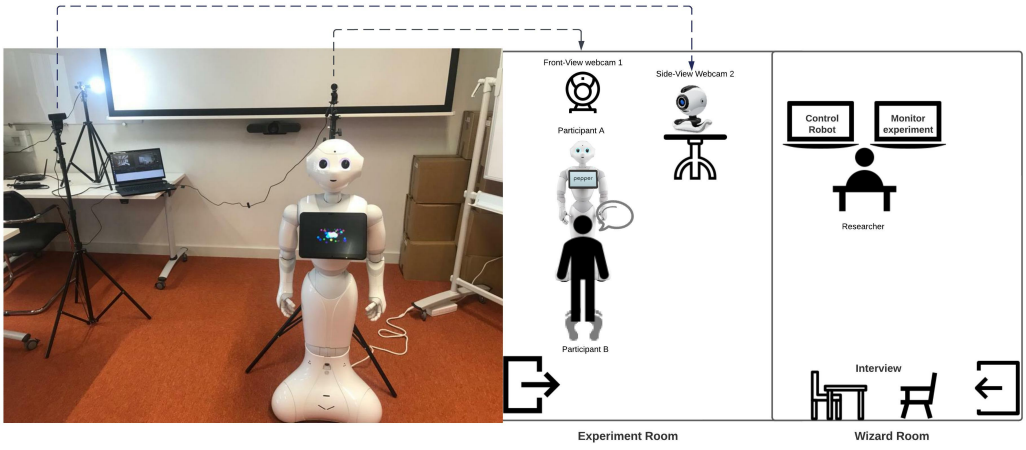


Fig. 4.10 WoZ HRI Experiment Laboratory
(left: the real experiment room; right: a mock experiment room and a mock wizard room)

The wizard room was designed for the researcher to monitor the real-time interaction between the participant and the Pepper robot in the experiment, as well as to control the

Pepper robot using a WoZ4U independent platform (Rietz *et al.*, 2021) (see Figure 4.11). The WoZ4U platform is an open-source WoZ interface that provides a graphical user interface (GUI) for the wizard to control Pepper movements, speech utterances, animated speech, gestures, autonomous mode, *etc.*. I also integrated conversation scripts and developed more specific behaviours for the Pepper robot in WoZ4U. These included a QR code for feedback that was made available on the Pepper tablet, the Pepper's animated speaking with anthropomorphic body language (happy, embarrassed, wave, *etc.*), and the use of the Pepper robot's lively colourful eyes. In addition, I also collected users' video data from the Pepper robot's forehead camera by using the camera viewer in a monitor application from the SoftBank Robotics company.

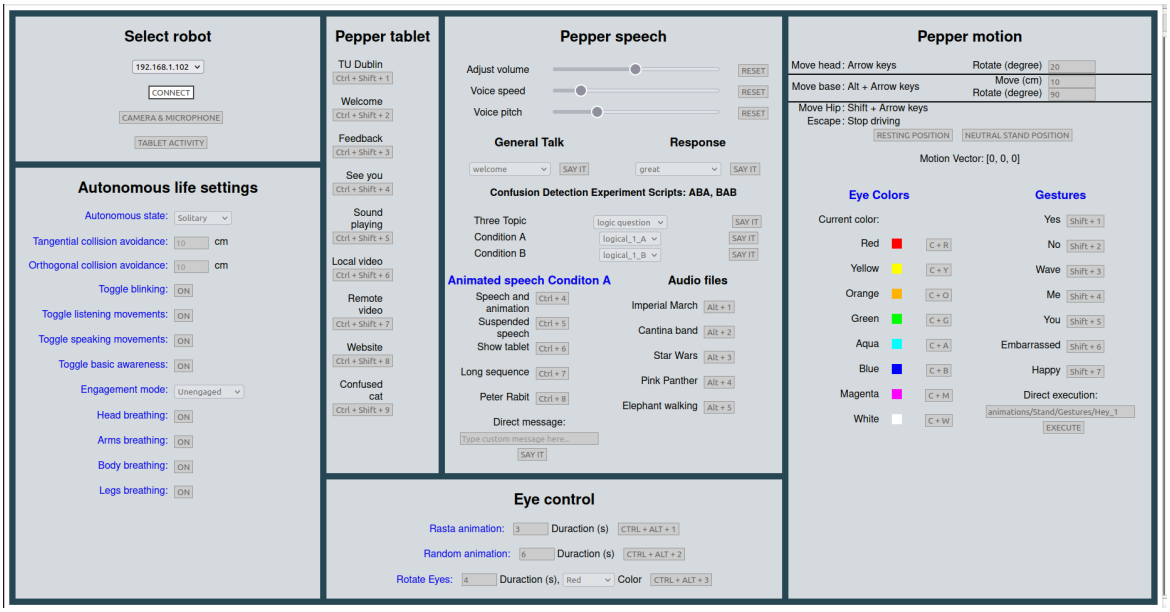


Fig. 4.11 WoZ4U Platform (Rietz *et al.*, 2021)

Figure 4.12 shows the procedure for this HRI experiment. All participants registered first for this walk-in experiment, including their email address, basic information, and the date and time to attend this experiment. Consistent instructions and consent forms were provided and signed before participants attended the experiment as well. Live participation was designed around two interaction sessions lasting more than 15 minutes. The first session

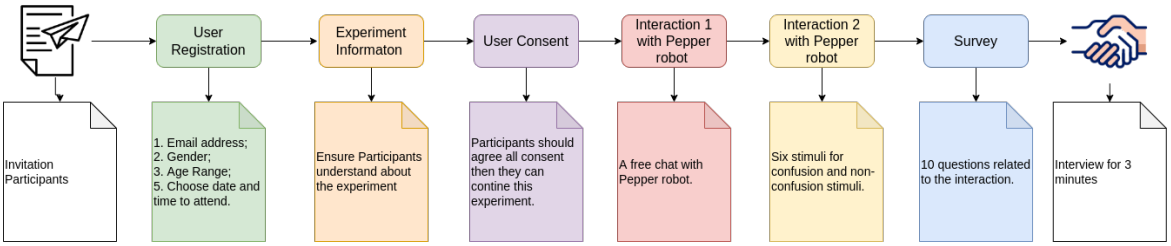


Fig. 4.12 HRI Experiment Process

was a casual talk because most of the participants had no experience of interaction with the Pepper robot prior to participating in this experiment. To help participants adapt to the mode of human-robot dialogue, I prepared 11 interactive topics that the participant could engage in (*e.g.*, "What is your name?", "Raising your arms", *etc.*) as a reference so that they could feel more comfortable and confident entering the second session. The second session was a 5-8 minute task-oriented conversation between the participant and Pepper (detailed later). The behaviours and speech of the participants were recorded in this session. The participant completed a post-study questionnaire, which was then followed by a 3-minute interview discussing this interaction.

The conversations were the same as the conversations from Study 1, namely that the Pepper robot asked three types of tasks (logical questions, word questions, and maths questions) to stimulate non-confusion and confusion states from each participant. The behaviour and speech of the participants were recorded. Also, as before, session 3 was a survey in which the participant had to answer 10 questions, which were similar to the questions used previously. The last session (session 4) was an interview session in which the researcher could interview the participant about the HRI experience. This was beneficial in improving the setup and design of the experiments.

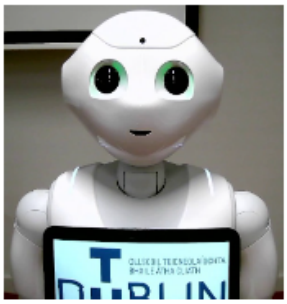
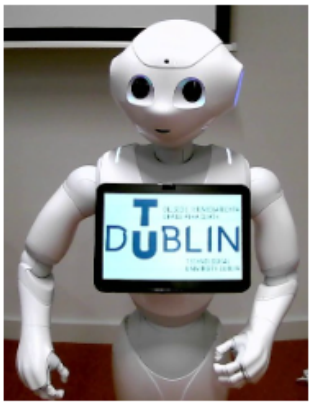
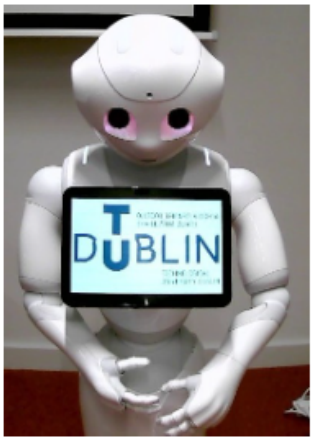
Conversational Phenomena	Communicative Behaviour		
	Face	Body	Details
Positive response			Eyes colour: green Head: face toward the participant. Body language: arms and hands are jogging swings, and head is swaying happily following the arms.
Negative response			Eyes colour: red Head: head down Body language: hands are close together in front of Pepper's body.

Fig. 4.13 The mapping of response status and behaviours of Pepper robot

4.3.2 Dialogue Design

The conversational HRI dialogue design includes four patterns of confusion stimuli. Meanwhile, to build a more natural interaction with participants, it is vital that the robot possess interactive nonverbal behaviours (Prasad *et al.*, 2020). Therefore, I designed a mapping of physical behaviours on the robot's head, eye colours, and body gestures to align with positive and negative responses (see Figure 4.13).

4.3.3 Data Analysis

30 individuals participated in this study. Among them, one participant helped to first dry run this study, so this participant's data was not analysed for research purposes. All participants were over 18 years of age, *i.e.*, 5 people in the 18 - 24 age group; 23 people were in the 25 - 44 age group; and 2 people are in the 45 - 59 age group. What's more, they were from at least six countries and were in university programmes, or were in industries, such that they were able to have a social conversation in English. Data from 29 participants (16 males, 12 females, and one was not stated) were made available for data analysis.

Data collection, for each participant, included speech data, facial video, postures and gestures. The data labelling strategy was the same as had already been applied in study 1 HAI (see table 4.2), namely three tasks with the two conditions (ABA or BAB) were used to stimulate participants' confusion and non-confusion states. I applied the same feature extraction algorithms as in Study 1 for emotion estimation, eye-gazing estimation, and head-pose estimation on the labelled facial frame data from 29 participants. The image data that I extracted had 5715 frames (3441 frames for Condition A, and 2274 frames for Condition B). Although that image data was from facial videos, it is necessary to recognise and align faces in a preprocessing data step. I applied a general approach for each frame from the centre crop in a region of 224×224 pixels which was then used to detect the face and remove the frame margins using the Multi-task Cascaded Convolutional Network (MTCNN)-based face detection algorithm (Savchenko, 2021). As facial video data also included high-quality audio in this study, I applied the FFmpeg framework, which is an effective video and audio converter, to extract audio tracks for analysis. The audio data had 85 audio files (45 waveform audio (wav) files for condition A and 40 wav files for condition B). The post-questionnaire with 10 questions on a Likert scale of 1-5 levels was completed by each participant after interacting with the robot. The 29 post-study questionnaires were split by the conditions independently such that I have prearranged sequences of conditions for each participant.

Table 4.4 The result of emotion estimation grouped by Condition A and Condition B

Condition	Anger	Disgust	Fear	Sadness	Happiness	Surprise	Neutral	Overall
A	40	62	1511	59	91	67	92	1922
B	19	46	1503	57	151	48	102	1926

I then combined the two independent files (one for condition A, another for condition B) into one file with the new "Condition" feature to mark the specific condition for subjective analysis. As there were two scores under the same conditions for each questionnaire, I calculated the average of the two scores as a new parameter.

Frame Data Analysis

For frame data analysis, I again applied the facial emotion detection algorithm to my preprocessed frame data, this algorithm used the MobileNet architecture and was trained on the AffectNet dataset (Savchenko, 2021; Howard *et al.*, 2017; Mollahosseini *et al.*, 2017). This resulted in estimates of the each of seven primary emotion categories (neutral, happy, sad, surprise, fear, anger, and disgust) for each frame. This estimation resulted in 2945 labelled frames for Condition A and 1941 frames for condition B. Table 4.4 showed the number of each of the seven emotion categories grouped by condition and normalised by total detection. I noticed that the number of fear emotions was much higher than the other six emotions. On investigation, I see this as a limit or bias in the algorithm, and subsequently removed the count of fear labels from further analysis.

For analysis, I grouped the 3 negative emotions (anger, disgust, and sadness), and note that the number of negative emotions in condition A is considerably greater than in condition B. Correspondingly, the number of predicted positive emotions (happiness) results for condition A is less than that for condition B. Similarly, surprise (either a negative or positive emotion in different contexts (Vrtika *et al.*, 2014)) was higher in condition A than in condition B. Finally, the predicted results of the neutral results for condition A are less than those for condition B (see Figure 4.14 for a summary of these aggregated results).

