Utrecht University

Master's Thesis Game And Media Technology

Human engagement state recognition for autonomous functioning of a robot in human-robot conversation

Faculty of Science
Department of Information and Computing Sciences

Author:
Kelly Griffioen
ICA-5496438

Supervisor:
Prof. dr. R.C. Veltkamp

August 19, 2020

Abstract

The goal of this thesis was to develop a model to classify the different states of engagement. We took on the definition of engagement as the process by which interactors start, maintain and end their perceived connection to each other during interaction and included the state where an interactor does not have or no longer has the intention to interact. Based on this, four states could be distinguished: *no interest*, *intention to interact* (or *interest*), *engaged* and *ending interaction*. The purpose of developing this model was to contribute to improving an informative conversation between human and robot by improving the way a robot determines who to engage with or pay attention to. Since engagement behaviour is not well understood in the human-human context, despite its apparent significance, we looked further into the research done both in human-human and human-robot interaction. Based on this, we have composed a set of features and set up a Naive Bayes classifier to classify the states of engagement. The features used are *distance from the robot*, *facing direction*, *gaze*, *position*, *sound direction* and *velocity*. The model can classify one person at a time, however the system is designed with the possibility to expand it for multiple people as well as additional features.

We intend to both choose the features and design the model in a way that it can be used regardless of the robot or platform as much as possible. However, to allow testing and to have a system that can be used in practice, we take the humanoid robot Pepper, developed by Softbank, as our main platform. This has given some limitations as to what features are chosen and to how the model is implemented.

Evaluation of the model gives promising results for the overall model and the states *no interest*, *intention to interact* and *engaged*, however the model performs badly for the state *ending interaction*. We discuss for the latter state specifically and for the model in general possibilities for improvement.

Contents

1 Introduction

In the last few years there has been a growing interest in the development of service robots that can aid humans and this is still an active area of study. Different types of applications have been developed, for instance in healthcare[20], museums[17, 27, 30] and in the management of people in need, such as the elderly[7, 8].

(a) Thorvald II by Saga Robotics.

(b) Tomato-picking robot by Root AI.

(c) Adlatus CR700 by ADLATUS Robotics.

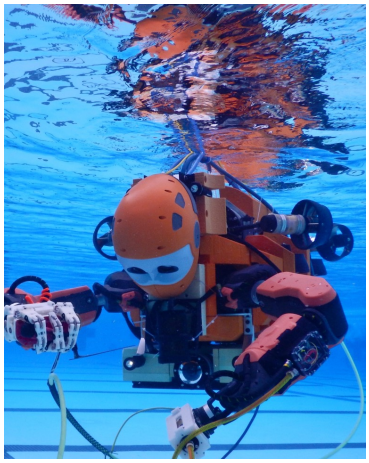
(d) Vest EXoskeleton or VEX by Hyundai Motor.

Figure 1: Examples of service robots.

The term 'service robot' is defined as a robot that performs useful tasks for humans or equipment excluding industrial automation applications[1]. Typically, they do the jobs that are dirty, dull, distant, dangerous or repetitive, including household chores. Their degree of autonomy ranges from partial autonomy, which includes human robot interaction, to full autonomy. The term therefore covers a wide range of robots such as agricultural robots (figure 1a and 1b), cleaning robots (figure 1c) and exoskeleton robots (figure 1d). Another type of service robots are the humanoid robots, which are often used for communication purposes. Many different models are available, each with a different focus, complexity and price. Ocean One (figure 2a) for example is designed to explore coral reefs and can reach

[1]Service Robots `https://www.ifr.org/service-robots/`

depths that most human beings cannot. ASIMO (figure 2b) is a humanoid robot designed to be a helper to people and Valkyrie (figure 2c) is designed to operate in degraded or damaged human-engineered environments. Pepper (figure 2d) is another example and is designed to interact with humans and is also suited for mobility. Additionally, it is more affordable than some of the alternatives, which makes Pepper a popular robot for companies. It is developed in 2014 by Aldebaran Robotics which was acquired by SoftBank Group in 2015 and rebranded as SoftBank Robotics. Currently, Pepper is used for several tasks which include being used as a receptionist, providing information at events as well as being subject to various fields of research.



(a) Ocean One by Stanford Robotics Lab.

(b) ASIMO by Honda.

(c) Valkyrie by NASA.

(d) Pepper by SoftBank Robotics.

Figure 2: Examples of humanoid robots.

A long-term goal in the field of autonomous robotics is to create systems that are capable of assisting humans in intelligent and versatile ways [3] with interfaces that preferably

require little or no training. This requires sophisticated cognitive abilities that incorporate perception, decision making and learning as well as the ability to interact with humans and inquire for information. Researchers in the field of Robot-Human Interaction (HRI) have been exploring the way to make humanoid robots interact with humans in a similar way as humans do, however there has been a lack of knowledge on human behaviour in interaction. Models exist for humans' conscious information processing, yet natural interaction involves a lot of humans' unconscious information processing [13]. A widespread assumption in the field of HRI is that interaction with a robot is good when it resembles natural (or human-human) interaction and communication [6, 18]. What humans consider natural behaviour for themselves is based on more than having a given or learnt behaviour repertoire and making rational decisions in any one situation on how to behave. It varies greatly depending on upbringing, context and the time in their lifetime. Robots do not experience the same circumstances and do not have the memory to learn from them the same way humans do. Therefore, any behaviour of a robot will be natural or artificial, solely depending on how the humans interacting with the robot perceive it.

Since HRI is such a wide field and human behaviour varies greatly depending on the situation, this thesis will aim to improve the interaction in a more specific setting. For companies and store owners, humanoid service robots focused on communication and interaction may be able to assist customers, clients or visitors by providing them with information or answering questions in stores, at branches or at events. This thesis will focus on events, which means the interaction will be in an open environment. In an open environment multiple people of different ages can interact with the robot at the same time. This however, bring a lot of complexity and therefore this thesis will focus on one person at the time. Since the ultimate goal is to include many different people, this will be taken into account with certain design choices and additions.

Research in an open environment has been done before, for example in a shopping centre by Tonkin et al. [31] and a science museum by Shiomi, Kanda, Ishiguro and Hagita [27]. As opposed to the first method, the robot in this thesis has to execute its tasks autonomously and as opposed to the latter method the robot has to be able to do so independently of its environment. It therefore cannot make use of components set up in other parts of the area such as the cameras and tags used in the science museum. To provide information and answer questions effectively, so that it can reduce the workload of a human operator, the robot needs to be able to interact autonomously with humans in a natural way according to the situation. What is considered natural behaviour depends on each persons experience, however by improving the general structure of an interaction, the naturalness of the conversation can be improved. An initial condition for having a successful natural conversation between a human and a robot is that the human is interested and wants to engage in the conversation. If the robot's task is to assist and provide information, the robot has to focus on the person who requires it. Thereafter, the conversation can continue as long as the person is engaged. If the detection and tracking of interest and engagement are improved, the robot can assist humans more efficiently. This thesis attempts to provide the basic tools to further build this desired behaviour on.

The objective of this thesis will be to contribute to improving an informative conversation between human and robot by improving the way a robot determines who to engage with. We will compose a set of features based on previous research and set up a computational model to classify the different states of engagement. Additionally we will incorporate a score for some of these states so that the robot can be directed to look at the right person. With this we will determine whether a person has the intention to interact or not, keep track of the engagement of a person with the robot and detect when a person ends the interaction when a person is within range of the robot's sensors. This can be over the course of an entire event, so it includes initiating, maintaining and ending multiple interactions and therefore the model is continuously updated. To build the model we take several features based on the behaviour humans exhibit for engagement. With these features we calculate a score on which we base the classification of the different states of engagement. Based on this model, other parts of the system can incorporate this classification to determine for instance which person to look at or how to set up directional listening to improve the robots attention. Pepper is currently often used in the event scenarios this thesis focuses on and therefore we take this robot as the primary platform.

In the next section we will discuss other research related to engagement classification in both human-human interaction and human-robot interaction, as well as the current state of Peppers engagement. We also elaborate on our research goals there.

2 Related work

An assumption often made in the field of human-robot interaction is that an interaction is natural when it corresponds with human-human interaction. We can therefore take inspiration from human-human interaction to have a robot better detect which person to focus on. In the paper by Sidner, Lee, Kidd, Lesh and Rich [28] they state that when individuals interact with each other face-to-face, they use gestures and conversation to begin their interaction, to maintain and accomplish things during the interaction and to end the interaction. The paper defines engagement as the process by which interactors start, maintain and end their perceived connection to each other during interaction. It combines verbal communication and non-verbal behaviours, all of which support the perception of connectedness between interactors. Evidence for the significance of engagement becomes apparent in situations where engagement behaviours conflict or are not present at all. One such example is given by Salem and Earle [24]. In their paper they emphasise the expressiveness of the avatar as a crucial improvement to the efficiency of their communication capabilities. They state that investigations into the structure of virtual world social encounters reveal that the process of interaction typically breaks down into three sequential stages [33, 10]. In the start part people seek an individual or a group for meaningful conversation. In the middle users interact and in the end they negotiate a way breaking out of the interaction. To improve the expressiveness of virtual characters, Salem and Earle describe a vocabulary of expressions to be implemented consisting of bodily actions that are described by acronyms, emotion icons and keywords that are found within the text messages used to

communicate in the virtual world. Despite the apparent significance of engagement behaviour, it is not well understood in the human-human context. This is partly because it has not been identified as a basic behaviour. Instead, behaviours such as looking and gaze, turn taking and other conversational matters have been studied separately, but only in the sociological and psychological communities as part of general communication studies. In artificial intelligence the focus lies more on language understanding and production, rather than on gestures or on the fundamental problems of how to start and maintain a connection. Only since conversational agents and better vision technologies started to be developed, studies started to address this. In human-robot interaction studies engagement is now commonly defined [31, 1, 23].

For this thesis, the definition of engagement by Sidner et al. will be adopted. Based on this, four states can be distinguished: *no interest*, *intention to interact* (also called *interest*), *engaged* and *ending interaction*.

