The Attention-Hesitation Model

a Non-Intrusive Intervention Strategy for
Incremental Smart Home Dialogue Management

Birte Richter

The Attention-Hesitation Model

a Non-Intrusive Intervention Strategy for Incremental Smart Home Dialogue Management

Birte Richter

2021

A doctoral thesis presented for the degree of
Doctor of Engineering (Dr.-Ing.) at

Faculty of Technology
Bielefeld University
Inspiration 1
33619 Bielefeld
Germany

Reviewers

Prof. Dr.-Ing. Britta Wrede
Prof. Dr. Katharina Rohlfing

Examination Board

Prof. Dr.-Ing. Ulrich Rückert
Dr. Olga Abramov

Defended and approved at May 20, 2021

Printed on permanent paper as per ISO 9706.

ACKNOWLEDGMENTS

Even if you write a dissertation alone, you are not a lone fighter. I received a lot of support and help while working on this dissertation. First of all, I would like to thank Britta Wrede and Gerhard Sagerer who gave me the opportunity to do my doctorate in the Applied Informatics Group and the Cluster of Excellence Cognitive Interaction Technology (CITEC) EXC 227 at Bielefeld University, founded by the German Research Foundation (DFG). Many thanks to my supervisor Prof. Britta Wrede, who gave me great support through many conversations and discussions. I would also like to thank Prof. Katharina Rohlfing for agreeing to be my second reviewer and Prof. Ulrich Rückert and Dr. Olga Abramov for their work in my examination board. My colleagues in the Applied Informatics Group and the CITEC welcomed me very warmly and were always available to offer advice and assistance. I will miss the exciting discussions in front of the coffee machine and during lunch. In addition, I was able to get to know many new perspectives through many meetings in Prof. David Schlangen's Dialogue System Group. Especially in the last hot phase, the people from our writing workshop repeatedly managed to dispel any lack of motivation, I would especially like to thank Eva, Jasmin, Mara, Marian, Sebastian, Thomas and Viktor for this from the bottom of my heart. Very special thanks go to the CSRA team, including Viktor, Marian, Christian, Patrick, Norman, Jasmin, Sabastian, Alex and Jan. Without you and your technical components, I would never have been able to do my work in the CSRA. I learned so much during that time - about my work but also about myself. It was fun with you guys! Also a big thank you to my assistants over the years, especially to Monika and Ayla, who have supported me in my many interaction studies. Without you I would probably still be looking for participants today. I would also like to thank my 190 participants for their time, courage and patience. In addition, a big thank you to my proofreaders Viktor, Jessica and Britta.

Apart from the professional help, I would like to thank my friends and family from the bottom of my heart. You always managed to get me thinking about something else in the evenings in our kitchen or the weekly family dinner. Thank you for your support, that you never doubted me, no matter how long it took and that you are always there for me! My biggest thanks go to Viktor—for the non-enumerable help and support in recent years and the many other little things that make my life worth living.

ABSTRACT

Smart homes are one of the most emergent research fields and provide fundamentally new means of interaction. So-called *Smart Personal Assistants (SPAs)* entered the household and assist us in our daily activities. Currently, these agents do not react to the attention of the smart home user. However, from *Human-Human Interaction (HHI)* research we know that humans coordinate their speech and adapt their behavior continuously, based on their interaction partner's actions and reactions. Therefore, the central question I ask in this dissertation is how human attention can be incorporated into dialogue management, to improve *Human-Agent Interaction (HAI)* in smart homes.

Research shows that speakers' hesitations are often produced as a reaction to the listener's inattentiveness in *HHI*. Furthermore, they can improve the listeners' comprehension. Therefore, I investigate whether it is possible to use system hesitations, based on the attention of the human interaction partner, as a communicative act for dialogue coordination in *HAI* within a smart-home environment. To this end, I develop a theoretical model based on observations from *HHI*, implement it in an autonomous agent and evaluate it in five interaction studies.

This document consists of three parts. In the first part, I develop a model which allows the dialogue management to incorporate the human attention: the *Attention-Hesitation Model (AHM)*. The model uses system hesitations as a non-intrusive *intervention strategy* to coordinate the human attention with system speech. This theoretical model is based on interdisciplinary literature from *HHI* and *HAI* research.

In the second part, I elaborate on the technical requirements implied by the integration of the *AHM* in an autonomous system. A technical realization of an incremental dialogue system in presented. Two main concepts for dialogue modeling are identified: (1) the use of interaction patterns with system task descriptions for generalizability and (2) the concept of the IU model to deal with the incremental nature of human dialogue. With the combination of the frameworks Pamini and inprotk both concepts are considered in my dialogue system. This allows autonomous *HAI* and the investigation of the effects of my *AHM* in interaction.