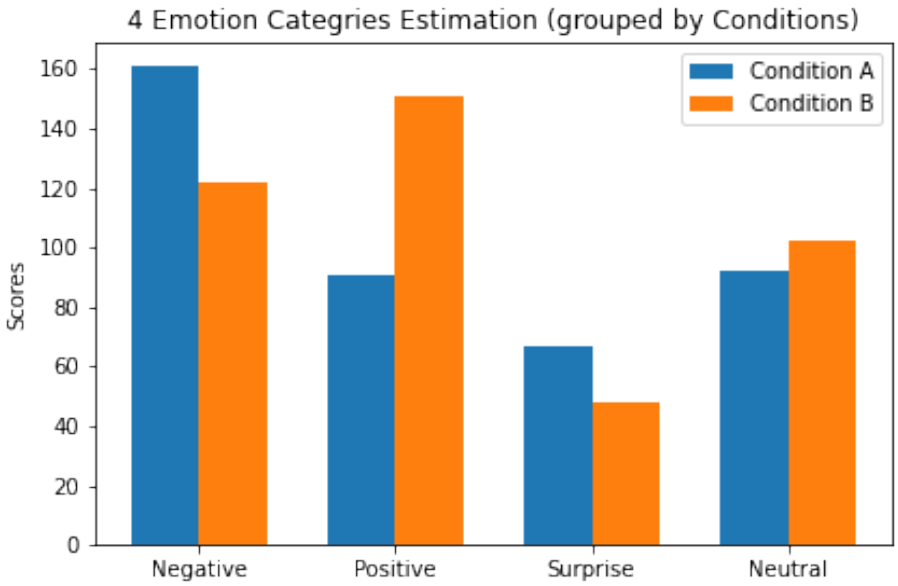


Fig. 4.14 The four emotion categories grouped by Condition A and Condition B

As a case study, I also applied the Facial Expression Recognition (FER) open-source framework to the preprocessed videos for exploratory purposes. The FER framework is built with MTCNN for facial recognition (Zhang *et al.*, 2016) and an emotion classifier (Arriaga *et al.*, 2017) that has been trained on the FER-2013 emotion dataset (Goodfellow *et al.*, 2013). To illustrate, I plotted the seven changes in emotions following the sequence of conditions for the three tasks (see Figure 4.15). It shows that emotion "happy" dominated most of the time in the two instances of condition B, with scores of "happy" approaching 1. Whereas the proportion of "happy" decreased, while each of "sad", "fear", "surprise", and "neutral" became dominant in periods of condition A.

Turning to eye gaze, I again applied a state-of-the-art eye gaze estimator, trained on the ETH-XGaze dataset (Zhang *et al.*, 2020), to predict pitch and yaw angles for each preprocessed frame. I summed the absolute two angles as a new feature for statistical analysis as a human has different angles of direction corresponding to positive or negative values of pitch and yaw, leading to the sum of values being 0. An independent-samples t-test was conducted to compare the angles of pitch and yaw for eye gaze under the two

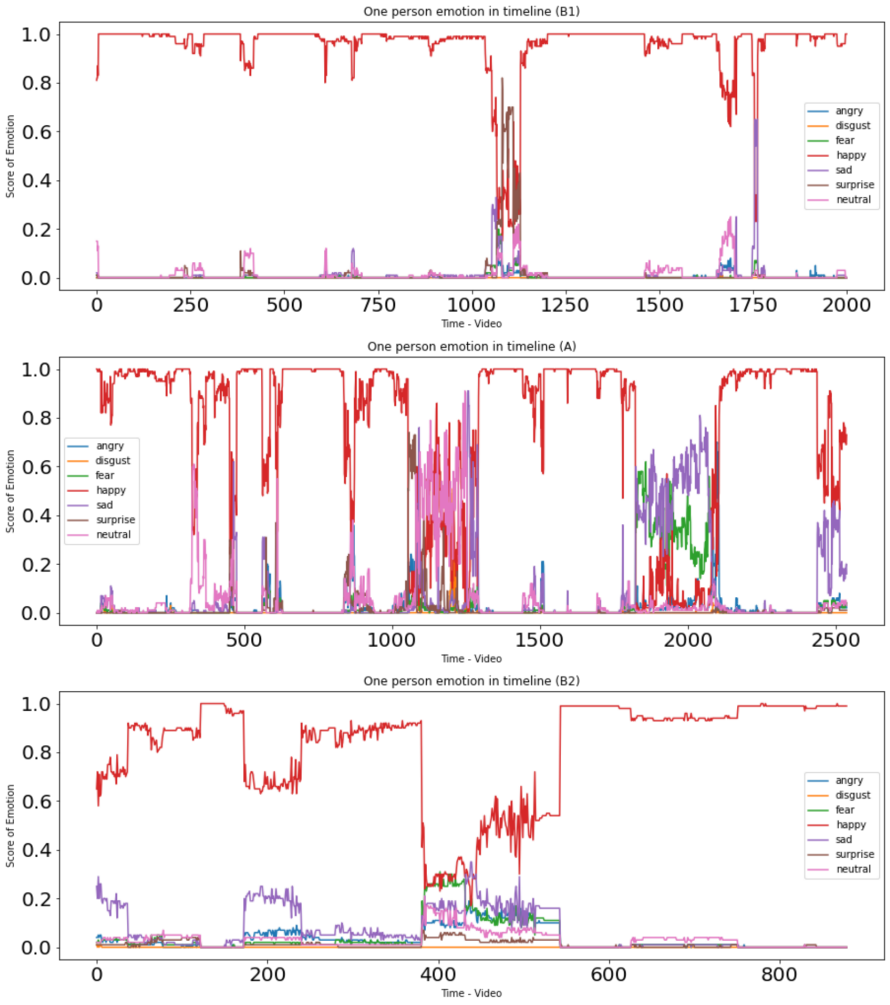


Fig. 4.15 The emotional changes for one participant during the three task with the two conditions

experimental conditions. A significant difference was found in the angles of pitch and yaw for eye gaze ($M = 0.44, SD = 0.15$ for condition A, $M = 0.46, SD = 0.15$ and condition B), $t(2587) = -2.27, \rho - value < 0.05$).

To illustrate the overall trend, Figure 4.16 shows the fluctuations of the pitch angle and the yaw angle with the two time periods labelled condition B and the one time period of condition A for an example user. I can see that the average area of two angles in condition A is greater than the average area of two angles in each of the two conditions B instances.

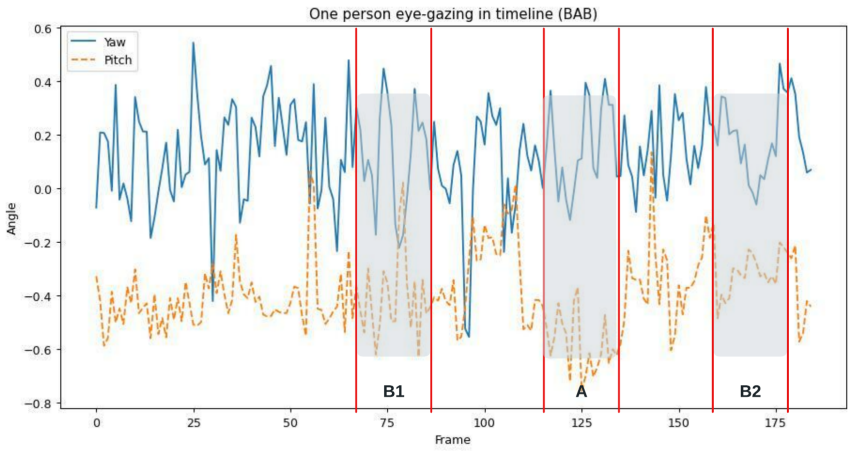


Fig. 4.16 Changes of one person's pitch and yaw angles of eye-gazing in timeline

For head-pose estimation, the model that I applied was designed using CNNs, dropout, and adaptive gradient methods (Patacchiola and Cangelosi, 2017), and trained on three popular datasets (*i.e*, the Prima head-pose dataset, the Annotated Facial Landmarks in the Wild (AFLW) dataset, and the Annotated face in the Wild (AFW) dataset) (Gourier *et al.*, 2004; Köstinger *et al.*, 2011; Zhu and Ramanan, 2012). The estimated metrics were the three angles of pitch, yaw and roll. Again, I calculated an aggregate value of the three absolute angles as a new variable. The result of an independent-samples t-test showed that there was, however, no significant difference in the angles of roll, pitch and yaw for head pose ($M = 24.52, SD = 11.72$ for Condition A, $M = 24.55, SD = 11.35$ for Condition B), $t(5713) = -0.08, \rho - value = 0.94$).

4.3.4 Audio Data Analysis

The phenomenon of silence during conversations has been analysed previously in pragmatic studies (Ohshima *et al.*, 2015). There are two types of silence for the specific state of the interlocutor: intentional silence in which the interlocutor refuses to respond to the speakers; and unintentional silence when the interlocutor psychologically cannot respond to

the speakers (Kurzon, 1998; Ohshima *et al.*, 2015). This can then be taken as a potential proxy or indicator of confusion. Given this, I calculated the silence duration time for each audio stream labelled with the specific condition. An independent-samples t-test was then conducted to compare the silence duration time between the two conditions. There was a significant difference between the silence duration time for the two conditions ($M = 36.22, SD = 13.52$ for condition A, $M = 25.68, SD = 10.82$ for condition B), $t(83) = 3.94, \rho - value < 0.05$).

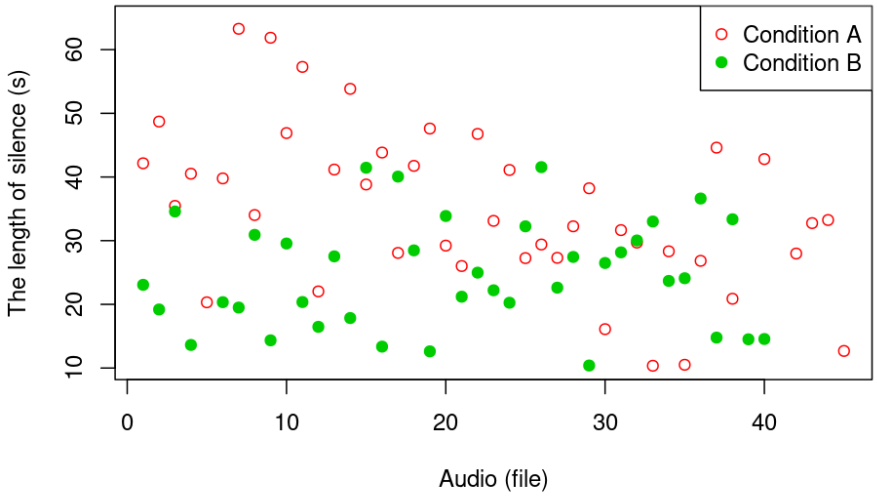


Fig. 4.17 Plotted silence duration time grouped by Condition A and Condition B

To illustrate, I plotted the silence duration time for the two conditions (see Figure 4.17). It shows that the silence duration values of condition A form a more discrete distribution than those of condition B. Meanwhile, for an individual observation, Figure 4.18 presents two changes in silence duration time for two participants in the three tasks performed with different sequences of conditions (*i.e.*, the blue chart is the "BAB" conditions, while the other is the "ABA" conditions). Here, I can see that the silence duration time for condition A is obviously longer than for each of the two condition B instances, either within a participant or between the two participants.

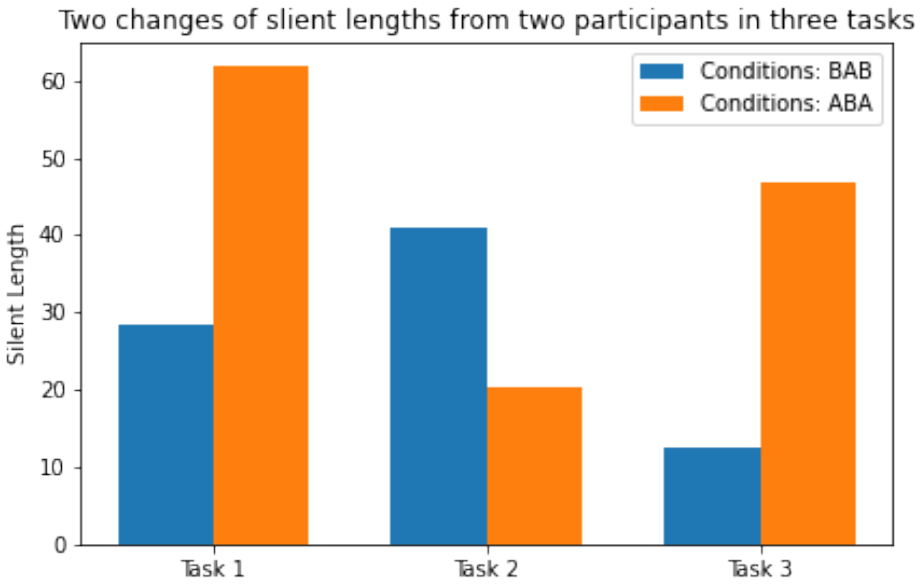


Fig. 4.18 Changes of silence duration time for an individual (BAB) and another individual (ABA) with different sequences of conditions

4.3.5 Subjective Analysis

I analysed the post-study questionnaire scores against the two independent groups divided by two controlled conditions for the stimuli. Two statistical questions were investigated: The first question investigated the relationship between the three task-centric confusion sub-question scores and the two conditions (A and B). The second question investigated whether there was a significant relationship between the average self-reported confusion scores for the three tasks and the two conditions.

An independent-samples t-test was conducted for each question. The results of the first question with the three sub-questions were: (1) There was an almost significant difference (trend) between the confusion score for task 3 (a Maths question) ($M = 2.27, SD = 0.99$ for condition A, $M = 2.31, SD = 1.28$ for condition B), $t(24) = 2.00, \rho - value = 0.056$). (2) There was no significant difference in the score of confusion for task 1 (a logic problem) ($M = 2.71, SD = 1.20$ for Condition A, $M = 2.50, SD = 1.61$ for Condition B), $t(24) = 0, \rho - value = 1.00$). And (3) there was no significant difference in the confusion score

for task 2 (a word problem) ($M = 2.46, SD = 1.30$ for Condition A, $M = 2.47, SD = 1.36$ for Condition B), $t(24) = -1.047, \rho - value = 0.31$). The result of the second question showed that there was no significant difference in the average of confusion scores for the three tasks performed and the two conditions ($M = 2.26, SD = 0.99$ for condition A, $M = 2.31, SD = 1.28$ for condition B), $t(50) = -0.12, \rho - value = 0.90$).