2.1 Human-Human Interaction

One study, though old, that does look into a stage of engagement solely in human-human context is done by Knapp, Hart, Friedrich and Shulman [16]. They have looked into what specific verbal and non-verbal behaviours are associated with the termination of a conversation. They've conducted an experiment in which a person was asked to interview someone to obtain information as quickly as possible. For each interview, four coders analysed verbal and non-verbal cues according to specific categories starting from forty five seconds prior to the interviewer rising from his seat until he left the room. The categories used were based on a review of literature, surveys, controlled observation and a pilot project all conducted by the authors of this paper. Table 1 shows an overall rank ordering of the frequency of the non-verbal categories used in terminating the conversation.

Rank	Non-verbal variables	Mean
1	Breaking eye contact	1.89
2	Left positioning	1.76
3	Forward lean	1.66
3	Nodding behaviour	1.55
4	Major leg movement	1.38
6	Smiling behaviour	1.31
7	Sweeping hand movement	1.23
8	Explosive foot movement	1.19
9	Leveraging	1.17
10	Major trunk movements	1.10
11	Handshake	1.09
12	Explosive hand contact	1.02

Table 1: Rank ordering of non-verbal leave-taking behaviours.

2.2 Human-Robot Interaction

Engagement is more commonly defined in Human-Robot Interaction and several studies have tried to shed some light on engagement by focusing on one part in different ways in various specific scenarios.

2.2.1 Initiating an interaction

When taking the definition of engagement made by Sidner et al., the first state is that of starting an interaction. This can be initiated from both sides, which gives researchers the choice for the robot to either wait until it is approached or approach someone itself. In both scenarios it is important for the robot to recognise whether a person wants to interact or not.

Proactive behaviour The following papers all take on a more proactive method by letting the robot approach a target to initiate an interaction.

In the paper by Satake et al.[25] for instance, they propose a model of approach behaviour with which a robot can initiate conversation with people who are walking. They predict the walking behaviour of people in a shopping mall, which they use to target a person the robot can approach. In their previous study they have collected and clustered people's trajectories and classified each trajectory with an Support Vector Machine (SVM) into four behaviour classes: fast-walking, idle-walking, wandering and stopping. Based on these classes, the robot could choose the person who is likely to be interested in conversation with the robot. The robot can then plan its approaching path and non-verbally indicate its intention to initiate a conversation.

Another approach is presented by Pourmehr, Thomas, Bruce, Wawerla and Vaughan [22]. They use a simple probabilistic framework for multi-modal sensor fusion that allows a mobile robot to reliably locate and approach the most promising interaction partner among a group of people in an uncontrolled environment. They focus on controlling a mobile robot's attention in multi-human robot interaction for distances greater than two meters. Which person is most interested in interacting with the robot is based on the idea that a person who is standing facing the robot and calling it will have the highest probability of being a potential interaction partner. To determine where this person is located they use three features. They detect legs to detect people, torsos to detect the orientation of the people and the direction of sound to determine if a person is speaking. For each modality they track detected humans using a bank of Kalman Filters. The output of these filters are converted into probabilistic evidence grids which are all centred over the robot so the results overlap. These grids are then fused to compute the integrated probability distribution.

In the paper by Kato, Kanda and Ishiguro [14] they model polite approaching behaviour based on the behaviour of staff members in a shopping mall. These staff members adjust their behaviour based on their estimation of a visitors' intention. To do this for their robot system, the authors collected pedestrians' trajectories around a robot. Additionally, they incorporated the distance from the robot, the fan shaped area in front of the robot that covers the direction of the trajectory when a person moves towards the robot, the stability of walking velocity and how long a person stands still. With these features they distinguished 'intention to interact' and 'other distinctive intention'. Every person not belonging to either class was classified as 'uncertain'. For classification they used an SVM.

Passive behaviour As previously stated, another method is to have a static robot which can be approached by people. It then needs to be taken into account that in a dynamic public spaces such as the ones seen in these papers, a robot needs to be able to determine the needs and intention of multiple people in a scene, so that it only interacts with people who intent to interact with it.

In the paper by Foster, Gaschler and Giuliani [11] they address the task of estimating the engagement state of customers for a robot bartender based on the data from audio-visual sensors and incorporate behaviour to deal with multiple customers at once. To confirm that the sensor data contains the information necessary to estimate user states, they have done an offline experiment using hidden Markov models. They then compared two strategies for online state estimation, namely a rule-based classifier based on observed human behaviour in real bars and a set of supervised classifiers trained on a labelled corpus. They found that all classifiers change their estimate too frequently for practical use, so to address this they presented a classifier based on Conditional Random Fields.

Klotz et al. [15] present the integration of an engagement model in an existing dialog system based on interaction patterns. As a sample scenario, this enables the humanoid robot Nao to play a quiz game with multiple participants. To determine the user's actions (e.g. if the user explicitly wants to start an interaction with the system), they use a set of possible utterances which are matched against the results of a speech recognition module. To get an estimation of the user's *intention to interact* they estimate the user's current visual focus of attention.

Within the context of engagement, non-verbal signals are used to communicate the intention of starting the interaction with a partner. Vaufreydaz, Johal and Combe [32] investigates methods to detect these signals in order to allow a robot to know when it's about to be addressed. Classically, spatial information like the human-robot distance and human's position and speed are used to detect engagement. In this paper however, they also integrate multi-modal features. They started with computing 99 features on their corpus, which they reduced based on social and cognitive science research on non-verbal communication cues, the available sensors and performance of algorithms used in experiments. Eventually they demonstrated however, that 7 selected features are sufficient to provide a good starting engagement detection score.

In another example they propose a detector for the intention-for-interaction which fuses multi-modal cues using a probabilistic discrete state Hidden Markov Model. The cues they use are line of sight, anterior body direction and vocal activity[19].

Yumak, van den Brink and Egges [35] do not make use of a robot, but present a gaze behaviour model for an interactive virtual character simulated in the real world. The sensors used however are used in other papers as well and the model is suitable to be implemented with a robot. For this they focus on estimating which user has an *intention to interact*. The model takes into account behavioural cues such as proximity, velocity, posture and sound. They use this to estimate an engagement score, which drives the gaze behaviour of the virtual character.

2.2.2 Maintaining and ending interaction

As defined by Sidner et al., initiating interaction is only one part of engagement. The other two parts, maintaining and ending an interaction, do not seem to have had as much attention in human-robot interaction as the first.

In the paper by Sidner et al. they have studied the effect of tracking faces during an interaction and applied the results to human-robot interaction. They studied how two humans tracked each other's faces in one interaction. From this they defined the principle of conversational tracking: a participant in a collaborative conversation tracks the other participant's face during the conversation in balance with the requirement to look away in order to: (1) participate in actions relevant to the collaboration, or (2) multi-task with activities unrelated to the current collaboration, such as scanning the surrounding environment for interest or danger, avoiding collisions, or performing personal activities. They then applied this principle on the behaviour of a humanoid robot that can participate in conversational, collaborative interactions with engagement gestures. Experiments showed that people found these interactions more appropriate than when the engagement gestures are absent.

While people are maintaining their connection to each other via conversation, people need to coordinate with one another on turn taking. This is done through both verbal and non-verbal cues such as establishing or breaking eye contact and the use of head and hand gestures. Some of these cues are similar to cues used to recognise other states of engagement. For example, in a conversation with multiple people, participants direct their gaze towards the person speaking. Directing their gaze away from the robot may be interpreted as losing interest, while this is not the case in this scenario. Research has been done on turn taking in human-robot interaction and what features are important. Bohus and Horvitz [2] for example have constructed a computational framework for managing multiparty turn taking in situated spoken dialog systems. They use face detection and head pose tracking software to detect and track multiple participants in the scene, as well as their focus of attention. They also capture the audio, perform speech recognition and perform sound localisation. From this they infer attention, engagement and turn taking among other conversational aspects.

2.3 Pepper

Despite its popularity, Pepper has its limitations when it comes to autonomously providing information, including in recognising who wants to interact with it. When only using the software Pepper has by default, Pepper responds to any detected stimulus (sound, movement or touch) by looking in its direction. Then it will check if the stimulus corresponds to a human. Pepper can detect a presence if it is within 1.5 meters. If the presence is from a human, Pepper is then engaged with this person. When engaged, there are three options that specify how focused the robot is on the engaged person. The default option for Pepper is that when it's engaged with a person, it can be distracted by any stimulus and engage with another person. Alternatively, as soon as Pepper is engaged with a user,

it can be set not to get distracted by anything. In this case, Pepper focuses on the first person until it loses him and can therefore only switch to another person when the first one has walked away. The third option is to have Pepper still listen and react to other stimuli, but always have it return to the person it is engaged with. Pepper tries, as much as possible, to establish and keep eye contact while engaged. This and more information about Pepper can be found on the Aldebaran documentation website[2].

2.4 Research goals

To improve Pepper's own engagement, it has to be able to recognise the engagement state or the absence of an intention to interact from the people around it, starting with a single person. Previous research has studied parts of engagement using both spatial and multimodal features. Based on this, the main goal of this thesis therefore is:

- Developing a model to detect engagement for a single person using spatial and multimodal features

To develop this model, several questions need to be answered:

- Which features sufficiently show when a person does not have the intention to interact?
- Which features sufficiently show when a person does have the intention to interact?
- Which features sufficiently show that a person wants to maintain their perceived connection?
- Which features sufficiently indicate a person is ending the interaction?
- Which approach can sufficiently classify the states of engagement using said features?

Note that due to the choice of robot these questions are all specific to the possibilities of the available sensors. Furthermore, in the future the goal is to use the output of this model to improve the behaviour of Pepper. It is therefore necessary to be able to use this model in real-time.

3 Approach

This section describes the decisions made regarding the use of hardware, software and approach for building an engagement model. The final approach consists of three parts. Pepper provides sensory data which is send via a client over the network. On the server side, this data is processed by a network to create features. These features are then used by the model to classify the different stages of engagement.

[2]Aldebaran documentation `http://doc.aldebaran.com/2-5/family/pepper_user_guide/interacting_pep.html`.