In the third part, I evaluate the effects of my *AHM* on the interaction (partner) in five *Evaluation Cycles (ECs)*, consisting of three pilot- and two *HAI* studies in a smart-home environment. In these cycles, I further enhance my model, its implementation, and the experimental design. Thereby, I investigate the effect of the *AHM* on the task

performance and the side effects in interaction: the subjective ratings of the agent and the visual attention of interlocutors.

With my investigations, I show that in short interactions without a change of discourse, the participants interacting with an agent that uses my *AHM* are significantly less inattentive than participants in the baseline (EC1). Furthermore, I show that the *AHM* can work fully autonomously (EC2, EC4). Regarding the task performance, I demonstrate that participants interacting with an agent that uses my *AHM* perform significantly better in some practical tasks than participants in the baseline (EC3-EC5). This effect is, however, accompanied by lower subjective ratings of the agent (EC2-EC4). The ratings show that repetitions can be perceived as annoying (EC2) and users may struggle with the differentiation of unfilled pauses from turn-ends in more complex scenarios (EC2, EC3). However, the use of lengthening may counteract this problem and enhance some subjective ratings (EC4). The final model uses mutual gaze and task related features to distinguish inattentiveness based on (1) missing engagement from (2) difficulties in understanding. To deal with inattention based on missing engagement, a cascade of lengthening, unfilled pauses, and hesitation vowels is used. For difficulties in understanding, the model uses repetitions with lengthening. This combination improves the task performance without negative side effects on the interaction (EC5).

CONTENTS

LIST OF FIGURES

LIST OF DEFINITIONS

AAR	Automatic Addressee Recognition. xiii, 76
ADHD	Attention Deficit Hyperactivity Disorder. xiii, 237
AHM	Attention-Hesitation Model. v, vi, xiii, 13, 15–17, 59, 62, 64, 65, 71, 74–76, 80, 81, 85–87, 89, 102, 106, 113, 117, 127, 128, 131, 135, 137, 139–142, 149, 151–153, 159, 160, 173–176, 187–190, 194, 197, 200, 203, 204, 206–212, 214–223, 225, 226, 228–231, 234, 236, 237
ANOVA	Univariate Analysis of Variance. xiii
ASR	Automatic Speech Recognition. xiii, 52, 72–74, 76, 82, 86, 92, 96, 98, 104
AVM	Attribute Value Matrices. xiii, 119
BCO	Base Cube One. xiii, 93, 94
BN	Bayesian Network. xiii, 79
CITK	Cognitive Interaction Toolkit. xiii, 93, 124
CSRA	Cognitive Service Robotics Apartment. xiii, 89–93, 100, 103, 106–110, 123, 124, 131, 133, 142, 145, 153, 155, 165, 174, 178, 180, 192, 194, 225, 228, 233
DM	Dialogue Management. xiii, 52, 61, 73–78, 83, 85–87, 93, 94, 98, 100, 107, 161, 194, 223
DNN	Deep Neural Networks. xiii, 74
DoF	Degrees of Freedom. xiii, 41, 90
DOR	Diagnostic Odds Ratio. xiii
EC	Evaluation Cycle. v, xiii, 15, 16, 117, 123, 126–129, 131, 153, 180, 209, 216, 225, 226, 228, 230, 233, 235
EC1	Evaluation Cycle 1: Self-interruptions as Attention-regain Strategy. xiii, 127, 137, 138, 146, 188, 209–212, 215, 216, 218, 220, 221, 225, 229, 230, 307
EC2	Evaluation Cycle 2: Introducing the Focus of Discourse Feature. xiii, 127, 142, 160, 190, 209–212, 214–216, 218, 220, 225, 226, 229–231, 234, 309
EC3	Evaluation Cycle 3: Exploration of a Practical Task during Interaction. xiii, 128, 189, 190, 195, 210–215, 218, 223, 225, 226, 229, 230, 236, 311