4.3.6 Discussion

Reflecting on the results presented in the previous section, I can identify the following six observations of note:

- Participants were not necessarily aware of being confused when presented with the confusion stimuli.
- Participant's emotions were more negative and more surprised in confusion conditions than in non-confusion condition states.
- Participant's ranges of eye gaze angles were greater in confusion than in non-confusion states.
- Generally, the silence duration time was longer in confusion than in non-confusion states.
- There was no strong correlation of their ranges of the angle of the head pose with confusion or non-confusion states.

Compared to existing human-like avatar interaction work (Study 4.2), in this HRI study, although the Pepper robot lacks anthropomorphic facial expression, the Pepper robot has advanced body language and appropriate automatic animated speech. Meanwhile, the quality of the data is guaranteed as I controlled the variables in the setup of the experiment environment. In a 3-minute interview, most participants are surprised that the Pepper robot

Table 4.5 The Result of emotion estimation grouped by two conditions in HAI and HRI

Modality	Condition	Anger	Disgust	Fear	Sadness	Happiness	Surprise	Neutral	Overall
HAI	A	262	282	136	677	702	65	1799	3923
	B	77	165	57	480	858	95	1502	3234
HRI	A	40	62	1511	59	91	67	92	1922
	B	19	46	1503	57	151	48	102	1926

has high-tech social interaction skills and friendly behaviours, which is better than the feedback from our earlier study.

In the case of emotions detected, in the HAI experiment, the estimations in Condition A corresponding to the four classes of negative emotions (anger, disgust, fear, and sadness) are stronger than in the case for these classes in Condition B. In contrast, the number of predicted results for the two positive emotions (happiness and surprise) in Condition A was less than in Condition B. In the HRI experiment, I can see that the results of the five main predicted emotions are slightly similar to the HAI experiments, except for the two special surprise and neutral emotions (see Table 4.5).

However, some limitations must be mentioned. First, 25 out of 29 participants had a technical background in computer science. Thus, the sample size and range of background may influence my interaction results compared with real-world HRI work. Second, given the estimations for facial emotion estimation, the pre-trained feature algorithms were not possible to apply reliably for the HRI studies. Third, most participants' voices were not loud enough (*i.e.*, the amplitude of those participants' voices was far lower than the amplitude of the Pepper robot voice). Figure 4.19 shows the amplitude difference between a participant and the Pepper robot for an audio conversation. The small bits labelled the participant was when the participant is talking, and the big bits correspond to when the robot is talking. Therefore, it is very challenging to separate the two speakers' speeches in the same video file for only analysing participants' speech data. Third, there was no control of conversation boundaries to reflect the different confusion states (*i.e.*, positive confusion or negative confusion) in these 5-minute confusion stimuli conversations.

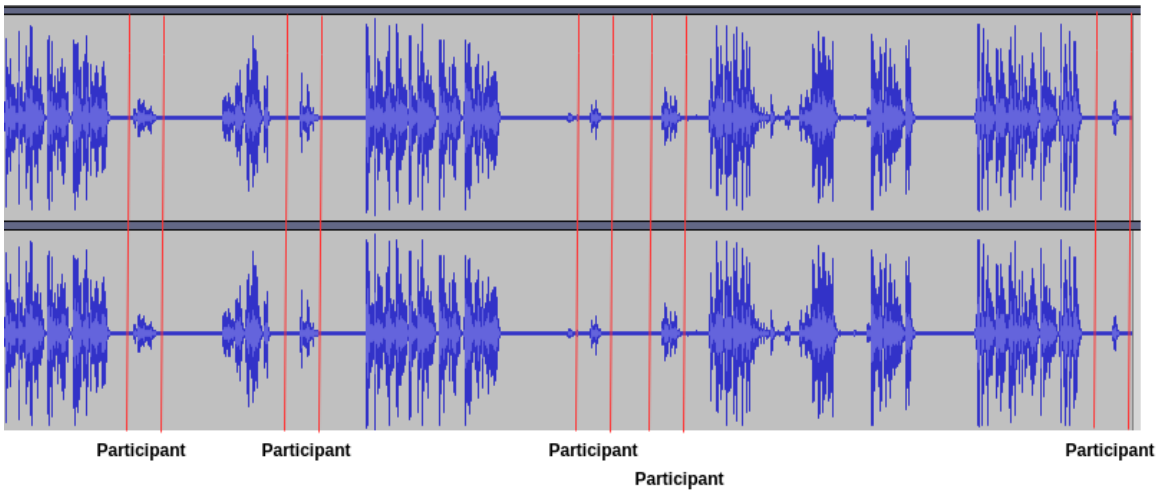


Fig. 4.19 A amplitude of one audio conversation between a participant and the Pepper robot

Despite these limitations, this study illustrated that even when users are not aware of being in a confused state, they present different interaction behaviours which may in principle be detected by automated systems such as social robots. This has the potential to increase the social task-oriented capabilities of dialogue-equipped robots in the medium to long term. This is my first study on modelling confusion states in a situated HRI task-oriented dialogue setting. Nevertheless, I see it as a firm foundation for further situated dialogue investigation for HRI, and in particular, where I focus on enhancing engagement through preemptive anticipation of disengagements.

This work is currently being prepared for submission to 18th Annual ACM/IEEE International Conference on Human Robot Interaction (HRI) 2023.

4.4 Comparing Two Embodied Interactions

Given the two studies just presented, it allows us to begin to answer the following two sub-questions for RQ4: **(1)** To what extent can we rely on human-avatar interaction (HAI) studies as a substitute for human-robot interaction (HRI) data collection efforts? **(2)** Does the choice of embodiment type (HAI vs. HRI) have a significant effect on users' perceptions and

mental state with respect to my central challenge of confusion detection? In the following, I present some specific analyses of the two studies to help to answer these questions.

4.4.1 Data Analysis

The HRI study (Section 4.3) analysis mirrors the approach taken with the earlier HAI study (Section 4.2) that includes emotion detection, eye gaze estimation, and head pose estimation for both facial frame data. Generally, I found that there were minimal data collection issues in the HRI study compared to the HAI study, since the researcher controlled all variables in the HRI setup.

Considering gaze estimation, through an independent-samples t-test of results from the HAI and HRI studies, a significant difference was found in the sum of absolute values of pitch and yaw across the two conditions for HAI and HRI experiments, respectively. In the case of head pose, the independent-samples t-test result shows a significant relationship between the sum of absolute values of the three angles (pitch, yaw, and roll) and two conditions for the HAI study; while there is no significant difference between the sum of absolute values of these three angles and two conditions for the HRI study.

Regarding the subjective self-reporting scores for the 48 participants, I analysed the self-estimated confusion scores of each participant and the user attitude towards the embodiment option (avatar vs. social robot). For both HAI and HRI options, a significant difference was found between confusing and non-confusing tasks only in the case of task 3 (maths problems). There was no significant correlation in the confusion scores for task 1 (logic problems) and task 2 (word problems) for conditions A and B in the HAI or HRI studies. However, what is more interesting is the analysis of the user experience questions: this included an examination of whether there is a significant difference between the average scores of negative feedback of the user's experiences in the two studies; an examination of whether there is a significant relationship between the average scores of positive feedback

of the users' experiences in the two studies; and finally, an examination of whether there is a statistically significant relationship between the score of participants wanting to abandon these conversations and these two studies.

I also found that there was a significant difference between the average negative feedback scores in the two studies ($M = 2.77, SD = 0.85$ avatar, $M = 1.91, SD = 0.62$ robot), $t(88) = 5.5547, \rho - value < 0.05$. Regarding the second question, there was a significant difference between the average positive feedback scores of user experiences and the two modalities ($M = 3.33, SD = 0.92$ avatar, $M = 4.09, SD = 0.54$ robot), $t(88) = -4.72$, $\rho - value < 0.05$. Lastly, the result indicated that there was also a significant difference between the scores for which participants want to abandon the conversations with the two studies ($M = 3.21, SD = 1.34$ avatar, $M = 1.34, SD = 0.62$ robot), $t(88) = 8.35$, $\rho - value < 0.05$.

4.4.2 Discussion

Based on these results, I observed that when participants are confused, the changes in their emotions and gaze movements after stimuli from the different conditions of the HAI study are similar to those of the HRI study, but that the changes in the range of head pose angles with different stimuli from the different conditions of the HAI study are different from those of the HRI study. Furthermore, participants prefer to engage in interaction with the robot platform rather than with the avatar in this research study and are more willing to continue to interact with the robot platform than with the avatar in this research study.

While these basic observations can be made, it is very notable that, while I attempted to unify my studies across embodiment types, it is hard in practise to achieve this. At a very technical level, my human-avatar studies were less controlled as users could participate from home – unlike in the case of my human-robot studies. Generally, there was a very low abandonment rate for the HRI study, and I can also ensure the same quality of the dataset that I collected.

Meanwhile, it should be mentioned that during the 3-Minute post-task interview, in the HAI experiment, the expectation of interaction from many participants was found to be much higher relative to the actual capabilities of the avatar. However, from the feedback of the participants in the HRI experiments, they felt fresh and curious talking to the robot, so most of them enjoyed and engaged in this HRI experiment. Also, they were surprised that the Pepper robot has high-tech social interaction skills when the Pepper robot vividly interacts with them.

This work was presented in the following paper at the MMAI2022 Workshop 2022:

- Li, N. and Ross, R. (2022) Transferring studies across embodiments: A case study in confusion detection in: 1st workshop (MMAI2022) that is a part of the conference on Hybrid Human-Artificial Intelligence 2022, Amsterdam, Netherlands

Chapter 5

Planned Work

The main objective of the next phase of my research is to iterate and extend my HRI studies to further investigate my work in situated conversational interaction. I begin by illustrating the planning before outlining work on potential dialogue policy modelling.

5.1 HRI Study 2: Detecting Interlocutor Confusion in Situated HRI

The first proposed study is an iteration of HRI Study 1 (Section 4.3). In the last HRI study, there were still some areas for improvement and increased control, for example, improving the methods of multimodal data collection, and enriching the interactive information on the Pepper robot, and so forth. In the second proposed study, I plan to focus on designing and launching new dialogues for investigating different states of confusion stimuli.

5.1.1 HRI Study 2.1: Iteration Study: Quality Improvement for Confusion Detection in Situated HRI

In HRI Study 1 (Section 4.3), some limitations were mentioned in terms of experiment design (*e.g.* low-quality recordings for participants' voices) and no significant differences result between conditions for certain measurements (*e.g.* user survey and head post estimation). Therefore, the goal of this iteration study, this study is to iterate HRI Study 1 in order to improve the quality of experiment design, data collection and dialogue design.

Study Design

As the HRI dialogue design for my study is a task-oriented one-to-one conversation, each participant in the new study will be assigned three tasks (*i.e.*, word problems, logic problems, and maths questions) that are the same as in the earlier study, but in order to enhance the interactivity between a participant and the Pepper robot, one more task will be introduced (see Figure 5.1). This task will be centred around a tabletop pick-and-place task. A table will be set up between a participant and Pepper, Pepper will be placed behind the experiment table, and the participant stands at the front of the table. On the table, there will be three coloured square boxes with numbers (*i.e.*, the red box is number 1, the blue box is number 2, and the yellow box is number 3) as well as four coloured blocks at the front of the three boxes (*i.e.*, two red and the green cubic blocks on the left, a blue cone block in the middle, and a yellow cylinder block on the right) toward to the participant's side. Following the Pepper robot's oral request, the participant should look for a particular block to put on a particular square box. One webcam will be used by the researcher to obverse the participant's operation result in each time. Each pick-and-place task has one condition (condition A or condition B) to trigger the participant's confusion or non-confusion. Therefore, each participant in this

study will have four tasks in total to complete, and the sequence of conditions in the four tasks will be designed as either condition-ABAB or condition-BABA.

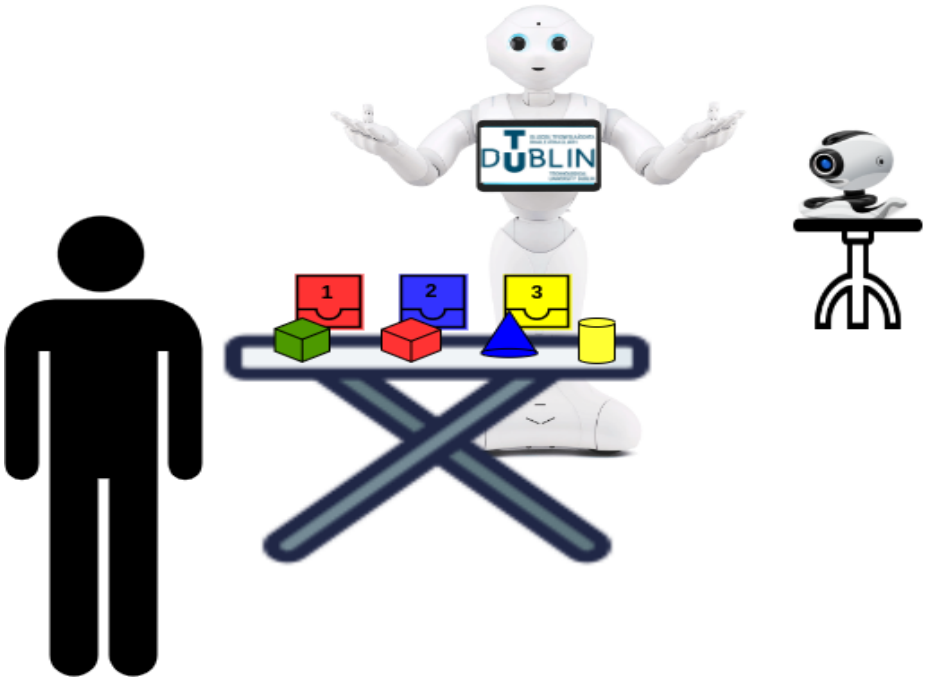


Fig. 5.1 Pick-and-place task setup

In order to improve control and reduce the chance of "leakage" across tasks, before the next task commences, the participant will be given a 1-minute break to reset their emotion from the previous interaction. Moreover, in the first 30 seconds of the break time, the participant will be required to rate their confusion scores in the post-study survey immediately. This should also help to ensure that users self report more accurately on their confusion scores, as in the earlier design, there was much long of a gap between stimuli and the reporting phase. Furthermore, from the early HRI study, I noticed it was very difficult to separate the two speakers' speech in the same video file for analysis of emotion estimation in speech. In the experiment room, the third webcam with a high-quality microphone will thus be set up on the left (see Figure 5.2) in this study.