3.1 Specifications Pepper

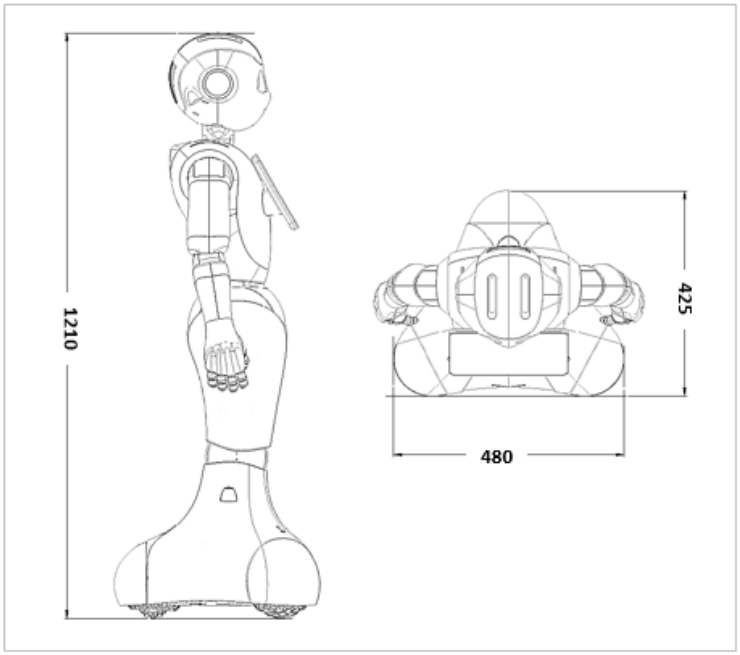


Figure 3: Dimensions of Pepper with its arms down.

Pepper is a humanoid robot developed by Aldebaran Robotics. It was first introduced in 2014 and is designed to interact with humans. Users can interact with Pepper through either speech and dialogue which is available in 15 languages or through its tablet. This section gives an overview of its hardware and relevant modules. Information about all the modules is available at the Aldebraran documentation website[3].

3.1.1 Hardware

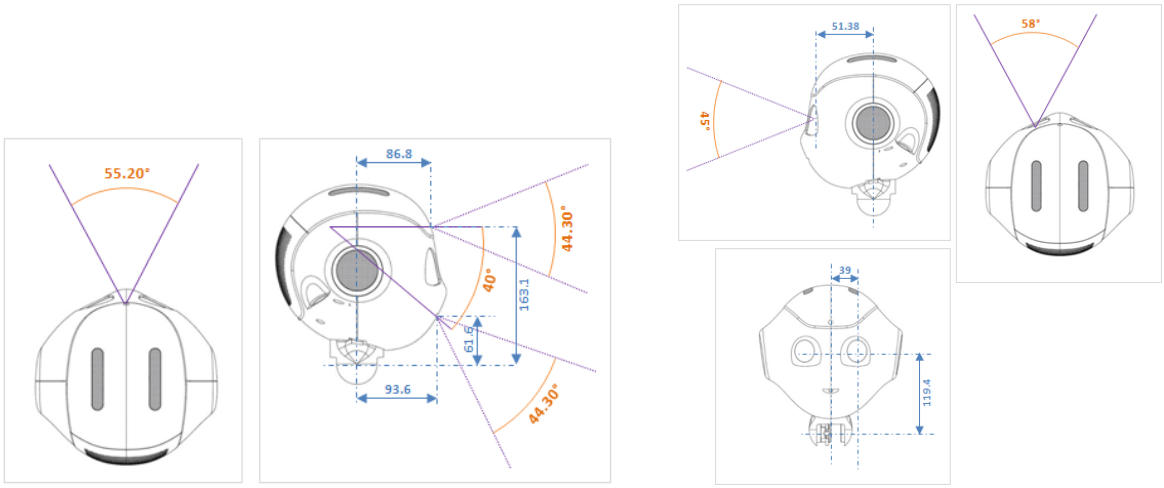
For this thesis, we have made use of Pepper version 1.6.

Dimensions Pepper is 120 cm tall and weighs about 28 kg. It has 20 degrees of freedom for expressive movements and can rotate 360 degrees. The dimensions of Pepper are displayed in figure 3. The joints of Pepper are displayed in figure 5.

Cameras The robot head is composed of three cameras:

- Two 2D Cameras
- One 3D Sensor

[3]NAOqi modules `http://doc.aldebaran.com/2-5/naoqi/index.html`.



(a) Location and field of view of the 2D cameras.

(b) Location and field of view of the 3D sensor.

Figure 4: 2D cameras and 3D sensor of Pepper.

All three of them are located in the forehead. Figure 4a shows the location and field of view of the 2D cameras. They provide a resolution up to 640*480 at 30 frames per second (fps) or 2560*1920 at 1 fps. The 3D sensor provides an image resolution up to 320x240 at 20 frames per second. Its location is shown in figure 4b.

Microphones Pepper has four microphones located on the top of its head.

Lasers Pepper is equipped with six laser line generators. Three of those produce lasers projected on the ground in front of Pepper and three of those are projected on the surroundings in front, at the left and at the right side of pepper.

Sonars The robot is equipped with two ultrasonic sensors (or sonars) which allow it to estimate the distance to obstacles in its environment.

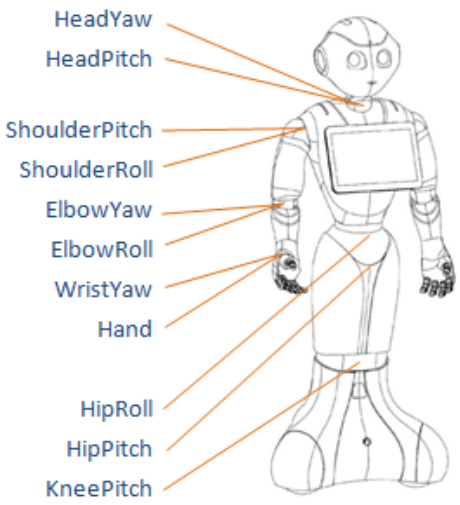


Figure 5: Joints of Pepper.

3.1.2 Software

Pepper runs on NAOqi, which is an embedded GNU/Linux distribution based on Gentoo. The programming framework is called the qi Framework (previously NAOqi Framework). NAOqi supports working with Python, C++,

Java, Javascript and ROS[4]. Additionally, a Python and C++ SDK can be used to work with the qi Framework.

Another option is Choregraphe. This allows a user to create applications containing Dialogs, services and powerful behaviors, such as interaction with people, dance, e-mails sending, without writing a single line of code[5]. All documentation on the robots of Softbank is available online[6].

Pepper defines three types of frames for itself and the objects around it. These three frames are illustrated in figure 6.

For this thesis, the qi Framework (version 2.5) is used with Python to gather the sensory data and send it through a WebSocket connection. The qi framework comes with a list of core modules which give access to the sensors and provide interpretation to a certain extent. This section further explains the modules and data that are used.

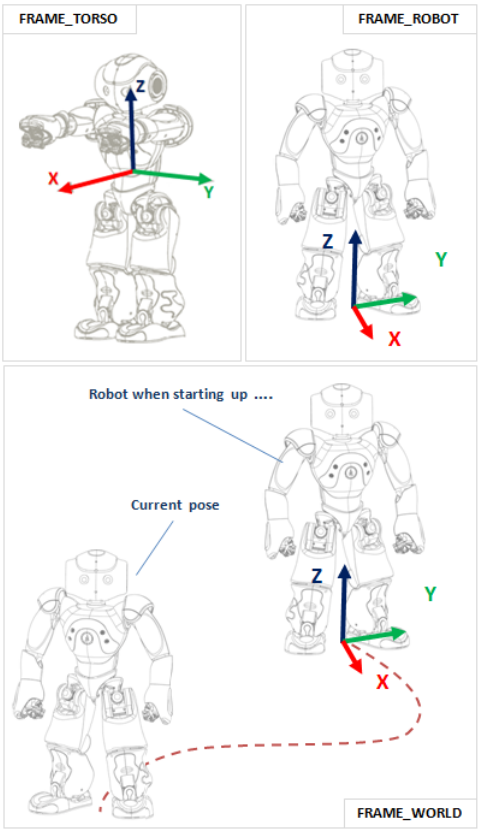


Figure 6: The three frames used by both the Nao and Pepper robot.

ALPeoplePerception This module is an extractor that keeps track of the people around the robot and provides basic information about them. It gathers visual information from the RGB cameras and the 3D sensor if available. Once people have been detected their attributes are constantly updated. The event that is used from this module is the 'PeopleDetected' event, which is raised whenever at least one person is visible. From this event the position of each person in the robot frame is collected.

ALSoundLocalization This module identifies the direction of any loud enough sound heard by Pepper. The sound wave emitted by a source is received at slightly different times on each of the Pepper's four microphones, from the closest to the farthest. By using this relationship, the robot is able to retrieve the direction of the emitting source.

ALVideoDevice This module provides images from the video cameras of Pepper.

[4]Programming of NAOqi `http://doc.aldebaran.com/2-1/dev/programming_index.html`.

[5]Choregraphe Suite `http://doc.aldebaran.com/2-5/software/choregraphe/index.html#choregraphe-suite`.

[6]Softbank robot documentation `http://doc.aldebaran.com/2-5/index_dev_guide.html`.

3.2 Platform for Situated Intelligence

A framework that is often used in research (e.g. in [22, 21, 23]) is the open-source Robot Operating System (ROS) framework. Its primary goal is supporting code reuse and it is both language and sensor independent. Additionally, ROS has access to more information from Pepper's sensors. The downside of ROS however, is that the infrastucture has to be build up from the ground up. This means that Pepper loses the processes that are already set up, such as the idle animation and stance. Additionally, newer versions of Pepper most likely will no longer support ROS. Since the extra information we can get through ROS does not outweigh the extra work we have to put in to get the functionality we need, we have decided to go with the PSI framework instead.

The Platform for Situated Intelligence (PSI) is an open, extensible framework developed by Microsoft. It enables the development, fielding and study of situated, integrative-AI systems[7]. It provides a *Runtime* that enables parallel coordinated computation, a set of *Tools* that provide development support and an ecosystem of *Components*. The latter provides a wide array of AI technologies encapsulated into PSI components, which can be easily developed by wiring them together. The initial set of components includes sensor components for cameras and microphones, audio and image processing components. These are used to create a network of relevant components for Pepper.