EC4	Evaluation Cycle 4: Introducing the Lengthening Feature and new Evaluation Approach. xiii, 128, 177, 190, 199, 209–218, 220, 223, 225, 226, 229, 230, 234
EC5	Evaluation Cycle 5: Bringing It All Together. xiii, 128, 191, 209–216, 218–220, 223, 226, 229–231, 236
FoD	Focus of Discourse. xiii, 16, 23, 28, 37, 41, 61, 87, 103–105, 127, 128, 133, 138–143, 145, 146, 150–152, 173, 184, 188–191, 203, 206, 209, 213, 220, 221, 223, 229
GMM	Gaussian Mixture Models. xiii, 74
GQS	Godspeed Questionnaire Series. xiii, 118, 126, 136, 138, 171, 184, 198, 216, 287, 311
HAI	Human-Agent Interaction. v, xiii, 7, 10–15, 19, 24, 26, 38, 39, 41, 44, 48–50, 53–55, 57, 58, 67, 77, 78, 80, 81, 85, 86, 88, 89, 92, 100, 106, 113, 117, 119, 121–124, 127, 133, 135, 136, 142, 143, 146, 158, 165, 168, 180, 210, 213, 221, 225–229, 231, 233, 234, 237
HHI	Human-Human Interaction. v, xiii, 6, 7, 9, 10, 12, 14, 15, 19, 24, 28, 31, 37–39, 48, 51, 54, 55, 57, 58, 62, 67, 76, 80, 117, 155, 158, 165, 166, 176, 206, 212, 213, 226, 227, 232, 234
HLRC	High Level Robot Control. xiii, 94
HMM	Hidden Markov Model. xiii, 48, 74, 167
HRI	Human-Robot Interaction. xiii, 6, 24, 25, 39, 42, 47, 54, 78–80, 87, 118–122, 126, 146
HSI	Human-System Interaction. xiii
IDL	Interface Definition Language. xiii, 93
Inprotk	Incremental Processing Toolkit. xiii, 83, 94, 96, 98, 100, 102–104, 107, 113, 133, 178, 228, 233
IP	Interaction Pattern. xiii, 100
IPA	Intelligent Personal Assistant. xiii, 5
IU	Incremental Unit. xi, xiii, 52, 81–83, 94, 98, 104, 107, 113, 178, 192, 228, 233, 234
MANOVA	Multivariate Analysis of Variance. xiii, 126, 127, 138, 150, 172, 184
MORSE	Modular Open Robots Simulation Engine. xiii, 91
NARS	Negative Attitudes Toward Robots Scale. xiii, 118
NLG	Natural Language Generation. xiii, 73, 75, 77

NLU	Natural Language Understanding. xiii, 73–77, 82, 86, 98, 232
Pamini	Pattern Based Mixed Initiative Interaction Toolkit. xiii, 78, 80, 87, 93, 94, 98, 100–102, 105, 107, 113, 154, 159, 161, 228, 233
POMDP	Partially Observable Markov Decision Processes. xiii, 79
ROS	Robot Operating System. xiii
RSB	Robotics Service Bus. xiii, 93, 96, 100, 104, 105
RST	Robotics Systems Types Repository. xiii, 93, 96, 104
SDS	Spoken Dialogue Systems. xiii, 119
SPA	Smart Personal Assistant. v, xiii, 5–9, 80, 106, 234, 235
T-Test	Welch two-sample t-tests. xiii, 126, 150, 167, 196, 197, 307, 309, 311, 314
ToM	Theory of Mind. xiii, 5
TSP	Task State Protocol. xiii, 100
TTS	Text-to-Speech Synthesis. xiii, 53, 74, 86, 94, 104, 217, 219, 232
TVA	Theory of Visual Attention. xiii, 22
VAD	Voice Activity Detection. xiii, 76
VFoA	Visual Focus of Attention. xiii, 9, 23, 26–28, 45–47, 57, 58, 60–62, 67, 96, 103, 120, 126, 132, 135, 137–139, 141, 142, 145, 148, 150–152, 175, 176, 184, 185, 190, 198, 203, 220, 222
WoZ	Wizard-of-Oz. xiii, 44, 52, 53, 57, 121, 123, 124, 127, 131, 132, 165, 233
WR-Test	Wilcoxon Rank Sum Test / Mann-Whitney U tes. xiii, 127, 137, 138, 147, 183, 184, 186, 198, 314, 316
WSR-Test	Wilcoxon Signed Rank Test. xiii, 127, 148, 205, 316

Part I

FROM HHI TO HAI: DEVELOPING A MODEL

1 INTRODUCTION

Interaction is a joint action. To exchange and share knowledge at least two interaction partners are needed who are willing and able to coordinate their interaction. In this respect, it makes no difference whether the conversational partners discuss an important issue, just do chitchat, or try to teach or learn something. As Goodwin pointed out:

> "To engage successfully in conversation, participants are required not only to produce sentences but also to coordinate, in a meaningful fashion, their talk with the talk of others present." [Goo81]

This is in line with Clark, who proposed that using language is a joint action [Cla96]. This action needs the coordination between the speaker and the listener, and for Clark, this always involves the speaker's meaning and listener's understanding [Cla96]. To understand each other and have a meaningful conversation, it is necessary to share knowledge and beliefs. In this context, a connection to the *Theory of Mind (ToM)* can be established. The *ToM* refers to the cognitive ability to attribute mental states to self and others [FF05]. These mental states include among others: beliefs, goals, perceptions and desires. According to this theory, people can recognize that another person's knowledge is different from their own, which allows them to influence other people's behavior by manipulating their beliefs. One of the critical precursors to these skills is joint (or shared) attention: the ability to selectively attend to an object of mutual interest [FF05].