Comparing to the avatar behaviours in the HAI study, the avatar has human-like facial expressions, but the Pepper robot does not pretend to be a human as it is without facial

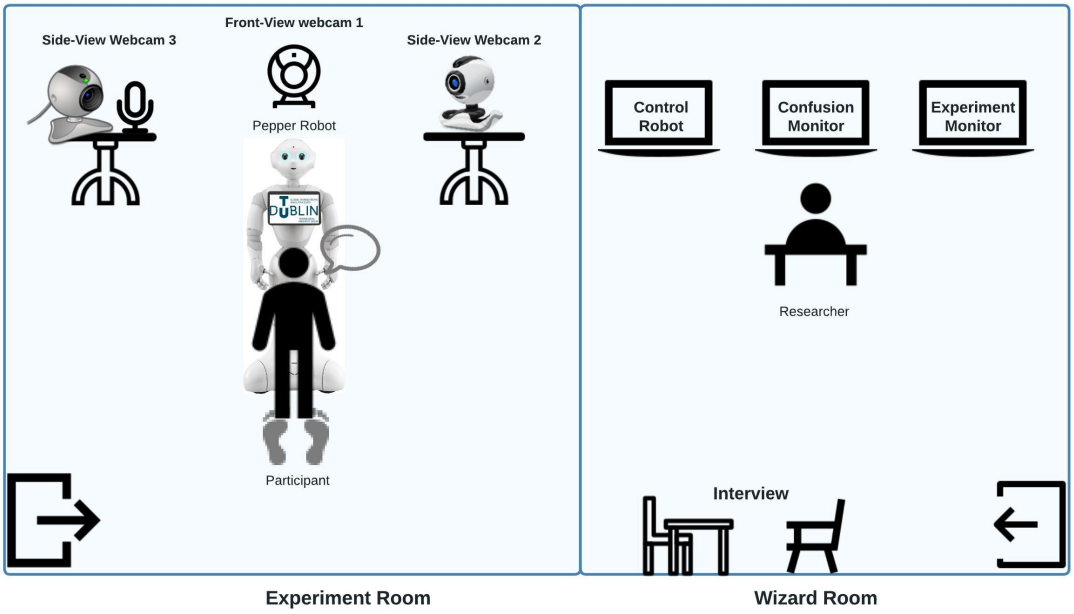


Fig. 5.2 Improved Experiment setup

expressions. To provide more multimodal output for the Pepper robot in order to improve interaction (Applewhite *et al.*, 2021; Fadhil *et al.*, 2018; Jam *et al.*, 2021), I will enable five facial emojis with five typical emotions (satisfied, cheerful, doubtful, surprised, and frustrated) to show on the Pepper robot screen during interactions with participants (see Figure 5.3). During the interaction between the Pepper robot and the participant, One cartoon emoji will be shown on the Pepper's tablet to provide an accurate response or feedback for each time to the participant's response.

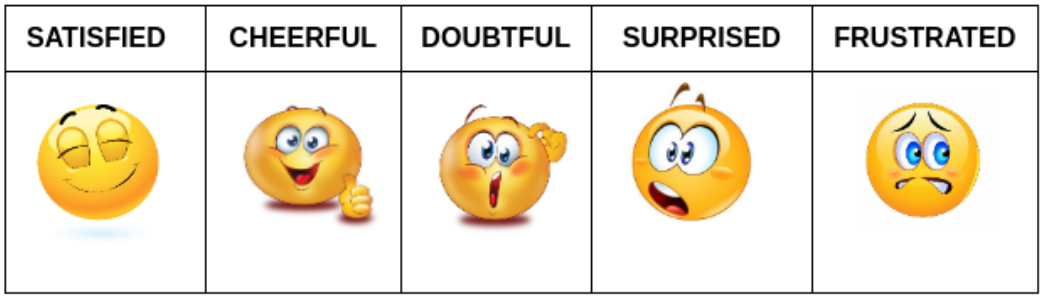


Fig. 5.3 Facial emojis with different emotions

5.1.2 HRI Study 2.2: Different States of Confusion Detection in Situated HRI

Although the study just introduced does aim to refine my experimental setup to allow testing of a number of RQ from Chapter 3, it does not in itself say anything about confusion except that there exist confused and non-confused states. Furthermore, it can also be seen that some papers in Chapter 2 only mentioned that confusion can, in fact, be a phenomenon with multiple levels, and in some cases even having four classes from very high to very low. In particular, Lodge *et al.* (2018) proposed two states of confusion *i.e.*, productive confusion and unproductive confusion through the ZOC and ZOSOC (Lodge *et al.*, 2018). It is reasonable to expect that the interpretation of the specific state of confusion based on an interlocutor's behaviours can potentially improve engagement in conversational HRI. However, these works do not have any constraints on the level of confusion belonging to either productive confusion or unproductive confusion, and also do not have definitions for the two states of confusion. Therefore, as a starting point for teasing these issues apart, we can define the two terms of productive confusion and unproductive confusion as follows:

- ***Productive confusion*** *is the first stage of confusion: An impasse in the flow of interaction is generated as a disequilibrium state; the human has meta-cognitive awareness of the confusion and will be engaged in solving this disequilibrium effectively.*
- ***Unproductive confusion*** *is the second stage of confusion: the disequilibrium state of the first stage of confusion is persistent and the impasse still cannot be solved during the interaction; then the interlocutor may become disengaged and may cease interacting with others and may enter negative emotions states (for example, frustration and boredom).*

In addition, the key difference between productive and unproductive confusion is that some attempts to overcome the confusion states failed. The action taken to overcome the

confusion state might be interaction, but it could also be taking time to reflect on the problem, looking for additional information or help *etc.*.

Study Design

This study is based around a semi-spontaneous one-to-one conversation between the Pepper robot and a user, based on my previous two HRI studies (HRI study 1 and further iterating study 2.2), I will explore different confusion states in this HRI study, that is, state A1 is intended to produce productive confusion, while state A1 is intended to produce unproductive confusion, while state B is non-confusion).

Table 5.1 Example of sequences of confusion stimuli in participants

Participant 1		
Dialogues	**Confusion Cause (CC)**	**States**
1st	CC1*	State B & State A1
2nd	CC2*	State B & State A1
3rd	CC3*	State B & State A1
4th	CC4*	State B & State A1
Participant 2		
Dialogues	**Confusion Cause (CC)**	**States**
1st	CC1*	State B & State A2
2nd	CC2*	State B & State A2
3rd	CC3*	State B & State A2
4th	CC4*	State B & State A2

* CC1: Complex information * CC2: Insufficient information
* CC3: Contradictory Information * CC4: False feedback

Unlike the previous studies, this study will focus on one task only (*i.e.*, a word problem); therefore, each participant will have four word problems that are designed, and each participant will be stimulated into two states (*i.e.*, non-confusion and productive confusion, or non-confusion and unproductive confusion) (see Table 5.1) by the four confusion causes. Figure 5.2 shows the mapping of the three confusion states with the four partitions of confusion causes, that is, complex information, insufficient information, contradictory information,

Table 5.2 A matrix of confusion causes and states of confusion for each task

Productive Confusion (State A1)	**Unproductive confusion (State A2)**	**Non-confusion (State B)**
Complex information		Simple information
Insufficient information		Sufficient information
Contradictory Information		Consistent Information
False feedback*		Correct-positive feedback

* False feedback includes false feedback and correct-negative feedback

and false feedback; Simple information, sufficient information, consistent information and correct-positive feedback are used for the non-confusion state.

Two post-questionnaires will be required for participants after interacting with the Pepper robot; the first questionnaire will be an 11 part-post interaction survey with a 5-level Likert scale questions; while the second questionnaire will be based on the standard survey "Your Thoughts About the Research" (Rubin, 2016), where there are four questions to measure the potential influence of demand behaviours in this study's situations (see Table B.4).

5.2 Mitigating User Confusion in Situated Human-Robot Interaction

Mitigating user confusion by addressing the last two sub-research questions (*i.e.*, RQ5 and RQ6), is a final goal for my PhD study. As I have collected the HRI data from those confusion detection experiments, applying generic feature models are the first step for confusion detection classification, and then modelling this classification is the second study against validating and evaluating the confusion detection classification as the final study. I begin by introducing the generic feature models.

5.2.1 Generic Feature Models

I have applied the pre-trained feature algorithms to analyse each unimodal data, *e.g.*, facial emotion estimation, head pose estimation, eye gazing estimation, *etc.*. There are a few limitations that I noticed: (a) the number of fear emotion has been detected is much more than other numbers of emotions in Study 4.3; and (b) There is no significant difference between the angles of pitch, yaw and roll of head pose and the two conditions in Study 4.3, *etc.*. Obviously, these existing algorithms are not satisfied for HRI study, specific confusion detection in HRI, it is necessary to retrain feature extraction models on each unimodal data. The unimodal data are from our multimodal HRI dataset and these public feature dataset. I plan to use self-supervised learning's pre-training to learn representations in the presence of variations and features on visual data for facial expression, eye gaze, and head pose from visual data (Spurr *et al.*, 2021; Roy and Etemad, 2021; Mahmud *et al.*, 2021), and on audio data for emotion speech recognition (Revathi *et al.*, 2022).

5.2.2 Confusion Detection Classification

The confusion detection classifier is a significant contribution that should detect three states of confusion in real-time human-robot interaction. Currently, most researchers have trained confusion level detection models with different types of classifiers (*e.g.*, decision tree, SVM, KNN, a random forest algorithm, logistic regression, a feed-forward neural network, recurrent neural networks and long short-term memory (LSTM)-based recurrent neural network (Hori *et al.*, 2016; Kavita Kelkar, 2021; Benlamine and Frasson, 2021)). Training data that they used was from online learning and driving systems without using a multimodal dataset. Despite the fact that the confusion level detection model that they trained with had four levels (*i.e.*, very low, low, medium, high, very high), they did not analyse and constrain which levels are productive confusion, unproductive confusion or even non-confusion.

This piece of work is intended to model confusion detection on the multimodal dataset (including audio data and visual data) from HRI Study 1, two future studies (*i.e.*, HRI Study 2.1 (Section 5.1.1) and HRI Study 2.2 (Section 5.1.2). Deep learning-based late fusion is one modelling strategy that may be of interest (Pandeya and Lee, 2021). All frame data and audio data will be labelled with confusion states. Two training phases will be approached: (1) feature extraction from mulitmodal data in the HRI studies; and (2) learning the integrative feature representation (see Section 5.2.1) to make the final prediction by integrating to synchronise the feature vectors and concatenate these synchronised unimodal feature vectors to build a confusion detection classification (Lee *et al.*, 2019).

5.2.3 Validated Confusion Mitigation in WoZ Study

To validate the final performance of confusion detection and explore the opportunities for mitigation, the situated conversations between the Pepper robot and a participant will be further explored. The study design will be based on two dialogues for two states of confusion stimulus that are either non-confusion and productive confusion, or non-confusion and unproductive confusion using the four confusion causes. It is worth mentioning that a key contribution of this work will be that different strategies of confusion mitigation will be designed to mitigate the confusion from each confusion cause, and I will apply these strategies to the Pepper dialogues and behaviours.