The data provided by the modules from the qi Framework is send to the PSI network through a WebSocket connection. Similar as to how the modules use events to send data, the server fires an event when data is received. Specific event source components are subscribed to the right event through which the data is send into the network. This network processes the data and produces what is needed to recognise the states of engagement by the engagement model.

3.3 Engagement model

Recognising the states of engagement is done by recognising the behaviour of humans. Researchers of the covered related work have done this by selecting or finding certain meaningful features of this behaviour. To classify the four states *no interest*, *intention to interact*, *engaged* and *ending interaction* for each person within the proximity robot, a model needs to be developed that can be continuously updated for multiple people who may enter and leave the vicinity. For this we need a set of features and a method to combine these in a way that the classes can be distinguished.

3.3.1 Features

To compose the set of features we take into account the following three aspects: the explored related work, the human frame and the limits brought by the choice of robot.

As for the related work explored in the previous section, engagement has been explored in several different situations with a different focus and therefore with a different set of

[7]PSI `https://github.com/microsoft/psi/wiki`.

	Facing Direction	Sound Direction	Gaze	3D Head Pose	Distance From Robot	Speech Recognition	Facial Expression	Body Language	Position	Velocity	Torso	Speech Activity Detection Tags
Approach Humans In Shopping Centre	x				x				x	x		
Find Person Calling		x									x	
Customers Of Bartender	x	x	x	x	x		x	x			x	
Human Assist Robot With Task												
Non-Verbal Language For Virtual Avatars	x		x	x			x	x				
Robot In Home Environment	x	x			x							x
Intention To Interact When User Sits 3m Before Robot	x			x								x
Quiz Game			x			x						
Turn Taking		x	x	x		x						
Virtual Character To Determine Who To Look At	x	x		x	x				x	x		
Ending Conversation Between Humans In Sitting Position	x		x			x	x	x				

Table 2: Features that are used in more than one situation found in the related work. Similar situations in different papers are combined.

features. By combining the similar situations and tallying which features are used for each, an overview is obtained that shows which features are used in the most situations and which are used for situations similar to the set-up in this thesis. Table 2 shows the features that were used in more than one situation. Some papers however, describe the used features in more detail than others, which makes it difficult to determine whether they use the same features or if they're different in any way. Additionally, most of the found papers focus on *intention to interact* (or *interest*) and fewer papers on the other classes. This means table 2 tells us more about the first then the latter and we might miss important features. Another observation we made is that several papers explored which features where more important than others ([25, 11, 32, 16, 35]) and Vaufreydaz et al. suggest that selecting features can be more relevant than combining them. The latter statement is substantiated by the model used by Klotz et al., which primarily uses the gaze of the participants as the only feature. We therefore will also take into account the validation given in the papers for the features individually. Furthermore, the choice of features is dependent on the robot and hardware used. This means that some papers may have left out features solely because

	Facing Direction	Sound Direction	Gaze	3D Head Pose	Distance From Robot	Speech Recognition	Facial Expression	Body Language	Position	Velocity	Torso	Speech Activity Detection Tags
No interest	x	x	x	x	x				x	x		
Interest	x	x	x	x	x		x	x	x	x		x
Engaged	x	x	x	x	x	x				x		
Ending interaction	x		x			x	x	x				

Table 3: Features that are used in more than one situation found in the related work, organised by engagement state.

they could not detect them (accurately).

Because of this last issue, we also take into account the human frame in general. Since in the field of HRI the assumption is made that human-robot interaction is natural when it compares to human-human interaction, we can use our personal experience and that of others, like they have done in other papers [31, 22, 11]. This way we can infer which parts of the body are prominent in showing engagement in conversation and how they are used. We can then look at the related work again from this perspective and consider how those researchers have considered or incorporated these parts. Furthermore, we can compare how those researchers have chosen their features based on the human frame and their personal experience with our ideas and perhaps derive other important features of this.

Since the focus of this thesis lies on classifying the four states we have divided engagement in, it is important that we have sufficient features for each state. We therefore have made an overview categorised by state of the features of table 2. This overview is shown in table 3. Note that not all papers specify which features are important for which state and some features are used for a different purpose then we want to use them for. Since we cannot use those for our model, we have left them our of table 3.

Our ultimate goal is to develop a general model with a general set of features that can be used regardless of the robot or platform. However, since we assess this initial direction with the Pepper robot, we have to take the sensors and hardware of Pepper into account, as well as the way we set up our PSI network and the capacity of the wireless network used to pass the sensor information to this PSI network. Taking all of the considerations explained in this section into account, we have narrowed the feature set down to the following six features:

- Distance from the robot
- Facing direction

- Gaze
- Position
- Sound direction
- Velocity

3.3.2 Method for combining features

The research covered in section 2 also describes several methods to classify the respective states of engagement each paper focuses on. For this thesis we aim to develop a model which is based on the understanding of the non-verbal behaviour involved in the different stages of engagement. A rule-based classifier seems to be a suitable choice, however since we have chosen six different features which are also to be categorised, a rule-based classifier will become too complex. Additionally, we aim to develop a model that can be extended with additional features relatively easily. For this purpose, we have decided to set up a Naive Bayes classifier, which classifies the states one frame at the time. This approach can naturally be extended to a situation requiring multinomial classification and it has been shown that the method performs well despite of the underlying assumption of conditional independence. This method can also provide a score for the *intention to interact* and *engaged* states in the form of the chance of the assigned state.

To distinguish between each state, we categorise each feature further to facilitate the differences between each state. We will elaborate on our choices based on the research done in both human-human and human-robot interaction that we have covered before.

Distance Research in social sciences investigated how people manage distance during social interactions [12]. It is considered along four zones: the intimate zone (0 to 0.15m), the close intimate zone (0.15 to 0.45m), the personal zone (0.45m to 1.2m), the social zone (1.2 to 3.6m) and public zone (more than 3.6m). According to Shi, Shimada, Kanda, Ishiguro and Nagita [26], when a person approaches, the first utterance would happen at a 2m distance in a small quiet room, however Kato, et al. found that in a larger and more noise environment it is more natural to use the first utterance at a farther distance and empirically decided it to be 3m. They have however set the social distance at which the conversation takes place to 1.5m.

Facing direction One of the papers that uses et al. have set the limit of the facing direction angle to 45°.

Gaze The gaze angle is mainly used in its raw form [11, 2], though Yumak et al. also have set the limit to 45°.

Position Yumak et al. define the position as 'closeness to the centre of field of view (FoV)' and calculate the 'field of view deviation'. When a person stands directly in front of the robot, this deviation is 0°. The limit is set to 35°to either side based on the limits of their setup. When a person stands directly in front of the robot, this person is considered to be more engaged than someone standing at 35°. They therefore map the deviation value to 1-0, where 1 is for a deviation of 0°and 1 for a deviation of 35°.

A similar approach is taken by Kato et al., where they define a fan shaped area to estimate whether the trajectory of a person would lead to the robot. If a smaller fan can fit the direction of the trajectory, it is more likely that the person is moving towards the robot.

Sound direction The main purpose of the sound direction in the covered research is to determine whether a person is speaking or calling to the robot.

Velocity Several papers have mentioned that they distinguish between standing still and moving, which is indicated by the threshold of 1m/s [14].

Feature categories Taking into account all of this information, we categorise our features as follows, taking into account our own interpretation and expectations based on the scenario we want to apply this model to, as well as the possibilities of our chosen robot and approach.

- Distance
  - Distance $>$ 3.6m
  - Distance $<$ 3.6m, $>$ 1.5m
  - Distance $<$ 1.5m
- Facing direction
  - Facing direction $>$ 45°
  - Facing direction $<$ 45°, $>$ 10°
  - Facing direction $<$ 10°
- Gaze
  - Gaze $>$ 90°
  - Gaze $<$ 90°, $>$ 45°
  - Gaze $<$ 45°, $>$ 10°
  - Gaze $<$ 10°
- Position (angle with respect to the FoV)

  - $0° \rightarrow 1$
  - $90° \rightarrow 0$
- Sound direction (or speaking)
  - Speaking
  - Not speaking
- Velocity
  - Velocity > 0.1m/s
  - Velocity < 0.1m/s

The way in which we define the position feature, it cannot be used like the other features for classification by the Naive Bayes classifier. The reason we have chosen this feature and defined it like this, is because we expect it to be of significant value for determining who to look at in a scenario with multiple people. We assume that interested people standing directly in front of Pepper have a higher chance of wanting to engage and that the engaged person of highest importance also stands closer to the centre of Peppers FoV. Additionally, this could give an indication regarding the number of people who are in conversation with Pepper, since we presume that two (or more) people in conversation share the space in front of the person (or robot) that has their focus. Therefore, instead of incorporating it in the classification, we use it to adjust the final score of the *engaged* and *interest* states.

Bayes' theorem Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be relative to the event. This theorem is mathematically stated as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

where

- A and B are events
- $P(B) \neq 0$
- $P(A|B)$ and $P(B|A)$ are conditional probabilities, i.e. the likelihood of event A occurring given that B is true and the likelihood of event B occurring given that A is true
- $P(A)$ and $P(B)$ are marginal probabilities, meaning the probabilities of observing A and B respectively

Naive Bayes classifier With Bayes' theorem we can calculate the probability of a class based on a feature. For multiple features, Bayes' theorem can be extended to Naive Bayes, provided that the features are independent. Naive Bayes can mathematically be written as follows:

$$P(C_k|x_1,\ldots,x_n) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{\prod_{j=1}^{n} P(x_j)} \tag{2}$$

where

- C_k is a class for which the probability is calculated
- the features are denoted by x

This formula can be used for classification of a set of features with multiple classes by calculating the probability of each class and assigning the class with the highest probability. Since the denominator is only dependent on the set of features, this can be omitted. The corresponding Naive Bayes Classifier is the function that assigns a class label $\hat{y} = C_k$ for some k as follows:

$$\hat{y} = \underset{k\in\{1,\ldots,K\}}{argmax} P(C_k)\prod_{i=1}^{n} P(x_i|C_k) \tag{3}$$

Probabilities The naive Bayes classifier requires several known probabilities. These are usually calculated based on the occurrences of the features and classes in data that has been obtained beforehand. Since we do not have this data yet, for the first test scenario we will set up we make use of estimated probabilities. From this scenario we will record the data, which will then be used to calculate the probabilities offline.