Interaction with intelligent agents is becoming more and more important to us. We already interact with robots in various contexts. In future-oriented projects, such as *Industry 4.0*, smart factories are supported and human–robot collaboration in industrial settings have already partially become reality [Vil+18]. In these smart factories, a robot changes from being solely a tool to being an assistant and partner. However, robots become more important in other areas of our lives too. They are used in healthcare [Van16], education [Bel+18], or crisis management [KG17]. Furthermore, they can be found in our household, for example in the form of cleaning robots (e.g., [Cor19b]) or entertainment tools (e.g., [Cor19a]). Besides these robots, other agents entered our households. Virtual agents and smart devices built and programmed to support us in our daily life and make our home smart. So-called *SPAs*[1], such as Google Assistant [Lim20], Amazon

1 Sometimes called *Intelligent Personal Assistants (IPAs)*.

Alexa [Ama], or Microsoft Cortana [Mic21] entered the household and assist us in our daily activities. According to Knote et al. these commercial assistants represent the vast majority of *SPAs* [Kno+19]. *SPAs* "rely on emerging technologies, such as natural language processing and artificial intelligence" [Kno+18]. Smart homes in general are one of the most emergent research fields and provide fundamentally new means of interaction [And+16]. Smart home devices exist for better energy management, improved security, and assisted living (see [Zai+18] for an overview). In the future, the interaction with artificial agents at home will increase.

We already interact with *assistive systems* in various ways, e.g., via GUIs, speech, gestures, or biological electro signals [LLM15]. Li et al. argue that

> "[i]n order to provide immersive using experience, voice and vision controllers are widely employed. Since the experience using these is more close to the interaction between humans, it is more natural and effective." [LLM15]

I agree that a form of interaction that is close to the interaction with humans has several benefits, although that needs to be investigated more closely. It is to be expected that a natural interaction is easier for us, as we do not have to read long operating instructions or work ourselves through tutorials in order to be able to interact with assistive systems. Rather, such systems should be intuitively usable. In their survey about socially interactive robots, Fong et al. list concepts from *HHI* which have already influenced the *Human-Robot Interaction (HRI)* in order to make the interaction more human-like. These concepts are manifold, starting from the general appearance of the robot, the use of facial expressions and body movements, up to the multimodal expression of emotions, to only mention a few aspects [FND03]. Referring to dialogue, Fong et al. state: "regardless of form, [it] is meaningful only if it is grounded, i.e., when the symbols used by each party describe common concepts" [FND03][2]. Dautenhahn argues that the requirements for social skills increase with the frequency and nature of contact with humans [Dau07]. Furthermore, Dautenhahn state that "social intelligence [...] might bring us closer to the goal of making robots smarter (in the sense of more human-like and believable in behaviour)" [Dau07]. This has been confirmed by several researchers. De Ruyter et al. for example showed that an agent with some social intelligence, e.g., that mimics facial expressions, enhances the user acceptance for the dialogue system [De +05]. Other researchers found that social abilities contribute to the sense of social presence [Hee+08], or that the application of the spatial model (gained from *HHI*) to a humanoid robot results in a system that is perceived as more interested

2 The concept of grounding by Clark, Brennan, et al. [CB+91] will be discussed further in section 2.1.2.

in the human, and additionally shows its attention and intentions earlier and to a higher degree [HPW11]. Also in the field of intelligent virtual agents, concepts known from *HHI* are evaluated in *HAI*, especially regarding the human-like expressiveness of the agent and its non-verbal behavior [Nor+18].

Verbal interaction has the advantage of being "hand free". In addition, in several studies users reported that their preferred interaction modality is speech [CKB18; Ber+16]. By now we can already control smart devices via speech and verbal requests for information are no longer a rarity. Alexa, Google and other *SPAs* can perform such tasks, but according to Reis et al.

> "[there] is a gap between what the IPAs can do and what they effectively do, when confronted with a user expecting to fully execute an interaction in a natural conversational manner." [Rei+18]

It is still an open question how the verbal interaction in smart homes can be made more natural. We need to further investigate which concepts of *HHI* are transferable to *HAI* and especially how this influences the modeling of dialogue coordination.

Currently, the dialogue management of artificial agents is based on simple command-based structures (see example 1.0.1). We state simple commands to, e.g., switch our smart home lights or ask for specific information.