Figure 5.4 depicts the WoZ experiment laboratory setup. In the wizard room, I plan to equip a state-of-the-art device such as the NVIDIA Jetson TX2 device (Greco *et al.*, 2019b; Amert *et al.*, 2017) to receive images from the camera on the robot and the webcam. The confusion detection classifier will be integrated into the NVIDIA Jetson TX2 device as an example, using an HTTP protocol to monitor the confusion detection score in real-time. The researcher also controls the Pepper robot by using the WoZ4U interface which is an open-source WoZ interface for the wizard to control the Pepper movements, animated speech,

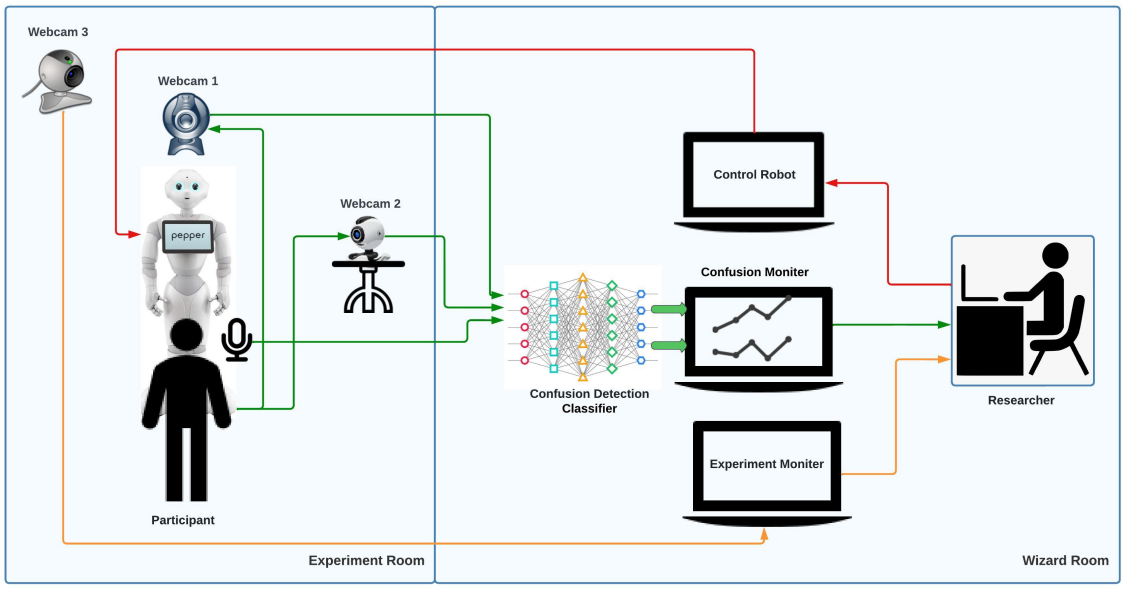


Fig. 5.4 Validation confusion mitigation in WoZ experiment

gestures *etc.* (Rietz *et al.*, 2021); Meanwhile, the researcher will monitor the experiment room. In the experiment room, three cameras are setup: Camera 1 is behind the Pepper robot and aimed toward the participant's face to collect the participant's facial expression; Camera 2 with a microphone is left next to the Pepper robot for the participant's speech and gesture collection; Camera 3 is for the researcher to monitor the interaction between the Pepper and the participant. The video data from Camera 1 and Camera 2 will be transferred to predict the confusion states online with pre-trained confusion detection classifier. Therefore, according to the confusion score and a specific confusion cause for one confusion stimuli, the researcher can adjust the Pepper robot's behaviours including body language and dialogues (more detail later) to validate whether the participant's confusion will be mitigated by the different states of confusion stimuli through observing the fluctuations of the confusion score. Finally, I will also apply the user survey and user feedback to evaluate the confusion mitigation approaches.

In addition, a linguistic design of dialogue policies is will be developed to allow me to design a dialogue framework for alleviating interlocutor confusion. I propose seven dialogue act types with a detailed general dialogue policy. Moreover, two sub-policies for the different

confusion states mitigation are designed respectively. Therefore, the seven dialogue act types are defined with relevance to the mitigation as follows:

1. **Restatement**: The agent repeats the information or question.
2. **Feedback request**: The agent asks for the participant's feedback and response.
3. **Information extension**: The agent provides more information to expand on the information or question already raised.
4. **Information supplement**: The agent provides comprehensive information or questions in different ways for participants to quickly understand easily.
5. **Response correction**: The agent provides the appropriate response in order to avoid confusion states on the participant.
6. **Confirmation**: The agent admits that the information or question has one or more issues leading to the participant being confused.
7. **Subject change**: The agent changes straightforward questions or other topics.

Following these dialogue act types, a general dialogue policy is designed based on a number of communicative rules (see Table 5.3). Figure 5.5 illustrates the operating dialogue policy as a control flow process, with each step corresponding to one of the detailed elements of the outline rules in Table 1. In this control flow policy, each step makes it possible to help users who are confused transfer to a non-confusion state. If after any one step, the user's confusion still cannot be mitigated, then the agent will move to the next step.

Based on this general framework policy, I am developing a set of sub-policies to apply in the specific cases of productive and unproductive confusion in the case of the four confusion induction types mentioned earlier. Table 5.4 outlines the first of these dialogue sub-policies that includes the dialogue act types and corresponding communication rules

Table 5.3 General dialogue policy for mitigating productive confusion

General dialogue policy of confusion mitigation	
Dialogue Acts	**Communication Rules**
Restatement	Repeat the information/question either at the same speed or more slowly.
Feedback request	Option 1: Ask the participant whether they can follow what the agent has said. Option 2: Ask the participant whether they want to continue to answer this question or complete the task with the agent's help.
Confirmation	Acknowledge that the information/question is difficult and that this has likely led to the participant being confused.
Information extension	Provide more explanations or information to fix the issued questions/information.
Information supplement	Provide the full information/question in different ways to easily understand without confusion.
Response correction	Provide a positive and correct response to remove the participant's source of confusion.
Subject change	Option 1: Raise a simple question that the participant can answer without confusion. Option 2: Raise another interesting topic arising the participant's engagement.

to reduce productive confusion according to the induction of a specific confusion method. The second sub-policy (shown in Table 5.5) addresses the case where the participant has reached an unproductive confusion state, where they may be frustrated or even want to drop the conversation. Therefore, this sub-policy helps the participant reengages in interacting with the agent from their unproductive confusion state.

Finally, I am building on this sketch to implement a physical test for those policies based on the WoZ experiment for validation confusion mitigation. I expect that this work can drive a true formalisation and evaluation of these policies. This proposal was presented in the following short paper at SemDial conference 2022:

- Li, N. and Ross, R. (2022) Dialogue Policies for Confusion Mitigation in Situated HRI in: 26th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2022) Technological University Dublin, Ireland

Table 5.4 Dialogue policy for mitigating productive confusion

Confusion Inductions	**Policy of confusion mitigation**	
	Dialogue Acts	**Communicative Detail**
Complex infor*	Restatement	The agent will reintroduce the complex information step by step.
	Feedback request	The agent will ask whether the participant is clear on the issue or question.
	Information supplement	The more and extra information will be told to the participant.
Insufficient infor	Information extension	The agent will provide the lost part of the question/information.
	Information supplement	The more and extra information will be told to the participant.
Contradictory infor	Response correction	The agent will show positive feedback.
	Confirmation	The agent will confirm the question/information is contradicted.
	Information extension	The agent will correct the question/information to consistent information.
	Information supplement	The more and extra information will be told to the participant.
False Feedback	Response correction	The agent will show positive and correct response or feedback.
	Subject change	The agent will talk about a simple question/information with positive feedback.

* Infor: Information

Table 5.5 Dialogue policy to mitigate unproductive confusion

Confusion Inductions	**Policy of confusion mitigation**	
	Dialogue Acts	**Communicative Detail**
Complex infor* Insufficient infor Contradictory infor False Feedback	Confirmation	With a positive response, the agent will confirm that the question/information is difficult which has led to the participant being confused.
	Feedback request	The agent will ask whether participants want to continue to answer this question or to continue the task with the agent's help.
	Subject change	The agent will ask straightforward questions to avoid confusion states.
		The agent will talk about another interesting topic *e.g.*, favourite food, movie *etc.*

* Infor: Information

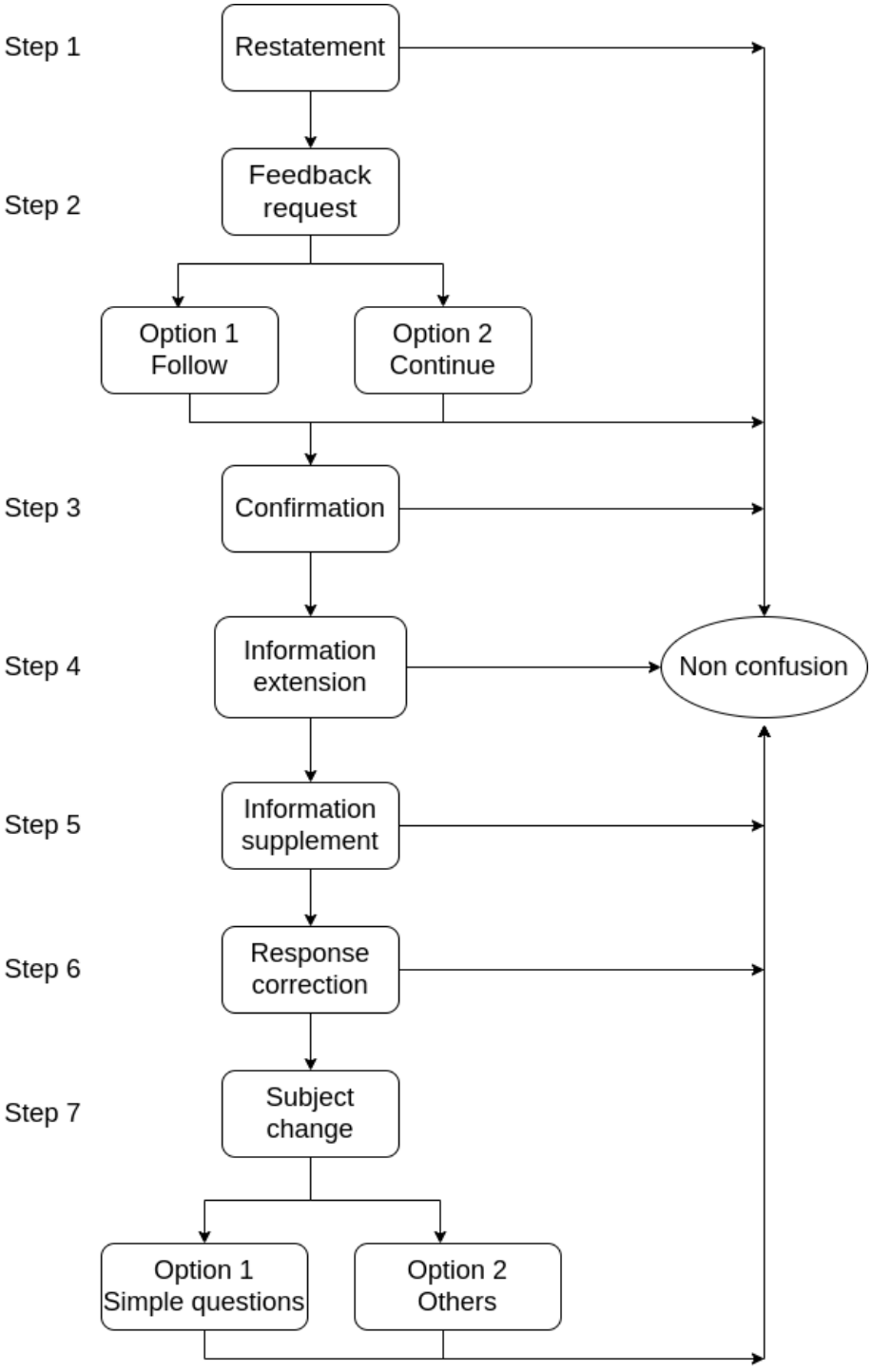


Fig. 5.5 General policy process of confusion mitigation

Chapter 6

Conclusion

This report illustrated the three studies conducted to date, and the studies planned to further explore the research topic of confusion detection in conversational HRI. First, I defined what confusion is in situated dialogue HRI as a guideline for designing the first two HAI and HRI studies based on four confusion causes. Meanwhile, feature analysis algorithms were used to explore different nonverbal and verbal human behaviours in confusion or non-confusion states. Next, a contrastive analysis of the reactions of users across the avatar and the physical robot embodiment types was conducted to investigate whether it remains feasible to leverage human-avatar data for human-robot interaction when the focus of these early two studies is on communication rather than, for example, physical cooperation.

As for the future study planning, two main further studies have been designed: To continue exploring confusion in HRI against limitations that were identified from the work-done-to-date, an improved confusion detection study based on HRI Study 1 was outlined, a new experimental study for detection of different states of confusion in HRI was also proposed. Another piece of future work is modelling and validating confusion detection classification along with generic feature algorithms to evaluate whether my study can ultimately improve the conversational interaction between a human and a robot.

The expected contributions of this research to the HRI community as a whole are as follows:

- A definition of confusion and two sub-definitions of states of confusion (*i.e.*, productive confusion and unproductive confusion).
- A Real-Time Online Chat platform is an open-source WoZ interface that integrates an avatar and provides a graphical user interface (GUI) for the wizard to control the avatar's interactive behaviours.
- Designs of interactive experimental scripts for confusion stimuli in HAI and HRI studies, respectively.
- Open-source interactive and animated speech code on the Pepper robot for stimuli confusion and non-confusion.
- Methods of multimodal data collection for human confusion analysis in two embodied modalities.
- Multimodal Datasets in HAI and HRI studies.
- A reference for comparison of two embodiments in the approach of confusion detection study.
- Unimodal data analysis for generic feature models.
- Mulitmodal deep learning model for confusion states detection.
- Application of confusion states mitigation in real-time HRI.
- Dialogue policies for confusion mitigation in conversational HRI.

The achievements of this research to date, with my supervisor, have one long paper and two round-table workshop position papers, which have been introduced in previous chapters.

In summary, the long paper was presented in an oral presentation format at the SemDial2021 venue, a short paper was presented orally at the MMAI2022 Workshop, and a short paper was posted at the SemDial2022 venue.

- Li, N., Kelleher, J.D. and Ross, R. (2021) Detecting interlocutor confusion in situated human-avatar dialogue: A pilot study in: 25th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2021) University of Potsdam, Germany
- Li, N. and Ross, R. (2022) Transferring studies across embodiments: A case study in confusion detection in: 1st workshop (MMAI2022) that is a part of the conference on Hybrid Human-Artificial Intelligence 2022, Amsterdam, Netherlands
- Li, N. and Ross, R. (2022) Dialogue Policies for Confusion Mitigation in Situated HRI in: 26th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2022) Technological University Dublin, Ireland

Finally, Figure 6.1 shows my study planning of this work including the future works, the submission planning of the paper, and the final Ph.D. thesis planning.