Online process When a person stands in the vicinity of the robot, we get a set of features. With these features the state is assigned with the naive Bayes classifier. If the state is either *intention to interact* (or *interest*) or *engaged*, the probability of this state will then be multiplied with the position with respect to the FoV, which is 1 for 0° and will decrease to 0 when it becomes 90°. The behaviour of the robot can then be determined on both the class label and the final probability.

3.4 Evaluation

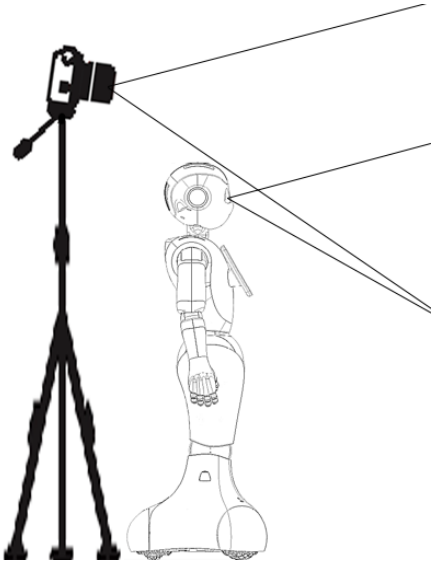
As mentioned before, we cannot collect a significant corpus to train and assess our model. We can however record the sensor information trough PSI in a smaller, controlled scenario.

The goal of the engagement model is to classify the different states of engagement based on the behaviour humans use to indicate these states. Therefore, the model is successful when it assigns the same state to a participant another human would assign. To evaluate the engagement model, we will compare the classification made by the model with the classification made by ourselves and one other person. To avoid disparity between the human classifications, we will discuss our decisions to come to an agreement on the state

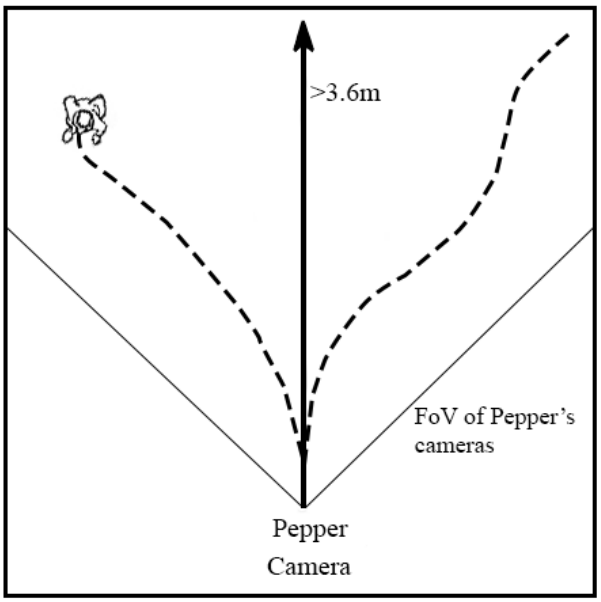
for every frame. This gives the opportunity to discuss which features the classification is mostly based on for future adjustment of the model.

The model's classification is done in real-time on a separate laptop with the sensors of the Pepper robot and is based on a frame rate of one frame per second. As explained, the four different states of engagement are defined as *no interest*, *interest*, *engaged* and *ending interaction*.

In a pilot study we found that the definition of *ending interaction* as defined by the participant differs from the concept we used, specifically it differs in where it starts, ends and what the telltale features are to recognise this. Since the model classifies each frame separately and therefore does not take into account what has happened before each frame, we have decided to define *ending interaction* as the action of turning away after the conversation ended. To assess its performance for all states of engagement, we will set up five scenarios. One shows all states in the order of *no interest*, *interest*, *engaged*, *ending interaction*, *no interest*. The other four scenarios only contain one state in an attempt to create an equal amount of frames for each state, since in a natural interaction, not all states are of equal duration. With our definition of the state ending interaction, it is impossible however to get an equal amount of frames to the other states while maintaining the naturalness of the behaviour. This is because *ending interaction* is a relatively short state compared to the other states and can't be extended by itself. We therefore execute this scenario a total of nine times. In an attempt to keep it more natural, the scenarios set up for this state are based on the participant's idea of the state rather than just the action of turning away. The scenarios therefore include a final closing sentence, turning



(a) An illustration of the setup of the extra camera used to record the experiments.



(b) An illustration of the setup of the experiments.

Figure 7: Illustrations of the setup of the experiments.

around and walking away, which means the scenarios consist of the states *engaged*, *ending interaction* and *no interest*. These scenarios are played out twice by a single person, once were all interactions take place in a straight line in the centre of Pepper's field of view and once with more variation in the position of the participant. All scenarios will entirely take place within Pepper's FoV. Since Pepper's 'autonomous life' software makes it move it's head often, this will be paused so that Pepper only looks straight ahead. Note that the ALPeoplePerception module only fires an event based on changes. It therefore is necessary for the participant to start outside of Pepper's FoV and walk up to the first position of the scenario before the recording can start. The states shown in the scenario are the same in the two times the scenario is enacted, however the details of the scenario will differ to take advantage of the participant's freedom in movement. This is all done by two different people, giving a total of 20 scenarios.

Since the purpose of this system is to classify in real-time, the entire network including the engagement model is run. Since the Naive Bayes classifier uses probabilities based on previously acquired data we don't have, we use probabilities that are manually estimated based on our own expectations. Alongside the respective frames, all final features (i.e. the raw data categorised as explained in section 3.3.2) are stored. This way we can afterwards evaluate our model using an offline version of the engagement model, while we will still see the performance and possible flaws of both the network and engagement model in the scenarios themselves. We will use k-fold cross validation to evaluate our model.

Additional to the system and the robot, a camera is set up to give a third-person perspective from behind Pepper. This view is used for the classification by the humans. The setup of this camera is illustrated in figure 7a. The room in which the experiment takes places is large enough to walk from one side to another at a distance larger than 3.6m from Pepper. It is devoid of anything that resembles a human or can reflect the participant, so that Pepper only classifies one person per frame. The setup of the room is illustrated in figure 7b.

To keep the scenarios as natural as possible we will explain beforehand that the goal of the experiment is to record the different stages of a conversation and will give instructions as tasks rather than as states. Additionally, the participant receives a script of a short conversation they can have with Pepper. This conversation contains questions to which we ourselves will respond through Pepper. The participant is informed at the start of the experiment how many scenarios need to be recorded and that instructions for each scenario are given before each scenario, rather than all at once. Afterwards, the participant is informed about the specific states we have divided interaction in and is asked to classify the videos of the other participant. We will classify the videos of both, so that each video is classified by two people.

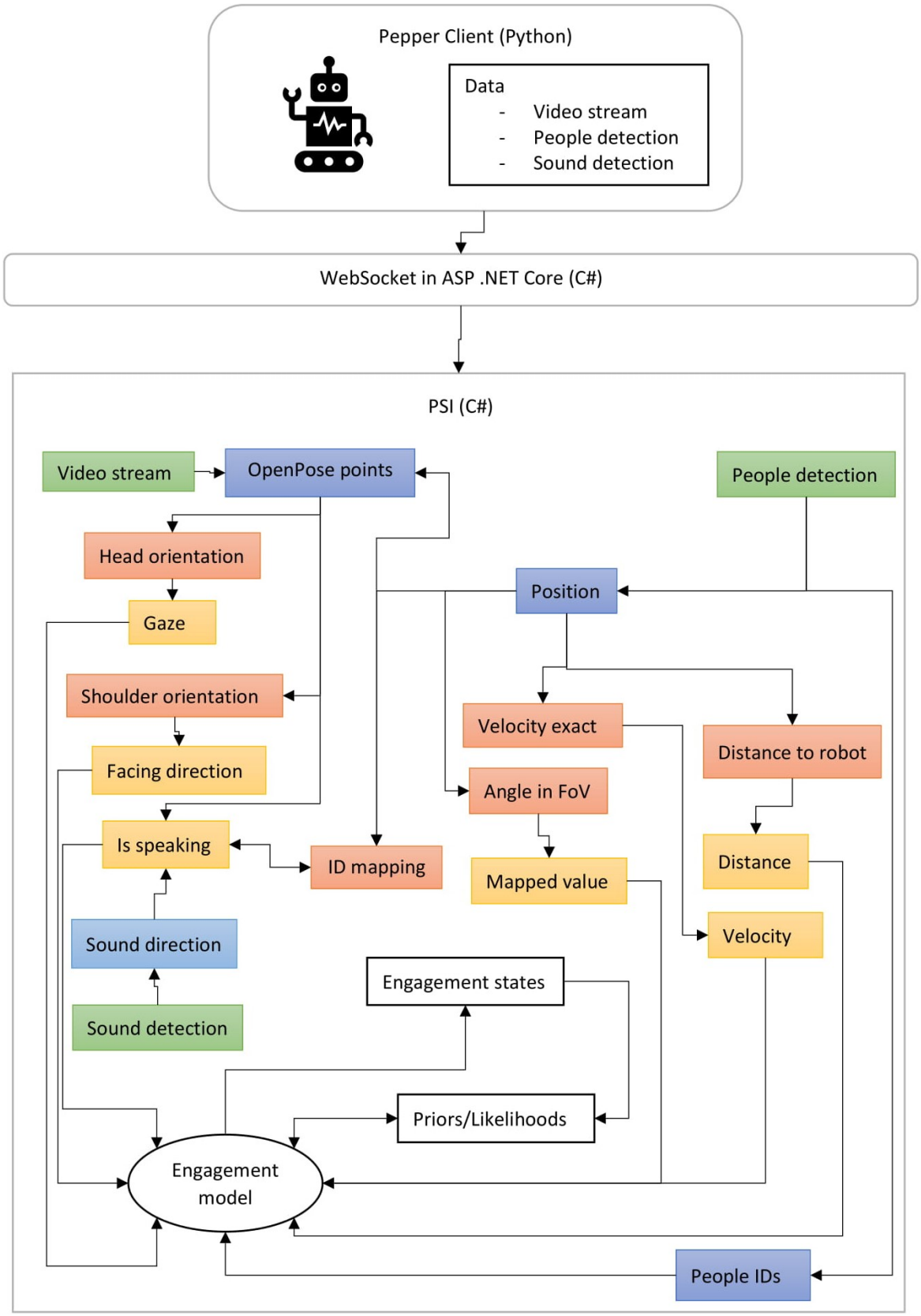


Figure 8: The system for running the engagement model with Pepper.