Example 1.0.1: Four example interactions between a human (H) and a SPA (A) in smart homes.

1)	H:	Alexa?
		SPA blinks.
	H:	Switch the light off.
		All lights go off.
2)	H:	Ok Google.
		SPA blinks.
	H:	How is the weather?
	A:	It's 18 degrees and raining.

3) H: Computer.
SPA blinks.
H: What does interaction mean?
A: According to Wikipedia, interaction is a kind of action that occurs as two or more objects have an effect upon one another. The idea of a two-way effect is essential in the concept of interaction, as opposed to a one-way causal effect. Closely related terms are ...

4) H: Computer.
SPA blinks.
H: How do I make muffins?
A: First, you need 2 medium eggs, 125ml vegetable oil, 250ml semi-skimmed milk, 250g golden caster sugar, 400g self-raising flour (or same quantity plain flour and 3 tsp baking powder), 1 tsp salt and 100g chocolate chips or dried fruit such as sultanas or dried cherries (optional)...

This structure works well for short command-response interaction scenarios and already has attention coordination mechanisms. Through the so-called hot-word, the agents are activated and show their readiness for the interaction through blinking lights and a short tone. However, besides their own visualization of attentiveness, these agents do not react to the attention of the user. As [Goo81] showed in his research, speakers adjust their speech based on listener's feedback signals, or do not start speaking unless they are sure they have the listener's attention [Goo81][3] I am convinced that if we want to change agents in smart homes from solely a tool to an *assistive system*, it is necessary to coordinate the system's speech with the human attention. To be able to support people with more complex tasks or to interact with people for longer, e.g., for longer explanations, the *assistive agent* should ensure that the person listens carefully and can understand. Currently, these agents simply continue speaking, regardless of whether they have the attention of the user or not. At the moment, *SPAs* can control parts of the smart home, but they do not use the several sensors within it. Therefore, a situated and multi-modal interaction is not possible. To assist people in their daily activities and have a natural, situated interaction, it is necessary to observe and monitor the interaction partner in more than the verbal modality. Other modalities provide important indicators, to find the right moment to act and react. Most assistants don't recognize, if the

3 This will be discussed in more detail late in this thesis.

user disengages, does not understand the given information, or is distracted and therefore not receptive. However, from *HHI* we know how important other modalities are in the interaction. For instance, with facial expressions we express emotions and cognitive states, e.g., understanding or thinking, which play an important role in everyday social interactions [OKJ06; Ekm04]. Other signals can also communicate information about a person's inner state or provide additional information, e.g., disambiguate referenced objects via pointing gestures [McN00]. The *Visual Focus of Attention (VFoA)* of the interlocutor plays an important role during conversations. It is used to manage the beginning of an interaction [Goo81], serves as a turn-talking signal [Ken67; ACC94] and provides additional information (e.g., for disambiguation) [AS17]. To achieve mutual understanding and a common grounding, it is crucial to have a representation of the attention state of the interaction partner. Based on gazing behavior or other social queues like facial expressions, and with the help of the measured task progress, it is possible to get insights into the human attention state. This multi-modality, the representation of the estimated human mental state, and an appropriate reaction on their basis, are—in my opinion—necessary to coordinate a situated and natural dialogue in smart homes. As Goodwin and Clark pointed out, the observation of the interaction partner and the reaction and adaptation of one's own behavior is a precondition for successful interaction [Goo81; Cla96]. In smart home interaction, this is currently not the case. Commercial *SPAs* mostly react solely on speech input and do not incorporate the human attention into their dialogue system. Especially to be able to perform more complex interactions, they should use additional sensors to improve the interaction. Such a situated dialogue is one of the design guidelines for future interaction that we have gained based on an investigation of user problems and improvement requests for current interaction capabilities of *SPAs* [Hux+19].

1.1 SMART HOME INTERACTION

The interaction with *assistive systems* in smart homes becomes more important to us [LLM15]. According to the Oxford Dictionary, interaction is defined as a "reciprocal action or influence", and further, a "communication or direct involvement with someone or something" [Pre19b][4]. The second part of this definition shows that interaction not only occurs between humans. It is also possible to communicate with *assistive systems*, such as robots, virtual agents, or a smart home speaker. I differentiate in this thesis between several forms of dyadic interaction depicted in fig. 1.1. Besides interacting with each other