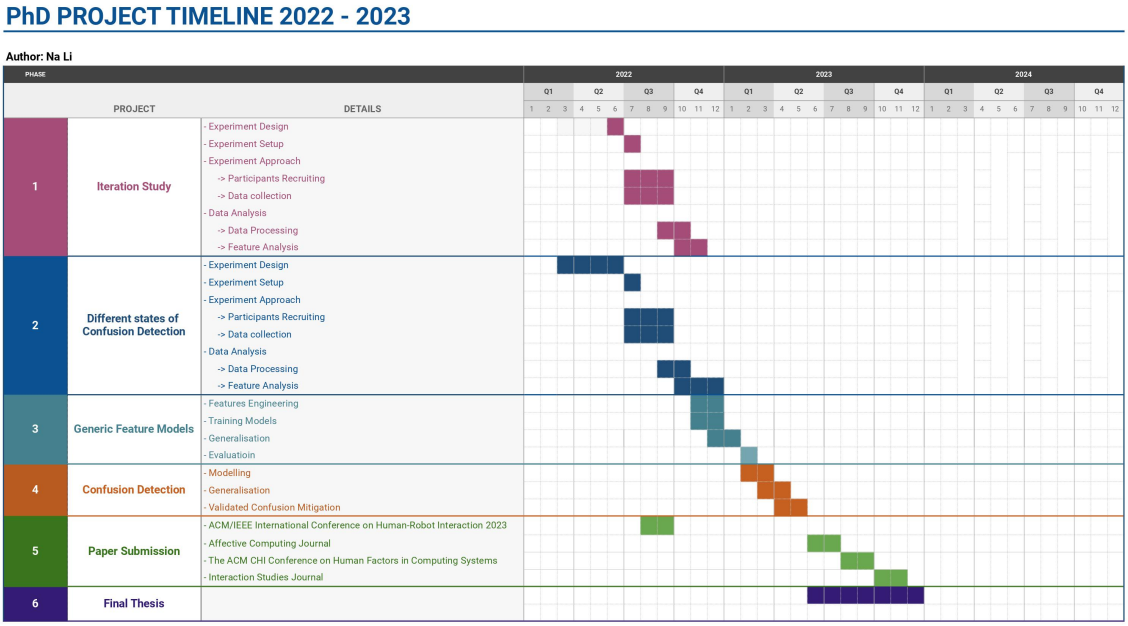


Fig. 6.1 A study planing until to the year of 2023

References

Abbas, T., Khan, V.J. and Markopoulos, P. (2020) Coz: A crowd-powered system for social robotics *SoftwareX* **11**, p. 100421

Admoni, H. and Scassellati, B. (2017) Social eye gaze in human-robot interaction: A review *J. Hum.-Robot Interact.* **6**(1), p. 25–63

Alonso-Martin, F., Castro-González, Á., Gorostiza, J.F. and Salichs, M.A. (2013) Multidomain voice activity detection during human-robot interaction in: *Social Robotics*, (Eds.) G. Herrmann, M.J. Pearson, A. Lenz, P. Bremner, A. Spiers and U. Leonards pp. 64–73 Springer International Publishing, Cham

Amert, T., Otterness, N., Yang, M., Anderson, J.H. and Smith, F.D. (2017) Gpu scheduling on the nvidia tx2: Hidden details revealed in: *2017 IEEE Real-Time Systems Symposium (RTSS)* pp. 104–115

Applewhite, T., Zhong, V.J. and Dornberger, R. (2021) Novel bidirectional multimodal system for affective human-robot engagement in: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* pp. 1–7

Arguel, A. and Lane, R. (2015) Fostering deep understanding in geography by inducing and managing confusion: An online learning approach *ASCILITE 2015 - Australasian Society for Computers in Learning and Tertiary Education, Conference Proceedings* (November), pp. 374–378

Arguel, A., Lockyer, L., Lipp, O.V., Lodge, J.M. and Kennedy, G. (2017) Inside out: Detecting learners' confusion to improve interactive digital learning environments *Journal of Educational Computing Research* **55**(4), pp. 526–551

Arriaga, O., Valdenegro-Toro, M. and Plöger, P. (2017) Real-time convolutional neural networks for emotion and gender classification

Aviezer, H., Trope, Y. and Todorov, A. (2012) Body cues, not facial expressions, discriminate between intense positive and negative emotions *Science* **338**(6111), pp. 1225–1229

Baker, S., Waycott, J., Carrasco, R., Kelly, R.M., Jones, A.J., Lilley, J., Dow, B., Batchelor, F., Hoang, T. and Vetere, F. (2021) Avatar-mediated communication in social vr: An in-depth exploration of older adult interaction in an emerging communication platform in: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* CHI '21 Association for Computing Machinery, New York, NY, USA

Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M. and Pollak, S.D. (2019) Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements *Psychological Science in the Public Interest* **20**(1), pp. 1–68 pMID: 31313636

Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M. and Šabanović, S. (2020) *References* Cambridge University Press

Bastianelli, E., Castellucci, G., Croce, D., Basili, R. and Nardi, D. (2014) Effective and robust natural language understanding for human-robot interaction in: *Proceedings of the Twenty-First European Conference on Artificial Intelligence* ECAI'14 p. 57–62 IOS Press, NLD

Bates, M. (1995) Models of natural language understanding. *Proceedings of the National Academy of Sciences* **92**(22), pp. 9977–9982

Ben Youssef, A., Clavel, C., Essid, S., Bilac, M., Chamoux, M. and Lim, A. (2017) UE-HRI: A new dataset for the study of user engagement in spontaneous human-robot interactions pp. 464–472

Ben Youssef, A., Varni, G., Essid, S. and Clavel, C. (2019) On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks *International Journal of Social Robotics* **11**

Benlamine, M.S. and Frasson, C. (2021) Confusion detection within a 3d adventure game in: *Intelligent Tutoring Systems*, (Eds.) A.I. Cristea and C. Troussas pp. 387–397 Springer International Publishing, Cham

Brabra, H., Báez, M., Benatallah, B., Gaaloul, W., Bouguelia, S. and Zamanirad, S. (2021) Dialogue management in conversational systems: A review of approaches, challenges, and opportunities *IEEE Transactions on Cognitive and Developmental Systems* pp. 1–1

Braun, D., Hernandez Mendez, A., Matthes, F. and Langen, M. (2017) Evaluating natural language understanding services for conversational question answering systems in: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* pp. 174–185 Association for Computational Linguistics, Saarbrücken, Germany

Breazeal, C. (2004) Social interactions in hri: the robot view *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **34**(2), pp. 181–186

Breazeal, C. and Velásquez, J.D. (1999) Toward teaching a robot "infant" using emotive communication acts

Brennan, S.E. (2010) *Conversation and dialogue* SAGE

Bruno, B., Chong, N.Y., Kamide, H., Kanoria, S., Lee, J., Lim, Y., Pandey, A.K., Papadopoulos, C., Papadopoulos, I., Pecora, F., Saffiotti, A. and Sgorbissa, A. (2017) The CARESSES eu-japan project: making assistive robots culturally competent *CoRR* **abs/1708.06276**

Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. and Narayanan, S.S. (2008) IEMOCAP: interactive emotional dyadic motion capture database *Language Resources and Evaluation* **42**(4), pp. 335–359

Cantrell, R., Scheutz, M., Schermerhorn, P. and Wu, X. (2010) Robust spoken instruction understanding for hri in: *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* pp. 275–282

Cassell, J. (2001) Nudge nudge wink wink: elements of face-to-face conversation for embodied conversational agents

Cassell, J. and Vilhjálmsson, H. (2004) Fully embodied conversational avatars: Making communicative behaviors autonomous *Autonomous Agents and Multi-Agent Systems* **2**, pp. 45–64

Celiktutan, O., Skordos, S. and Gunes, H. (2017) Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement

Chang, J.H., Kim, N.S. and Mitra, S. (2006) Voice activity detection based on multiple statistical models *IEEE Transactions on Signal Processing* **54**(6), pp. 1965–1976

Cohn, J.F. (2007) Foundations of human computing: Facial expression and emotion in: *Artifical Intelligence for Human Computing*, (Eds.) T.S. Huang, A. Nijholt, M. Pantic and A. Pentland pp. 1–16 Springer Berlin Heidelberg, Berlin, Heidelberg

Crovari, P., Pidó, S., Garzotto, F. and Ceri, S. (2021) Show, don't tell. reflections on the design of multi-modal conversational interfaces in: *Chatbot Research and Design*, (Eds.) A. Følstad, T. Araujo, S. Papadopoulos, E.L.C. Law, E. Luger, M. Goodwin and P.B. Brandtzaeg pp. 64–77 Springer International Publishing, Cham

Dewan, M.A., Murshed, M. and Lin, F. (2018) Engagement detection in online learning: a review *Smart Learning Environments* **6**, pp. 1–20

D'Mello, S. and Graesser, A. (2014) Confusion and its dynamics during device comprehension with breakdown scenarios *Acta Psychologica* **151**, pp. 106–116

D'Mello, S., Lehman, B., Pekrun, R. and Graesser, A. (2014) Confusion can be beneficial for learning *Learning and Instruction* **29**, pp. 153–170

Doherty, K. and Doherty, G. (2018a) Engagement in hci: Conception, theory and measurement *ACM Comput. Surv.* **51**(5)

Doherty, K. and Doherty, G. (2018b) Engagement in hci: Conception, theory and measurement *ACM Comput. Surv.* **51**(5)

Duchetto, F., Baxter, P. and Hanheide, M. (2019) Lindsey the tour guide robot - usage patterns in a museum long-term deployment pp. 1–8

Emery, N. (2000) The eyes have it: the neuroethology, function and evolution of social gaze *Neuroscience & Biobehavioral Reviews* **24**(6), pp. 581–604

Esterwood, C. and Robert, L. (2020) Personality in healthcare human robot interaction (h-hri): A literature review and brief critique

Fadhil, A., Schiavo, G., Wang, Y. and Yilma, B.A. (2018) The effect of emojis when interacting with conversational interface assisted health coaching system in: *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare* PervasiveHealth '18 p. 378–383 Association for Computing Machinery, New York, NY, USA

Fischer, K., Jung, M., Jensen, L.C. and aus der Wieschen, M.V. (2019) Emotion expression in hri – when and why in: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* pp. 29–38

Forsberg, M. (2003) Why is speech recognition difficult?

Geller, S.A., Gal, K., Segal, A., Sripathi, K., Kim, H.G., Facciotti, M.T., Igo, M., Hoernle, N. and Karger, D. (2021) New methods for confusion detection in course forums: Student, teacher, and machine *IEEE Transactions on Learning Technologies* **14**(5), pp. 665–679

Ginevra Castellano, Hatice Gunes, C.P. and W.Schuller, B. (2004) Multimodal affect recognition for naturalistic human-computer and human-robot interactions *The Oxford Handbook of Affective Computing, New York, NY: Oxford University Press* pp. 246–257

Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z. and Bengio, Y. (2013) Challenges in representation learning: A report on three machine learning contests

Gordon, G., Spaulding, S., Westlund, J.K., Lee, J.J., Plummer, L., Martinez, M., Das, M. and 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 p. 3951–3957 AAAI Press

Gourier, N., Hall, D. and Crowley, J. (2004) Estimating face orientation from robust detection of salient facial structures

Grafsgaard, J.F., Boyer, K.E. and Lester, J.C. (2011) Predicting Facial Indicators of Confusion with Hidden Markov Models Technical report

Greco, A., Roberto, A., Saggese, A., Vento, M. and Vigilante, V. (2019a) Emotion analysis from faces for social robotics in: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* pp. 358–364

Greco, A., Roberto, A., Saggese, A., Vento, M. and Vigilante, V. (2019b) Emotion analysis from faces for social robotics in: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* pp. 358–364

Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.P. and Zimmermann, R. (2018) Conversational memory network for emotion recognition in dyadic dialogue videos p. 2122–2132

Heerink, M., Kröse, B., Evers, V. and Wielinga, B. (2010) Assessing acceptance of assistive social agent technology by older adults: the almere model **2**

Herath, D.C., Binks, N. and Grant, J.B. (2020) To embody or not: A cross human-robot and human-computer interaction (hri/hci) study on the efficacy of physical embodiment in: *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)* pp. 848–853

Heyselaar, E., Hagoort, P. and Segaert, K. (2017) In dialogue with an avatar, language behavior is identical to dialogue with a human partner **49**(1), pp. 46–60

Hirschman, L. and Gaizauskas, R. (2001) Natural language question answering: The view from here *Nat. Lang. Eng.* **7**(4), p. 275–300

Hoffman, G., Cakmak, M. and Chao, C. (2014) Timing in human-robot interaction in: *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* HRI '14 p. 509–510 Association for Computing Machinery, New York, NY, USA

Hoffmann, L. and Krämer, N.C. (2013) Investigating the effects of physical and virtual embodiment in task-oriented and conversational contexts *International Journal of Human-Computer Studies* **71**(7), pp. 763–774

Hori, C., Watanabe, S., Hori, T., Harsham, B.A., Hershey, J., Koji, Y., Fujii, Y. and Furumoto, Y. (2016) Driver confusion status detection using recurrent neural networks in: *2016 IEEE International Conference on Multimedia and Expo (ICME)* pp. 1–6

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications *CoRR* **abs/1704.04861**

Hussen Maulud, D., Zeebaree, S.R.M., Jacksi, K., Mohammed Sadeeq, M.A. and Hussein Sharif, K. (2021) State of art for semantic analysis of natural language processing *Qubahan Academic Journal* **1**(2), p. 21–28

Ibrahim, F., Mutashar, S. and Hamed, B. (2021) A review of an invasive and non-invasive automatic confusion detection techniques *IOP Conference Series: Materials Science and Engineering* **1105**(1), p. 012026

Ikeuchi, T., Sakurai, R., Furuta, K., Kasahara, Y., Imamura, Y. and Shinkai, S. (2018) Utilizing social robot to reduce workload of healthcare professionals in psychiatric hospital: a preliminary study *Innovation in Aging* **2**(1), pp. 695–696