4 Implementation

To send as little data as possible through the Wi-Fi network, we collect the following data on Pepper from the modules mentioned in section 3.1.2:

ALPeoplePerception
- Timestamp
- IDs of people in view
- Position of each person in the robot frame

ALVideoDevice
- Frame

ALSoundLocalization
- Timestamp
- Azimuth values of all heard sounds
- Elevation values of all heard sounds

This gets sent through the PSI network. The system is set up according to figure 8 and consists of three parts: the client, the server and the PSI network. The PSI network in turn consists of four groups. The green blocks represent the modules and the data we have send from those, the red blocks represent the exact values of each feature and the orange blocks represent the features categorised as discussed in section 3.3.2. All of the features then converge in the engagement model where they are used to calculate the engagement state of the person in view. The system is designed to calculate the engagement state for one frame at a time.

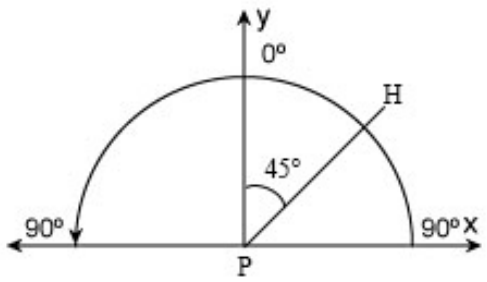


Figure 9: Top view of the set-up to illustrate the angle with respect to the FoV of Pepper.

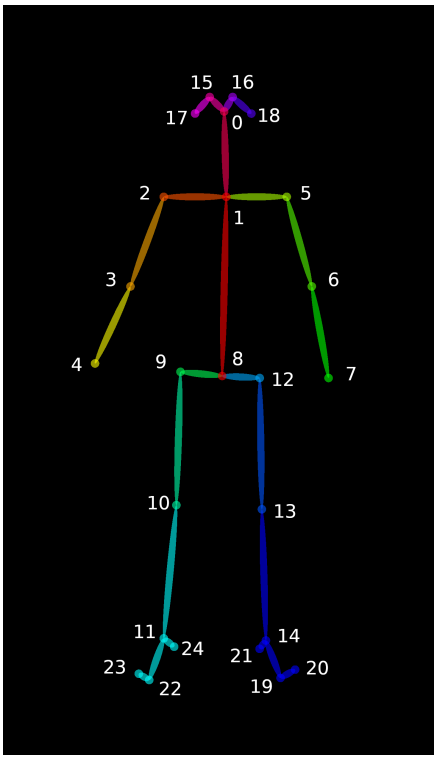
Video stream The frame from the video is processed by OpenPose[8], which is authored by Gines Hidalgo, Zhe Cao, Tomas Simon, Shih-En Wei, Hanbyul Joo, and Yaser Sheikh [4, 29, 5, 34]. To use this in our PSI network, we make use of an OpenPose wrapper written in C# by Takuya Takeuchi[9]. OpenPose gives 25 points representing the skeleton of the person visible in the frame. An example of this is shown in figure 10a, whereas the format and the identification of the points is shown in 10b. From the points given by OpenPose we calculate the gaze angle using the points representing the eyes and ears (15, 16, 17, 18). The facing direction is calculated using the shoulder points (2, 5). On the machine used for this thesis, OpenPose takes almost one second to generate the key points, so the frame rate is set to one frame per second so the classification can still be done in real-time.

People detection With the position we get from the people detection module we can calculate the velocity by comparing it with the position of the person in the previous frame. The distance is defined as the straight line distance between the origin of the robot frame

[8]OpenPose `https://github.com/CMU-Perceptual-Computing-Lab/openpose`.
[9]OpenPose wrapper in C# `https://github.com/takuya-takeuchi/OpenPoseDotNet`.

and the position of the person. This is then used to calculate the angle with respect to Pepper's field of view. Note that the field of view is based on the orientation of Pepper's base and does not change with the orientation of Pepper's head. The angle shows whether a person stands in front of Pepper, or further to either side. This is illustrated in figure 9 showing a top view of the area in front of Pepper (denoted with P). A human (H) is standing to the right of Pepper's base at an angle of 45°. When a person stands directly in front of Pepper, this is considered 0°and we've set the limit to 90°. It does not matter whether a person stands to the left or to the right.

(a) An example of the points representing the skeleton of people in an image.



(b) The format of the points.

Figure 10: Illustration of the key points given by OpenPose

Sound direction Pepper raises an event every 170ms if one or several sounds have been localised during that time frame. The angles provided by the sound source localisation engine match the real position of the source with an average accuracy of 10 degrees[10].

ID mapping To determine whether a sound originated from the person in view, we match it using the nose point we get from OpenPose (point 0). If one sound originates from within 10°of this point, we deem that the person is speaking. Aside from the sounds, we also have to map the IDs given by OpenPose with the IDs given by Pepper, since both those systems have a different way of assigning these. When classifying one person, we don't need explicit mapping. However, as mentioned before, this system is meant to classify one person at the time, yet is designed with the possibility to be extended to classify multiple people at once and therefore contains components with this in mind. The ID mapping component is one such component.

[10] ALSoundLocalization `http://doc.aldebaran.com/2-5/naoqi/audio/alsoundlocalization.html`

Engagement model This component is an implementation of the Naive Bayes classifier explained in section 3.3.2. Since Pepper sends its data and the PSI processes the data asynchronously, yet the classifier needs all features to make a classification, this component stores the features until it has received them all. Whether data has to be overwritten or ignored when new data arrives is handled at various stages in the network. The white components are the components we have setup an offline version of, which we use to evaluate the model.

5 Results

To evaluate our engagement model we use the k-fold cross validation techniques with $k = 10$. Our final data set consists of 1410 frames of which 728 are from one participant and 682 are from the other. Each participant has classified the frames of the other participant and we have classified the frames of both participants ourselves. We have discussed the differences in the classification and have come to an agreement for each frame. These classifications are considered the ground truth. The number of frames for each state in this ground truth are as follows: 469 are classified as *no interest*, 384 as *interest*, 504 as *engaged* and 53 as *ending interaction*. The confusion matrix of the 10-fold cross validation result is shown in table 5. From this matrix the performance metrics precision, recall and F1 score can be calculated. For each state separately, precision is the number of frames correctly classified as state S out of all of the frames classified as state S. Recall is the number of frames correctly classified as state S out of all the ground truth frames classified as state S. The F1 score, defined as $F1 = 2 * (precision * recall)/(precision + recall)$, is also calculated for each class separately. These results are shown in table 6. For the model as a whole we can calculate the macro-averaged precision, macro-averaged recall and macro-averaged F1 score as well as the weighted-average precision, weighted-average recall and weighted-average F1 score. Additionally, we can calculate the overall average. The macro-averaged metrics are computed as a simple arithmetic mean of our per-class precision, recall and F1 score respectively. The weighted metrics are the arithmetic means of our per-class metrics, for which each metric is weighted by the number of samples from its corresponding class. This is all shown in table 4.

	Precision	Recall	F1 score
Macro	0.701	0.705	0.698
Weighted	0.838	0.843	0.839
Accuracy		0.841	

Table 4: Performance metrics for the overall model.

6 Conclusion

In this thesis we have looked into the non-verbal behaviours that are involved in engagement, defined as the process by which interactors start, maintain and end their perceived connection to each other during interaction. In a practical setting, this definition calls for a

		True			
		No interest	Interest	Engaged	Ending interaction
Predicted	No interest	423	36	3	24
	Interest	38	294	38	12
	Engaged	1	32	458	6
	Ending interaction	7	22	5	11

Table 5: The confusion matrix of the 10-fold cross validation results.

	No interest	Interest	Engaged	Ending interaction
Precision	0.870	0.770	0.922	0.244
Recall	0.902	0.766	0.909	0.208
F1 score	0.886	0.768	0.915	0.224

Table 6: Performance metrics for each state.

way to distinguish between engagement and non-engagement (i.e. when a person has no intention to start an interaction or when a person considers the interaction to be concluded). We have therefore defined four states of engagement: *no interest* (or *no intention to interact*), *interest* (or *intention to interact*), *engaged* and *ending interaction*. The objective was to develop a model of the behaviour shown to indicate each of these states with the purpose of improving the way a robot determines who to engage with to contribution to improving an informative conversation between human and robot. As stated in section 2.4, to develop such a model several questions needed to be answered. Based on our literature study, developed model and experiments we can determine whether we can answer those questions and to what extend we have achieved our main goal. The first few questions regard the behaviour for each state. We can answer those based on the explored related work and the human frame, however due to the limits brought by the choice of robot and development of the network we can only comment on the sufficiency of a select few.

Which features sufficiently show when a person does not have the intention to interact? The features we consider important to classify the *no interest* state are *distance from the robot*, *facing direction*, *gaze* and *velocity*. Based on the results from our experiments we can conclude that these features are sufficient to classify *no interest*. Both the majority of the actual *no interest* frames are classified as such and few other states are classified as *no interest*. For this state, the precision and recall values differ more than for the other states.

Which features sufficiently show when a person does have the intention to interact? The features we have chosen for the state *interest* are *distance from the robot*, *facing direction*, *gaze* and *sound direction*. These features are also sufficient to classify the state, since the precision and recall values are also quite high and only differ slightly. Both values are a bit lower than for the *no interest* and *engaged* states however, so we will

elaborate on that in section 7.