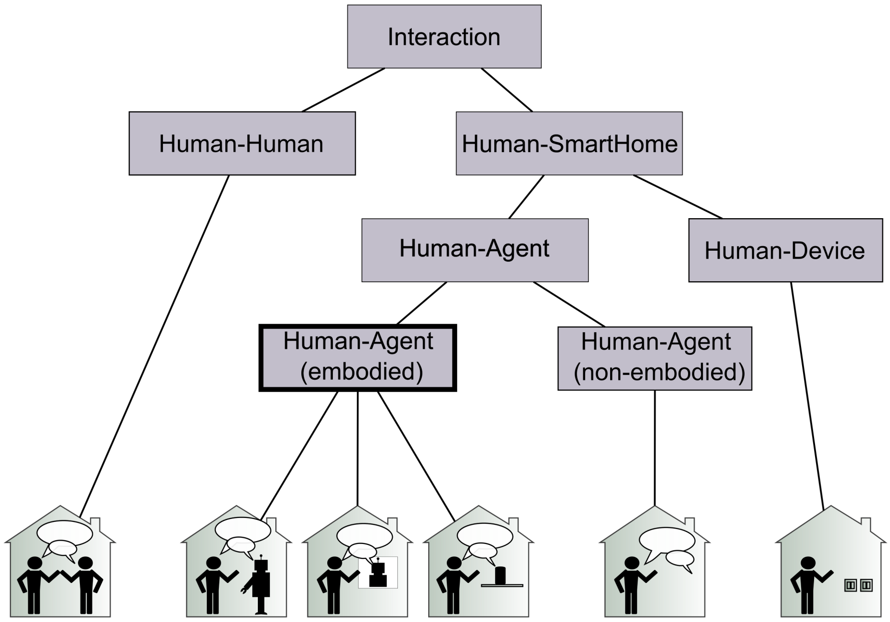


Figure 1.1: Interaction divided into several forms, depending on the type of the interaction partner (not completed).

in *Human-Human Interaction (HHI)*, humans can interact with smart homes through embodied or non-embodied agents or other smart devices, such as graphical user interfaces or smart switches. Verbal interaction is mostly realized through *HAI*. This is in line with the definition by the *HAI* community, where an agent "is an object or technology that people interact with as if it is able to act with its own purposes, motivations, and intentions"[Int19]. Of course, an agent may not necessarily have an embodiment. It is also possible to have an agent without an embodiment[5]. However, usually these artificial intelligences need an interface, e.g., a smartphone. In such cases, the distinction between embodied and non-embodied is more difficult. Furthermore, it is possible to interact with the smart home nonverbally in several ways through other smart devices, e.g., smart switches. Such

4 Other definitions focus only on sub-aspects, e.g., [Pre19a; Mer19c].
5 One example from science fiction is Marvels J.A.R.V.I.S..

a distinction is quite vague, as these devices can also be interpreted as part of the agent in smart homes. When people start talking into the room to start their coffee machine, it is difficult to distinguish whether they intend to interact with an agent in the form of the coffee machine, the microphone near the coffee, or the smart home itself.

Of course, this classification of interaction is incomplete. Various other interactions in smart homes involving humans (e.g., with wearables, or other living beings) and interactions without a human (e.g., robot-robot interaction) are not addressed in this thesis. I focus mainly on the embodied part of human interaction with an artificial agent, whether it is a robot, or a virtual agent, e.g., in the shape of a smart speaker. For simplification, in this thesis the term *HAI* describes interaction with embodied agents unless otherwise stated.

Working definition: Human-Smart-home Interaction

Interaction with smart homes occurs through embodied, non-embodied agents, or smart devices. In the following, HAI refers to an interaction between a human and a (virtual) agent, robot, or other embodiment, e.g., in the shape of a smart speaker.

The other important aspect of the definition of interaction in Press is that it is a reciprocal action. Interaction partners have an influence on each other. How interaction partners behave—how, what, and when they say something or which other actions they perform to communicate—has a direct influence on the other(s). For interaction design, we must always keep this in mind. If we want to incorporate the human attention into *HAI*, we need strategies to (re-)act to missing attention. Throughout this dissertation, the term *intervention strategy* will be used to refer to such reactions.

Working definition: Intervention strategy

Intervention strategies are reactions to unexpected behavior of the interaction partner. A re-attention intervention strategy should coordinate the human attention, by dealing with or regaining the missing attention of interlocutors.

1.2 RESEARCH QUESTION AND HYPOTHESIS

It is still an unsolved question how the coordination of natural *HAI* can be modeled. Currently, it is not possible to communicate with *assistive systems* in the exact same manner as with other humans. However, by enhancing our models, we can get closer to natural *HHI*. As repeatedly shown in the literature on *HAI*, concepts from *HHI* can be successfully transferred to interactions with artificial agents. In my thesis, I want to achieve this for the modeling of dialogue coordination. Therefore, I investigate the following research question, by taking two perspectives into account: (i) the **cognition motivated** research and the (ii) **software engineering** research perspective:

Research question: *How can human attention be incorporated into dialogue management, to improve the human-agent interaction in smart homes?*

My research question can be further split into the following sub-questions:

RQ 1: Model

How to model the coordination of human attention and system speech?