Jaimes, A., Lalmas, M. and Volkovich, Y. (2011) First international workshop on social media engagement (some 2011) in: *Proceedings of the 20th International Conference Companion on World Wide Web* WWW '11 p. 309–310 Association for Computing Machinery, New York, NY, USA

Jam, G.S., Rhim, J. and Lim, A. (2021) Developing a data-driven categorical taxonomy of emotional expressions in real world human robot interactions *CoRR* **abs/2103.04262**

Jarvela, S. (2011) *Social and emotional aspects of learning* Elsevier

Jo, W., Kannan, S.S., Cha, G., Lee, A. and Min, B. (2020) Rosbag-based multimodal affective dataset for emotional and cognitive states *CoRR* **abs/2006.05102**

Kavita Kelkar, J.B. (2021) Random forest algorithm for learner's confusion detection using behavioral features in: *International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*

Khan, M.Q. and Lee, S. (2019) Gaze and eye tracking: Techniques and applications in adas *Sensors* **19**(24)

Kidd, C. and Breazeal, C. (2004) Effect of a robot on user perceptions in: *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)* vol. 4 pp. 3559–3564 vol.4

Kleinsmith, A. and Bianchi-Berthouze, N. (2013) Affective body expression perception and recognition: A survey *IEEE Transactions on Affective Computing* **4**(1), pp. 15–33

Kontogiorgos, D., Pereira, A., Sahindal, B., Waveren, S.v. and Gustafson, J. (2020) Behavioural responses to robot conversational failures in: *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* pp. 53–62

Kozima, H., Michalowski, M.P. and Nakagawa, C. (2009a) Keepon: A playful robot for research, therapy, and entertainment *International Journal of Social Robotics* **1**(1)

Kozima, H., Michalowski, M.P. and Nakagawa, C. (2009b) A playful robot for research, therapy, and entertainment

Kumar, H., Sethia, M., Thakur, H., Agrawal, I. and P, S. (2019) Electroencephalogram with Machine Learning for Estimation of Mental Confusion Level *International Journal of Engineering and Advanced Technology* **9**(2), pp. 761–765

Kurzon, D. (1998) *Discourse of Silence* John Benjamins

Köstinger, M., Wohlhart, P., Roth, P.M. and Bischof, H. (2011) Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization in: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* pp. 2144–2151

Lee, G., Kang, B., Nho, K., Sohn, K.A. and Kim, D. (2019) Mildint: Deep learning-based multimodal longitudinal data integration framework *Frontiers in Genetics* **10**

Lee, K.M., Peng, W., Jin, S.A. and Yan, C. (2006) Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction *Journal of Communication* **56**(4), pp. 754–772

Lehman, B., D'Mello, S. and Graesser, A. (2012) Confusion and complex learning during interactions with computer learning environments *The Internet and Higher Education* **15**(3), pp. 184–194 emotions in online learning environments

Lehman, B.A., D'Mello, S.K. and Graesser, A.C. (2013) Who benefits from confusion induction during learning? an individual differences cluster analysis in: *AIED*

Lehmann, H. and Svarny, P. (2021) Using a social robot for different types of feedback during university lectures *Education Sciences & Society - Open Access* **12**(2)

Lian, Z., Liu, B. and Tao, J. (2021) Decn: Dialogical emotion correction network for conversational emotion recognition *Neurocomputing* **454**, pp. 483–495

Lodge, J.M., Kennedy, G., Lockyer, L., Arguel, A. and Pachman, M. (2018) Understanding Difficulties and Resulting Confusion in Learning: An Integrative Review *Frontiers in Education* **3**

Lubitz, A., Valdenegro-Toro, M. and Kirchner, F. (2021) The VVAD-LRS3 dataset for visual voice activity detection *CoRR* **abs/2109.13789**

Lukosch, H., Lukosch, S., Hoermann, S. and Lindeman, R.W. (2019) Conceptualizing fidelity for hci in applied gaming in: *HCI in Games*, (Ed.) X. Fang pp. 165–179 Springer International Publishing

Mahmud, Z., Hungler, P. and Etemad, A. (2021) Gaze estimation with eye region segmentation and self-supervised multistream learning *CoRR* **abs/2112.07878**

Mavridis, N. (2015) A review of verbal and non-verbal human–robot interactive communication *Robotics and Autonomous Systems* **63**, pp. 22–35

McNamara, N. and Kirakowski, J. (2006) Functionality, usability, and user experience: Three areas of concern *Interactions* **13**(6), p. 26–28

Menne, I.M. and Lugrin, B. (2017) In the face of emotion: A behavioral study on emotions towards a robot using the facial action coding system in: *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* HRI '17 p. 205–206 Association for Computing Machinery, New York, NY, USA

Mollahosseini, A., Hassani, B. and Mahoor, M.H. (2017) Affectnet: A database for facial expression, valence, and arousal computing in the wild *CoRR* **abs/1708.03985**

Morales, C.G., Carter, E.J., Tan, X.Z. and Steinfeld, A. (2019) Interaction needs and opportunities for failing robots in: *Proceedings of the 2019 on Designing Interactive Systems Conference* DIS '19 p. 659–670 Association for Computing Machinery, New York, NY, USA

Mubin, O., Henderson, J. and Bartneck, C. (2014) You just do not understand me! speech recognition in human robot interaction in: *The 23rd IEEE International Symposium on Robot and Human Interactive Communication* pp. 637–642

Murphy-Chutorian, E. and Trivedi, M.M. (2009) Head pose estimation in computer vision: A survey *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(4), pp. 607–626

Nassif, A.B., Shahin, I., Attili, I., Azzeh, M. and Shaalan, K. (2019) Speech recognition using deep neural networks: A systematic review *IEEE Access* **7**, pp. 19143–19165

Nehaniv, C., Dautenhahn, K., Kubacki, J., Haegele, M., Parlitz, C. and Alami, R. (2005) A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction in: *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* pp. 371–377

Nourbakhsh, I.R., Bobenage, J., Grange, S., Lutz, R., Meyer, R. and Soto, A. (1999) An affective mobile robot educator with a full-time job *Artificial Intelligence* **114**(1), pp. 95–124

Novoa, J., Mahu, R., Wuth, J., Escudero, J.P., Fredes, J. and Yoma, N.B. (2021) Automatic speech recognition for indoor hri scenarios *J. Hum.-Robot Interact.* **10**(2)

Novoa, J., Wuth, J., Escudero, J.P., Fredes, J., Mahu, R. and Yoma, N.B. (2018) Dnn-hmm based automatic speech recognition for hri scenarios in: *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* pp. 150–159

O'Brien, H.L. and Toms, E. (2008) What is user engagement? a conceptual framework for defining user engagement with technology *J. Assoc. Inf. Sci. Technol.* **59**, pp. 938–955

Ohshima, N., Kimijima, K., Yamato, J. and Mukawa, N. (2015) A conversational robot with vocal and bodily fillers for recovering from awkward silence at turn-takings in: *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* pp. 325–330

Omar Mubin, C.J.S., Shahid, S., Mahmud, A.A. and Dong, J.J. (2016) A review on the use of robots in education and young children *Educational Technology & Society* **19**(2), p. 148–163

Pachman, M., Arguel, A., Lockyer, L., Kennedy, G. and Lodge, J.M. (2016) Eye tracking and early detection of confusion in digital learning environments: Proof of concept Technical Report 6

Pan, Y. and Steed, A. (2016) A comparison of avatar-, video-, and robot-mediated interaction on users' trust in expertise *Frontiers in Robotics and AI* **3**

Pandeya, Y.R. and Lee, J. (2021) Deep learning-based late fusion of multimodal information for emotion classification of music video *Multimedia Tools Appl.* **80**(2), p. 2887–2905

Patacchiola, M. and Cangelosi, A. (2017) Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods *Pattern Recognition* **71**, pp. 132–143

Picard, R.W. (2003) Affective computing: challenges *International Journal of Human-Computer Studies* **59**(1), pp. 55–64 applications of Affective Computing in Human-Computer Interaction

Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017) A review of affective computing: From unimodal analysis to multimodal fusion *Information Fusion* **37**, pp. 98–125

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E. and Mihalcea, R. (2019) MELD: A multimodal multi-party dataset for emotion recognition in conversations in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* pp. 527–536 Association for Computational Linguistics, Florence, Italy

Pranto, S.I., Nabid, R.A., Samin, A.M., Mohammed, N., Sarker, F., Huda, M.N. and Mamun, K.A. (2021) Human-robot interaction in bengali language for healthcare automation integrated with speaker recognition and artificial conversational entity in: *2021 3rd International Conference on Electrical Electronic Engineering (ICEEE)* pp. 13–16

Prasad, V., Stock-Homburg, R. and Peters, J. (2020) Advances in human-robot handshaking in: *Social Robotics*, (Eds.) A.R. Wagner, D. Feil-Seifer, K.S. Haring, S. Rossi, T. Williams, H. He and S. Sam Ge pp. 478–489 Springer International Publishing, Cham

Pustejovsky, J. and Krishnaswamy, N. (2021) The role of embodiment and simulation in evaluating hci: Theory and framework in: *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior*, (Ed.) V.G. Duffy pp. 288–303 Springer International Publishing, Cham

Rajapakshe, T., Latif, S., Rana, R., Khalifa, S. and Schuller, B. (2020) Deep reinforcement learning with pre-training for time-efficient training of automatic speech recognition

Rajapakshe, T., Rana, R., Latif, S., Khalifa, S. and Schuller, B. (2019) Pre-training in deep reinforcement learning for automatic speech recognition *ArXiv* **abs/1910.11256**

Refat, C.M.M. and Azlan, N.Z. (2019) Deep learning methods for facial expression recognition in: *2019 7th International Conference on Mechatronics Engineering (ICOM)* pp. 1–6

Revathi, A., Neharika, B. and G, G. (2022) Emotion recognition from speech using multiple features and clusters in: *Evolution in Computational Intelligence*, (Eds.) V. Bhateja, J. Tang, S.C. Satapathy, P. Peer and R. Das pp. 265–277 Springer Nature Singapore, Singapore

Rho, D., Park, J. and Ko, J.H. (2022) NAS-VAD: neural architecture search for voice activity detection *CoRR* **abs/2201.09032**

Riek, L. (2012) Wizard of oz studies in hri: a systematic review and new reporting guidelines in: *HRI 2012*

Rietz, F., Sutherland, A., Bensch, S., Wermter, S. and Hellström, T. (2021) Woz4u: An open-source wizard-of-oz interface for easy, efficient and robust hri experiments *Frontiers in Robotics and AI* **8**

Robins, B., Dickerson, P., Stribling, P. and Dautenhahn, K. (2004) Robot-mediated joint attention in children with autism : A case study in robot-human interaction *Interaction Studies* **5**, pp. 161–198

Romeo, M., Cangelosi, A. and Jones, R. (2018) Developing a deep learning agent for hri: Dataset collection and training pp. 1150–1155

Roy, S. and Etemad, A. (2021) *Self-Supervised Contrastive Learning of Multi-View Facial Expressions* p. 253–257 Association for Computing Machinery, New York, NY, USA

Rubin, M. (2016) The perceived awareness of the research hypothesis scale: Assessing the influence of demand characteristics

Salloum, S.A., Khan, R. and Shaalan, K. (2020) A survey of semantic analysis approaches in: *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, (Eds.) A.E. Hassanien, A.T. Azar, T. Gaber, D. Oliva and F.M. Tolba pp. 61–70 Springer International Publishing, Cham

Samani, C. and Goyal, M. (2021a) Confusion detection using neural networks in: *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* pp. 1–6

Samani, C. and Goyal, M. (2021b) Modeling student confusion using fuzzy logic in e-learning *Proceedings of International Conference on Communication and Computational Technologies. Algorithms for Intelligent Systems.*

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P.W. and Paiva, A. (2011) Automatic analysis of affective postures and body motion to detect engagement with a game companion in: *Proceedings of the 6th International Conference on Human-Robot Interaction* HRI '11 p. 305–312 Association for Computing Machinery, New York, NY, USA

Savchenko, A.V. (2021) Facial expression and attributes recognition based on multi-task learning of lightweight neural networks *CoRR* **abs/2103.17107**

Sharkawy, A. (2021) Human-robot interaction: Applications *CoRR* **abs/2102.00928**

Sidner, C.L., Kidd, C.D., Lee, C. and Lesh, N. (2004) Where to look: A study of human-robot engagement in: *Proceedings of the 9th International Conference on Intelligent User Interfaces* IUI '04 p. 78–84 Association for Computing Machinery, New York, NY, USA

Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N. and Rich, C. (2005) Explorations in engagement for humans and robots *Artificial Intelligence* **166**(1), pp. 140–164

Silvia, P. (2010) Confusion and interest: The role of knowledge emotions in aesthetic experience *Psychology of Aesthetics, Creativity, and the Arts* **4**, pp. 75–80

Singh, A.P., Nath, R. and Kumar, S. (2018) A survey: Speech recognition approaches and techniques in: *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* pp. 1–4

Singh, K.J., Kapoor, D.S. and Sohi, B.S. (2021) Chapter 11 - all about human-robot interaction in: *Cognitive Computing for Human-Robot Interaction*, (Eds.) M. Mittal, R.R. Shah and S. Roy Cognitive Data Science in Sustainable Computing pp. 199–229 Academic Press

Skantze, G. (2021) Turn-taking in conversational systems and human-robot interaction: A review *Computer Speech & Language* **67**, p. 101178