Which features sufficiently show that a person wants to maintain their perceived connection? For the *engaged* state, we consider all of the features (*distance from the robot*, *facing direction*, *gaze*, *sound direction* and *velocity*) to be of importance. This state has the most similarities with the three other states. The significant features are therefore not only the prominent features based on the behaviour of the state itself, but also the features that key for humans in distinguishing between classes (e.g. whether someone is *speaking* can give the difference between *interest* and *engaged*). Based on these features, the results for this state are also promising. The overall score is even slightly higher than from the *no interest* state and the difference between precision and recall is slightly less.

Which features sufficiently indicate a person is ending the interaction? Setting up a suitable definition of the state *ending interaction* with clear boundaries which both covers the perception of humans and fits the structure of a Naive Bayes classifier proved to be impossible, since the state as defined by humans is more of a succession of actions rather than a state defined by one type of behaviour. Even the sole action of turning away, as which we have now defined this state, is not entirely independent of the state preceding it. Choosing features suitable for the classification of this state likewise proved to be difficult, partly by the lack of past research for this part of engagement. We have decided to use the features *distance from the robot* and *facing direction*. The results show that the classification of our model based on these features is vastly insufficient for classifying this state, as the majority of the frames are classified as *no interest*. Based on the definition and studied behaviour in past research as well as the difficulty of setting up a suitable definition of the state for our model, we argue that the main fault for this state lies in the setup of the model as opposed to the choice of features, however we do have reason to believe the chosen set of features is lacking as well. In section 3.4 we namely explain that we have discussed the classifications made by the participants. This discussion included how they would define the *ending interaction* state. We will further discuss this result and propose possible improvements in section 7.

Which approach can sufficiently classify the states of engagement using said features? To classify the four states using the chosen features, we have chosen to set up a Naive Bayes classifier. With an overall accuracy of 84% this model shows promise for three of the four states. The macro F1 score is 0.698, though when taking a closer look at the underlying scores per state, we can see that the model scores well on the states *no interest*, *interest* and *engaged* (with 0.886, 0.768 and 0.915 F1 scores respectively). The model however performs poorly on the state *ending interaction* (F1 score 0.224). Based on the definitions in past research, our own interpretation in human-human interaction as well as the discussion conducted with the participants of our experiment, we argue that the main cause of this poor performance is that the basic Naive Bayes classifier we have set up lacks the ability to take into account the temporal context of both the *engaged* state

that precedes *ending interaction* as well as the actions that humans consider make up the *ending interaction* state. We will further discuss this in section 7. Since the state *ending interaction* occurs less and for a far shorter period of time compared to the other states, we have taken this into account by also calculating weighted precision, weighted recall and weighted F1 score. These again shows that the model overall performs well.

Developing a model to detect engagement for a single person using spatial and multi-modal features Each previously answered question contributes to what extend we have achieved our main objective. We have developed a classification model using Naive Bayes, which can classify the four states of engagement using a set of five features. The overall performance of this model is good and it can accurately classify three of the four states. We argue that for the one state our model has a poor performance on, possible adjustments and extensions made to the model as well as a redefinition of the state can improve the performance sufficiently. Since the main purpose of the model is to contribute to improving an informative conversation between human and robot, we can conclude we have achieved our main objective.

7 Discussion and future work

Results The first point of discussion is the reason for the lower precision and recall values for the *interest* state. Both values are close to each other, which means about as much *interest* frames get classified as other states as the other way around. Looking at table 5 we can see that the number of false positives and false negatives are about equal for each state and that the most wrong classifications are generally from *no interest* and *engaged*. Additionally, since the model performs poorly on the *ending interaction* state, we can disregard this state. We speculate that the lower performance of *interest* is due to a fault and limit we discovered during our experiments. We have not been able to test this assumption in any way, however we can argue as to why we assume this. When we analysed our results, we found that for none of the frames, the sound direction feature indicated that the participant was speaking, even though we have numerous frames containing a conversation with the robot. We have tested the component that matches the sound with the person based on the frame on its own, however we were not able to test each component in the network, so we expect something is wrong with the integration of the component. In our discussion with the participants after our experiments, they both indicated that to determine when a person was *engaged* as opposed to in the state *interest*, they used whether the person was speaking as their main feature. Based on this we expect that whether a person is speaking or not is an important indication for whether a frame is considered *interest* or *engaged* for our model as well. We therefore think that this will largely solve the false positives and false negatives between those states. Regarding the misclassification of *no interest* with *interest*, we believe that this is partly due to the limit of Pepper's detection and tracking range being 2.5m. As mentioned before, the zone further than 3.6m is called the public zone and at an event, will be the zone in which the state

no interest can be mostly found. Of course it is possible and not unnatural for someone to stand closer while not having the intention to interact with Pepper, however the limit of Pepper required each scenario in our experiment setup to take place closer than 2.5m. Both participants have indicated that if they would not have been limited to 2.5m, they would have taken advantage of the extra space for some of the (parts of the) scenarios classified as *no interest*. They did note that not every part of the scenarios felt unnatural. We therefore expect that being able to expand the area in which people can be detected, will improve the performance for both the *interest* state as well as the *no interest* state, tough will not eliminate all false positives and false negatives.

The most important issue to discuss regarding the results is the poor performance of the *ending interaction* state. As we have briefly explained previously, we expect this poor performance to have two main reasons, namely the definition of the state and the inability to take into account the temporal context of the model as we have set it up. During our experiments, we asked the two participants to classify each others videos. We discussed their classifications afterwards to come to an agreement on each frame and hear they're interpretation of engagement and its states. The definition of *no interest* that resulted from this is as follows. *Ending interaction* starts when one of the interactors starts a closing sentence (e.g. 'Thank you for your time, goodbye.'), includes turning away from each other and end after a person has walked a certain distance away or switches to another action (e.g. looking at their phone or talking to someone else). We have not determined which distance is far enough, however, both participants indicated that it would be farther than the 2.5m limit to feel natural. This definition cannot be described by one type of behaviour and is more of a succession of three actions (speaking, turning away, walking away). This is where the inclusion of the temporal context is important. To accurately classify *ending interaction*, the model has to know which action or state preceded the current action. We intended that the model can be relatively easily expanded and therefore consider the inclusion of the temporal context a useful option for future research. Changing the definition of the state also has consequences for whether our current choice of features is sufficient. Currently we only determine whether a person is speaking or not, but do not register what exactly is said. To know whether a sentence is a closing sentence, we need at least some degree of speech recognition. We therefore need to incorporate a new feature. We assume that if we take on this new definition, once the model is expanded with the inclusion of the temporal context and the *speech recognition* feature, the performance for the *ending interaction* state will significantly increase.

Regarding the overall performance of the model, an accuracy of 0.84 means 16% of the frames is wrongly classified. Since the intent is to use this model at an event, it is important to know the implications on the scenarios that occur in this setting. We have not performed any tests to specifically look into this, however we speculate that the most errors occur around the transitions between states. We assume based on our own experience that humans take some time to (unconsciously) make the transition between states. In these moments, the occurrence of certain feature categories will differ less between states and deviate more from the general occurrence for the respective states. The difference between

the probabilities of both classes therefore also decreases, which decreases the certainty of the assigned state. Taking the increase in error, or decrease in certainty over time into account, may give the opportunity to predict a transition and adjust the behaviour of Pepper accordingly.

Future work Being able to incorporate the temporal context is not only useful for classifying *ending interaction*, but can also be incorporated in the classification of all the other states. Both due to human behaviour as well as physical limits (e.g. in distance or the way humans can move their body) one cannot transition from each state to each of the other states. For one person, the possible transitions we expect are illustrated in figure 11. Taking the previous state into account while classifying the current state can improve the classification. Understanding the relations between the states can also potentially give the opportunity to predict states to a certain extent. The presumed transition behaviour might support this as well. In turn, this might improve the behaviour of Pepper.

If the inclusion of the temporal context will not be implemented for between the states, it might improve the performance to implement a certain bias. Since we want to improve the way Pepper determines who to engage with or pay attention to, the states *interest* and *engaged* are more important than the other two. After all, it would be less rude to look at a person who is not interested than to look away from someone who is. Of course it would be even more rude to turn away from someone you are engaged in conversation with.

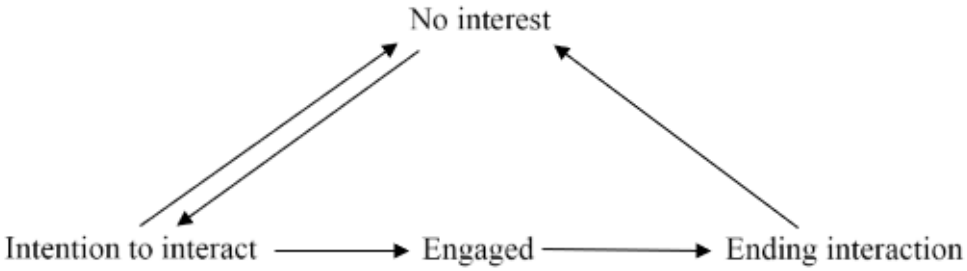


Figure 11: The engagement states and the transitions between them.