The component that coordinates dialogue in speech systems is the dialogue manager. It has two main responsibilities, (1) decide *when to (re-)act* and (2) *how to (re-)act*. In this thesis, I develop a model which serves as a module within the dialogue management to incorporate the human attention into it: I call it the Attention-Hesitation Model (AHM). The model uses system hesitations as a non-intrusive *intervention strategy* for the coordination of the human interaction partner's attention. Human speech is full of pauses, repetitions and repairs [Fox95]. There is evidence that disfluent speech can improve the listener's comprehension [Fox01]. Furthermore, some researchers postulate that humans use disfluent speech to express themselves and that it has a communicative function [CF02]. In addition, disfluent speech often occurs in combination with missing mutual gaze [Goo81]. Nonetheless, there is a lack of research on the topic of disfluent speech as a possible way to (re-)act in *HAI* systems. The goal of my thesis is to bring interdisciplinary research results together to improve the future design of *HAI* dialogue coordination in smart homes. I claim that the concepts of attention and hesitation are closely related and need further investigation in tandem.

RQ 2: Implementation

Which requirements does such a model pose to the design of dialogue (management) systems?

It needs to be further investigated if it is possible to use hesitations as a tool for the coordination of dialogue. From the system engineering perspective, it is important to know the requirements of such a model. Even though it is not the goal of this thesis to evaluate the correct function of single parts of the *AHM*, it must be ensured that it is possible to implement the whole system in an autonomous *HAI* with the current technical state of the art.

RQ 3: Evaluation

How does such a model affect the human-agent interaction in a smart home?

The evaluation of my model is an important part of my thesis. I discuss several ways to evaluate a dialogue system. However, the method of interaction study is the only way to keep the human "in the loop". Therefore, I evaluate my *AHM*, which uses system-hesitations as an *intervention strategy* for the coordination with the human interaction partner's attention to improve the *HAI* in the smart home. An agent that uses the *AHM* should be able to deal with an inattentive interaction partner. This should affect the interaction in terms of a better performance in the task at hand for this interaction. My hypothesis is therefore:

Hypothesis: *The Attention-Hesitation Model (AHM) increases the task performance in human-agent interaction.*

However, even if the goal is to improve the task performance, other side effects on the interaction should not be neglected.

To investigate my research hypothesis three aspects are especially important (see fig. 1.2):

1. a **model** that is based on *HHI* research,
2. its **implementation** in a *HAI* scenario in which the agent performs as autonomous as possible, and
3. its **evaluation** in a real *HAI* interaction study.

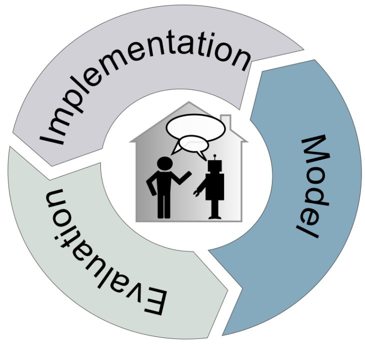


Figure 1.2: Three aspects of the research procedure to consider for answering the research question.

In the following section, I go further into detail on how these research questions are investigated throughout this thesis.

1.3 THESIS OUTLINE AND CONTRIBUTION

This thesis is composed of three parts. In the first part "*From HHI to HAI: Developing a Model*", I elaborate on the motivation for my work and investigate RQ 1 based on literature from *HHI* and *HAI* research. To this end, I focus on research from *HHI* in chapter 2. In section 2.1.1, I reveal considerations for the incorporation of human attention in *HAI* based on the *capacity theory of attention*. Furthermore, a concept of attentional state is developed, similar terms are differentiated and an overview of the role of gaze is presented in *HHI*. The evidence from the literature shows that the human gaze is a reliable indicator for interlocutor's attention and higher cognitive processes in *HHI*. Afterwards, I discuss disfluent speech and its use in interaction in section 2.2. Findings from the linguistic and psychological research on *HHI* are presented to investigate the claim that hesitations in speech can improve the listener's comprehension in *HHI* and are often produced as a reaction to the listener's inattentiveness. In the third chapter Summary of Research on Attention and Hesitations in HHI, I present findings from research focusing on attention and hesitations in *HAI*, which shows that these concepts are also important in *HAI*. An overview of the incorporation of the visual attention in *HAI* is given, which shows that the human gaze also plays an important role in *HAI*. In addition, existing systems and studies which deal with hesitations in *HAI* are presented, with special focus on hesitations as an intervention strategy. At the end of this first part, I present my model to incorporate the attention of the human interaction partner into the dialogue system which uses hesitations as an intervention strategy: the Attention-Hesitation Model (AHM).