Sloan, J., Maguire, D. and Carson-Berndsen, J. (2020) Emotional response language education for mobile devices in: *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* MobileHCI '20 Association for Computing Machinery, New York, NY, USA

Song, C.S. and Kim, Y.K. (2022) The role of the human-robot interaction in consumers' acceptance of humanoid retail service robots *Journal of Business Research* **146**, pp. 489–503

Spezialetti, M., Placidi, G. and Rossi, S. (2020) Emotion recognition for human-robot interaction: Recent advances and future perspectives *Frontiers in Robotics and AI* **7**, p. 145

Spurr, A., Dahiya, A., Wang, X., Zhang, X. and Hilliges, O. (2021) Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* pp. 11230–11239

Tapus, A., Peca, A., Amir, A., Pop, C., Jisa, L., Pintea, S., Rusu, A. and David, D. (2012) Children with autism social engagement in interaction with nao, an imitative robot – a series of single case experiments *Interaction Studies* **13**

Tran, N., Mizuno, K., Grant, T., Phung, T., Hirshfield, L. and Williams, T. (2020) Exploring mixed reality robot communication under different types of mental workload *International Workshop on Virtual, Augmented, and Mixed Reality for Human-Robot Interaction* **3**

Tripathi, S. and Beigi, H.S.M. (2018) Multi-modal emotion recognition on IEMOCAP dataset using deep learning *CoRR* **abs/1804.05788**

Tseng, B., Cheng, J., Fang, Y. and Vandyke, D. (2020) A generative model for joint natural language understanding and generation *CoRR* **abs/2006.07499**

VanLEHN, K. (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems *Educational Psychologist* **46**(4), pp. 197–221

Vrtika, P., Lordier, L., Bediou, B. and Sander, D. (2014) Human amygdala response to dynamic facial expressions of positive and negative surprise *Emotion* **14 1**, pp. 161–9

Wainer, J., Feil-seifer, D.J., Shell, D.A. and 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

Wei, H. (2016) When hci meets hri: the intersection and distinction *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* **1**, pp. 1–8

Williams, J.D. and Young, S. (2007) Partially observable markov decision processes for spoken dialog systems *Computer Speech & Language* **21**(2), pp. 393–422

Xu, J., Broekens, J., Hindriks, K. and Neerincx, M.A. (2014) Robot mood is contagious: Effects of robot body language in the imitation game AAMAS '14 p. 973–980 International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC

Yang, D., Kraut, R.E., Rosé, C.P. and Rosé, R. (2015) Exploring the effect of student confusion in massive open online courses Technical report

Zhang, K., Zhang, Z., Li, Z. and Qiao, Y. (2016) Joint face detection and alignment using multitask cascaded convolutional networks *IEEE Signal Processing Letters* **23**(10), pp. 1499–1503

Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S. and Hilliges, O. (2020) Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation in: *European Conference on Computer Vision (ECCV)*

Zhao, Y.J., Li, Y.L. and Lin, M. (2019) A review of the research on dialogue management of task-oriented systems *Journal of Physics: Conference Series* **267**(1), p. 012025

Zhou, Y., Xu, T., Li, S. and Shi, R. (2019) Beyond engagement: an EEG-based methodology for assessing user's confusion in an educational game *Universal Access in the Information Society* **18**(3), pp. 551–563

Zhu, X. and Ramanan, D. (2012) Face detection, pose estimation, and landmark localization in the wild in: *2012 IEEE Conference on Computer Vision and Pattern Recognition* pp. 2879–2886

Appendix A

Dialogue Design

A.1 Dialogue scripts for confusion A and confusion B stimuli

This table shows detail dialogues scripts for the agent (Avatar / Pepper) for HAI Study and HRI Study 1.

Table A.1 Conversation scripts for confusion stimuli

Tasks*	**Condition A**	**Condition B**
1	A*: Suppose Anna's mother admires Anna, Anna admires her mother, everyone admires her mother, so everyone admires Anna, right? B*: <user-response> A: Does it make sense Anna's friend admires Anna but her mother, doesn't it? B: <user-response> A: Thank you for your answer.	A: Suppose everyone over the age of 30 is a liar, William is a liar, so the question is, is William over 30? B: <user-response> A: Do you agree that not everyone under 30 is not a liar? B: <user-response> A: Great, you are correct.
2	A: There are 66 people in the playground including 28 girls, boys and teachers. How many teachers were there in total? B: <user-response> A: Please try again. B: <user-response> A: Thank you for your answer.	A: There are 5 groups of 4 students, how many students are there in the class? B: <user-response> A: You are correct. A: Each group has 2 pairs of scissors, how many pairs of scissors are there in total? B: <user-response> A: Well done, you are so smart.
3	A: If x = 4 and x + b + log(1) = 10, the question is, is $b = 6$ or $b = 12$? B: <user-response> A: Please try again. B: <user-response> A: Sorry, maybe this question is too difficult.	A: If $x = 4$ and $x + b = 10$, the question is, is b equal 12? B: <user-response> A: Great, you are correct.

* A: Agent; B: Participant

* Task 1: logic problem, Task 2: word problem, Task 3: math question

Appendix B

User survey

B.1 A user survey for HAI Study (see Table B.1)

Table B.1 User Survey for HAI Study

No.	Questions
1	Did you enjoy talking to Julia overall?
2	Was the conversation with Julia fluent?
3	Was the conversation with Julia easy?
4	Was the conversation with Julia frustrating?
5	Was the conversation with Julia boring?
6	Did you feel confused most of the time talking with Julia?
7	Did you feel confused most of the time when you answered logical questions to Julia?
8	Did you feel confused most of the time when you answered word problems to Julia (Including Julia's responses may make you confused)?
9	Did you feel confused most of the time when you answered Mathematics questions to Julia?
10	Did you want to drop this conversation with Julia when you were confused continually?

B.1.1 User Surveys for HRI Studies

Table B.2 User Survey for HRI Study 1

No.	Questions
1	Did you enjoy talking to Pepper overall?
2	Was the conversation with Pepper fluent?
3	Was the conversation with Pepper easy?
4	Was the conversation with Pepper frustrating?
5	Was the conversation with Pepper boring?
6	Did you feel confused most of the time talking with Pepper?
7	Did you feel confused when you answered logical questions to Pepper?
8	Did you feel confused when you answered word problems to Pepper (Including Pepper's responses may make you confused)?
9	Did you feel confused when you answered Mathematics questions to Pepper?
10	Did you want to give up this conversation with Pepper?

Table B.3 User Survey for HRI Study 2.2

No.	Questions
1	Did you enjoy interacting to Pepper overall?
2	Was the conversation with Pepper fluent?
3	Was the conversation with Pepper easy to understand?
4	Was the conversation with Pepper boring?
5	Did you feel confused most of the time talking with Pepper?
6	Did you feel confused when you talked about the logic problem to Pepper?
7	Did you feel confused when you talked about the word problem (Including Pepper's responses may make you confused)?
8	Did your confusion mitigate finally talking about the word problem (Including Pepper's responses)?
9	Were you still confused at last talking about the logic problem (Including Pepper's responses).
10	Were you frustrated this conversation with the Pepper at last?
11	Did you want to give up this conversation with the Pepper at last?

Table B.4 Your Thoughts About the Research for HRI Study 2

No.	Questions
1	I knew what the researchers were investigating in this research.
2	I wasn't sure what the researchers were trying to demonstrate in this research.
3	I had a good idea about what the hypotheses were in this research.
4	I was unclear about exactly what the researchers were aiming to prove in this research.

Appendix C

Applying Deep Reinforcement Learning in HCI and HRI

A pre-trained model in deep reinforcement learning (RL) has been used for automatic speech recognition in HCI. Rajapakshe *et al.* (2019) show that a pre-training in Deep RL can reduce the training time of speech classification and improve speech recognition performance. The deep neural network is a speech command recognition model (see Figure C.1) that is integrated into Deep RL as a deep neural network in the agent module (see Figure C.2. Convolutional Neural Networks (CNNs), LSTM RNNs combined into the speech command recognition model as CNNs can diminish frequency variations effectively as the top layer, and LSTM RNNs are strong to learn the temporal structure of a feature map. To approve how beneficial the pre-trained model is, they compared performance and standard deviation scores (see Figure C.3 and Figure C.4) between the "with pre-training" model and the "without pre-training" model. The performance scores with the pre-trained model are higher than the without pre-training, which means the training time is decreased by using pre-training. Meanwhile, the scores of standard deviation decreased dramatically with pre-training compared without pre-training, which means the pre-training accelerated the training processing (see Figure C.4). Similarly, Romeo *et al.* (2018) also involves the deep

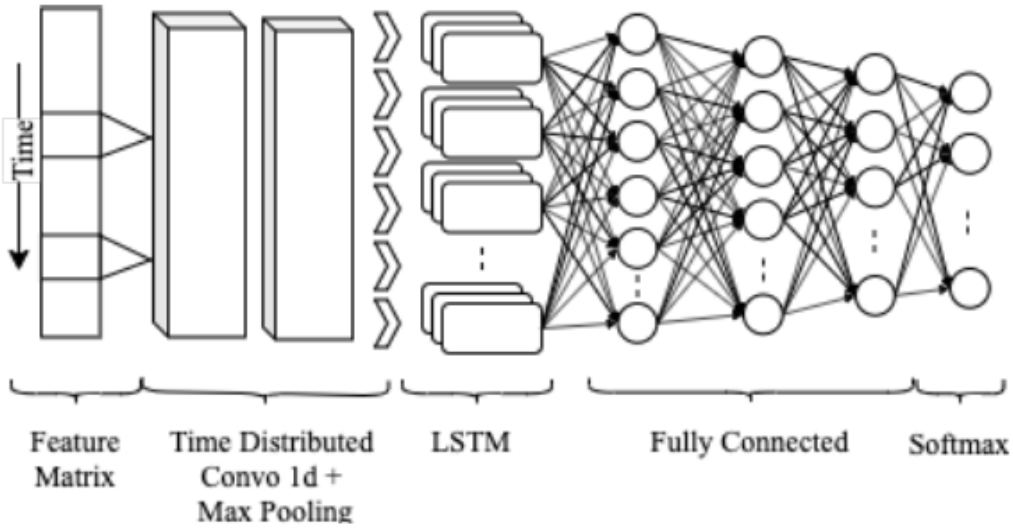


Fig. C.1 A Speech command recognition model architecture (Rajapakshe *et al.*, 2019)

Q-network framework using the Q-Learning algorithm, which is part of an algorithm in Deep RL and CNN to predict the probability of actions (waiting, calling for attention, and stating the interaction) in HRI.

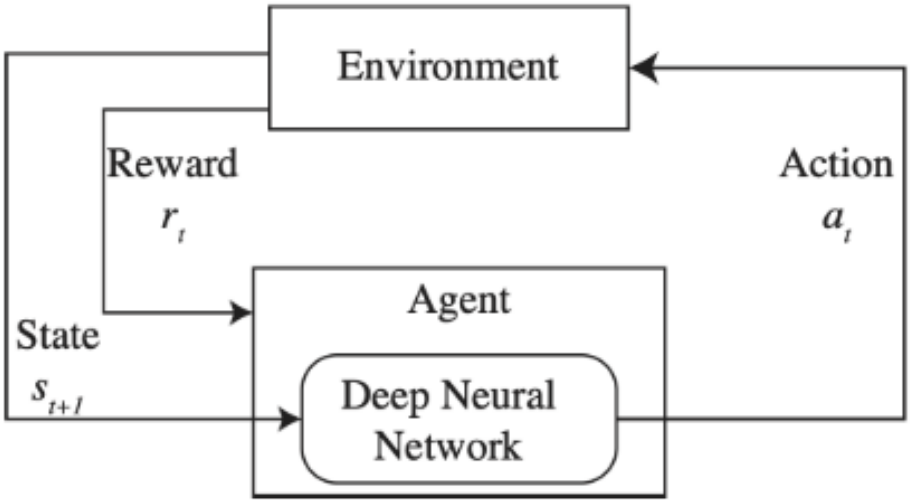


Fig. C.2 A Deep RL Framework
(Rajapakshe *et al.*, 2020)

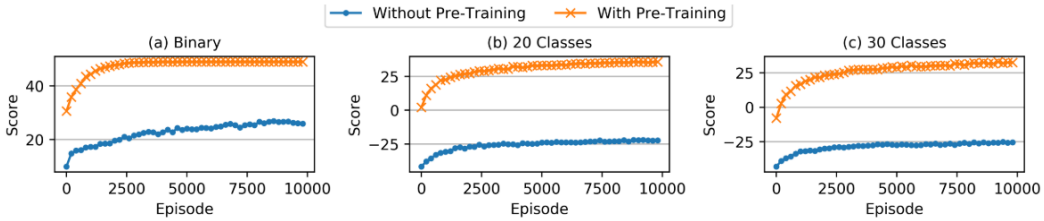


Fig. C.3 A Performance comparison in three scenarios

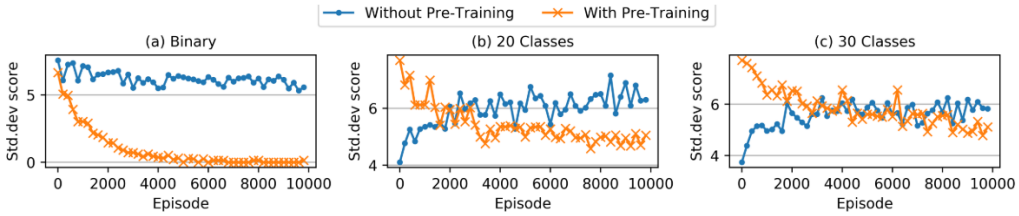


Fig. C.4 Standard deviation of the score in three scenarios
(Rajapakshe *et al.*, 2019)