Possibly the most important next step for this system however, is to expand the model's ability to classify one person to classifying multiple people at the same time. The main reason we have reduced our scope to one person for this thesis, is that multiple people add multiple layers of complexity and given our time constraint as well as the amount of work the initial setup of the system required, classifying multiple people was not feasible. Wherever possible though, we have already incorporated the extensions we thought needed for multiple people both in the design and technical implementation. Each component in the network should be able to handle multiple people if all our assumptions hold true. In our experience however, in the field of computer science, that is never the case, so we would suggest to first (after making some final adjustments) run the model on multiple people at once, to see where those assumptions fall through. Once the technical side of the system works for multiple people, we need to determine which layer of complexity is most important

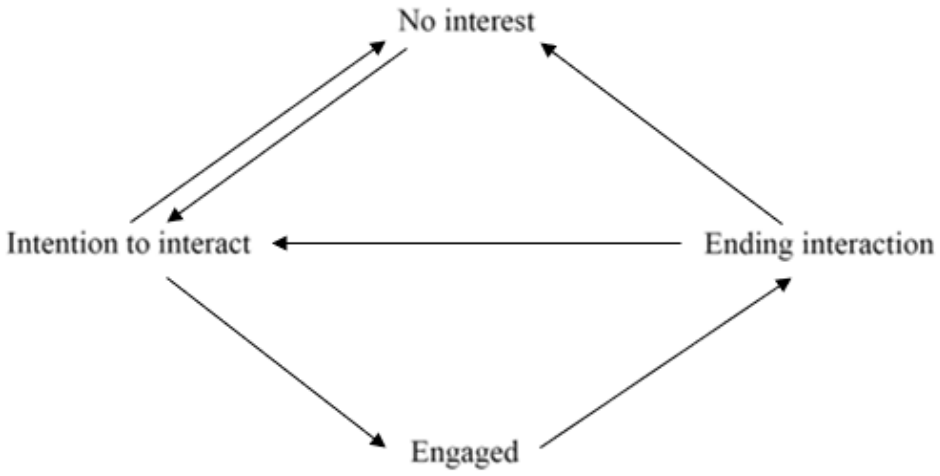


Figure 12: The engagement states and the transitions between them in a scenario with multiple people.

for the performance of the engagement model and what exactly this layer entails. One such layer adds to the transitions between states as illustrated in figure 12. When alone, it would be unnatural to not walk away at the end of an interaction, however when at least one other person is present, it is possible that the first person stays after ending their interaction (e.g. to listen to the conversation of the other person). This therefore adds another transition from *ending interaction* to *intention to interact* (or *interest*). Even though the person just finished an interaction, we can not rule out that the person might join in. This also means that the *ending interaction* state only consists of a closing sentence, which adds complexity to the classification of that state. An added benefit of classifying multiple people is that other humans may provide clues as to what state others are in. For example, in turn-taking, when one person 'holds the floor', the *gaze* of the other interactors is directed towards this person [2]. This can help Pepper determine who is the most prominent speaker. Since several of the systems in the papers covers in section 2 are able to classify multiple people at once, we expect them to prove useful, together with our personal experience in human interaction to expand the system while keeping its performance.

We have mentioned several times that our model has to contribute to improving an informative conversation between human and robot. As it stands, the output of the model is not used by Pepper. We can send the classified engagement state back to Pepper and develop a way to base Pepper's engagement behaviour on this state. This way we can evaluate whether the model has the desired effect of improving the naturalness of Pepper's behaviour in an informative conversation at an event and whether this can sufficiently add to Pepper's ability to interact with humans autonomously.

Another improvement we could make is taking over something Pepper already has incorporated to some extend in its own functionality. The 'PeopleDetected' event that we use to get the position of the person visible, only fires this event when the person is

visible by the cameras. The ALPeoplePerception[11] module however has other functionality. The module tries to find potential humans around the robot using visual cues. All new people detected in the current video frame are associated (when possible) with previously known people. This helps track someone and update his attributes. When somebody gets out of the field of view, he/she is not immediately removed form the people list, as this disappearance may be temporary and the result of the robot movements. This concept would be useful for the engagement model as well, especially if the temporal context becomes more important and the scenarios Pepper is put in contain more people.

Naive Bayes' conditional independence The reason the Bayes classifier we use is called naive is because it assumes that the features are independent. It is verified that the the Bayesian classifier performs quite well in practice even when strong attribute dependencies are present [9], however these dependencies do influence the probability estimates, possibly to the point where its no longer usable for comparison between the humans in Pepper's presence. We have assumed that the features used are independent enough to use the Naive Bayes classifier, considering that each of the features does not directly depends on the result of another. However we comprehend that human behaviour may create unconscious dependencies, for example between the *facing direction* and *gaze*. This could potentially cause a deviation of the final score from the actual probability estimate, which in turn could cause the wrong person to be picked as the prominent engagement partner. Since engagement behaviour is not well understood in the human-human context, it is best to take this into account while testing the model on multiple people.

Discussion with the experiment participants At the end of our experiments we discussed the classifications made by the participants and asked them which features they based those classifications on. Since we have not told them which features we have used for our model, this gave us some insight in the behaviour other humans pay attention to when it comes to engagement. Aside from the comments about the distance and the insights given on the definition of *ending interaction* and the incorporation of speech recognition, they both mentioned that the vertical gaze angle is an important feature to distinguish between the *no interest* and *interest* states. It suggest a person shows *no interest* either by looking down at something (e.g. their phone) or looking over Pepper's head at something behind the robot. In turn it suggests *interest* when the gaze angle is directed towards Pepper. Both participants also confirmed our assumption that whether a person is speaking or not is an important feature to distinguish between *interest* and *engaged*.

8 Acknowledgements

We would like to thank the ING and its interns for contributing to parts of our system and the client.

[11] ALPeoplePerception module documentation `http://doc.aldebaran.com/2-5/naoqi/peopleperception/alpeopleperception.html`.

References

[1] Salvatore M Anzalone, Sofiane Boucenna, Serena Ivaldi, and Mohamed Chetouani. Evaluating the engagement with social robots. *International Journal of Social Robotics*, 7(4):465–478, 2015.

[2] Dan Bohus and Eric Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 1–8, 2010.

[3] Rodney A Brooks. New approaches to robotics. *Science*, 253(5025):1227–1232, 1991.

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[6] Kerstin Dautenhahn and Claude Ghauoi. The encyclopedia of human-computer interaction, 2014.

[7] Alessandro Di Nuovo, Frank Broz, Filippo Cavallo, and Paolo Dario. New frontiers of service robotics for active and healthy ageing, 2016.

[8] Alessandro Di Nuovo, Frank Broz, Ning Wang, Tony Belpaeme, Angelo Cangelosi, Ray Jones, Raffaele Esposito, Filippo Cavallo, and Paolo Dario. The multi-modal interface of robot-era multi-robot services tailored for the elderly. *Intelligent Service Robotics*, 11(1):109–126, 2018.

[9] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2-3):103–130, 1997.

[10] N Earle and C Beardon. The role of obligation within virtual encounter. In *Proceedings of CVE*, pages 57–66, 1998.

[11] Mary Ellen Foster, Andre Gaschler, and Manuel Giuliani. Automatically classifying user engagement for dynamic multi-party human–robot interaction. *International Journal of Social Robotics*, 9(5):659–674, 2017.

[12] Edward Twitchell Hall. *The hidden dimension*, volume 609. Garden City, NY: Doubleday, 1966.

[13] Takayuki Kanda. Natural human-robot interaction. In Noriaki Ando, Stephen Balakirsky, Thomas Hemker, Monica Reggiani, and Oskar von Stryk, editors, *Simulation, Modeling, and Programming for Autonomous Robots*, pages 2–2, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[14] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. May i help you?: Design of human-like polite approaching behavior. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 35–42. ACM, 2015.

[15] David Klotz, Johannes Wienke, Julia Peltason, Britta Wrede, Sebastian Wrede, Vasil Khalidov, and Jean-Marc Odobez. Engagement-based multi-party dialog with a humanoid robot. In *Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, number CONF, 2011.

[16] Mark L Knapp, Roderick P Hart, Gustav W Friedrich, and Gary M Shulman. The rhetoric of goodbye: Verbal and nonverbal correlates of human leave-taking. *Communications Monographs*, 40(3):182–198, 1973.

[17] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. Receptionist or information kiosk: how do people talk with a robot? In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 31–40. ACM, 2010.

[18] Nikolaos Mavridis. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35, 2015.

[19] Christophe Mollaret, Alhayat Ali Mekonnen, Isabelle Ferrané, Julien Pinquier, and Frédéric Lerasle. Perceiving user's intention-for-interaction: A probabilistic multimodal data fusion scheme. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2015.

[20] Iroju Olaronke, Ojerinde Oluwaseun, and Ikono Rhoda. State of the art: a study of human-robot interaction in healthcare. *International Journal of Information Engineering and Electronic Business*, 9(3):43, 2017.

[21] Vittorio Perera, Tiago Pereira, Jonathan Connell, and Manuela Veloso. Setting up pepper for autonomous navigation and personalized interaction with users. *arXiv preprint arXiv:1704.04797*, 2017.

[22] Shokoofeh Pourmehr, Jack Thomas, Jake Bruce, Jens Wawerla, and Richard Vaughan. Robust sensor fusion for finding hri partners in a crowd. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3272–3278. IEEE, 2017.

[23] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L Sidner. Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 375–382. IEEE, 2010.

[24] Ben Salem and Nic Earle. Designing a non-verbal language for expressive avatars. In *Proceedings of the third international conference on Collaborative virtual environments*, pages 93–101, 2000.

[25] Satoru Satake, Takayuki Kanda, Dylan F Glas, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. How to approach humans?: strategies for social robots to initiate interaction. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 109–116. ACM, 2009.

[26] Chao Shi, Michihiro Shimada, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Spatial formation model for initiating conversation. *Proceedings of robotics: Science and systems VII*, pages 305–313, 2011.

[27] Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Interactive humanoid robots for a science museum. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 305–312. ACM, 2006.

[28] Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164, 2005.

[29] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[30] Fumihide Tanaka, Kyosuke Isshiki, Fumiki Takahashi, Manabu Uekusa, Rumiko Sei, and Kaname Hayashi. Pepper learns together with children: Development of an educational application. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 270–275. IEEE, 2015.

[31] Meg Tonkin, Jonathan Vitale, Suman Ojha, Mary-Anne Williams, Paul Fuller, William Judge, and Xun Wang. Would you like to sample? robot engagement in a shopping centre. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 42–49. IEEE, 2017.

[32] Dominique Vaufreydaz, Wafa Johal, and Claudine Combe. Starting engagement detection towards a companion robot using multimodal features. *Robotics and Autonomous Systems*, 75:4–16, 2016.

[33] Hannes Högni Vilhjálmsson. *Autonomous communicative behaviors in avatars*. PhD thesis, Massachusetts Institute of Technology, 1997.

[34] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[35] Zerrin Yumak, Bram van den Brink, and Arjan Egges. Autonomous social gaze model for an interactive virtual character in real-life settings. *Computer Animation and Virtual Worlds*, 28(3-4):e1757, 2017.