In the second part of this thesis "*Fundamentals for Autonomous HAI*", I investigate RQ 2. After a brief introduction into dialogue modeling—particularly regarding the dialogue management component—in chapter 6, the requirements posed by its technical realisation are illustrated (section 6.2). Furthermore, the choice of the research platform is constituted(section 7.1). In chapter 7, the technical realization of a dialogue system which allows further investigation of my research question is described. Furthermore, scenarios and research studies, which use this dialogue system, including a first integration of human gaze, are presented.

In the third part "*Learning from Experiments*", I evaluate and enhance the model in five *ECs*, consisting of three pilot- and two HAI studies in a smart-home environment to investigate RQ 3. During these cycles, I investigate my hypothesis that the *AHM* increases the task performance of *HAI* in a smart home scenario. In doing so, I do not only look at the task performance itself, but also consider the side effects that the *AHM* can have on the interaction. Therefore, my model and the evaluation approach are iteratively improved and different hesitation features are explored to find an intervention strategy that can improve the task performance without having negative side effects on the interaction. In the end, my findings are summarized and conclusions for modeling the coordination of human-agent dialogue are drawn. Furthermore, the limitations of this work and the consequences for smart home interactions are discussed. I conclude with future research questions that follow from or can be investigated based on my findings.

In this thesis, I make several contributions to both cognitive interaction research and system engineering research of dialogue modeling. I present the first model, which uses hesitations within a speech act as a conversational signal for inattentive interlocutors that improves the interaction in a smart home environment. In contrast to various other systems, it uses hesitations not to "buy time" for the system, but rather to give the inattentive interlocutors the time they need and thereby acknowledge that the human attention is a valuable resource. In addition, it distinguishes between two different reasons for inattention: missing engagement or difficulties in understanding and deals with it with dedicated intervention strategies.

Furthermore, I make a technical contribution to incremental dialogue modeling through a combination of two frameworks and the resulting modular dialogue system perform as test bed for the evaluation of the *AHM*. In addition, I illustrate requirements posed by the technical realization of the model, which can serve as guidelines for further research on this topic.

Combining both sides, I make a contribution to the evaluation of such models and the investigation of the effect on the interaction. Through a cascade of evaluation cycles in a smart home environment, I improve not only the model itself, but also its evaluation process that shows that the task of the interaction is important for the effectiveness of the intervention strategy. Furthermore, I show that the *AHM* can have a positive effect on the interaction. During the *ECs*, I investigate different features for the recognition of inattention and the corresponding hesitation intervention strategy. For the attention concept, I show that already with the feature of mutual gaze the hesitation strategy can improve the task performance, but at the cost of less positive subjective ratings. With additional task related information, such as incorporating information about the task progress or the *Focus of Discourse (FoD)* in combination with the dialogue history, the model distinguishes between different reasons for inattention. I show that for the hesitation intervention strategy already unfilled pauses are useful to improve the performance in a practical task. However, participants struggle with the differentiation of pauses and turn-ends. The use of lengthening counteracts this problem. My final *AHM* uses mutual gaze and task related features to distinguish inattentiveness based on missing engagement or difficulties in understanding. Furthermore, it uses different strategies to deal with these different reasons for inattention. To deal with inattention based on missing engagement, it uses a cascade of lengthening, followed by an unfilled pause, followed by a hesitation vowel. For understanding problems, it uses repetitions with lengthening. I show that this final *AHM* improves the task performance without negative side effects on the interaction.

2 HUMAN-HUMAN INTERACTION

In the first part of this thesis, I investigate how the coordination of human attention and system speech can be modeled. As repeatedly shown in the literature on HAI, concepts from HHI can be successfully transferred to interactions with artificial agents. To this end, I focus this chapter on research on *attention* and *hesitations* in HHI and conclude with the interplay of these concepts.

2.1 ATTENTION IN HHI

It is important to be clear about the definition of *attention*. To this end, a short overview of different attention concepts and resulting theories are presented in the next section (section 2.1.1). Based on the *capacity theory of attention*, I reveal considerations for the incorporation of human attention in HAI. Afterwards, in section 2.1.2, a closer look at the attention and similar terms–which are often defined only vaguely—are taken and classified into observable features or concepts for mental states. In section 2.1.3, an overview of the role of gaze in HHI is given. Evidence from the literature shows that the human gaze is a reliable indicator for interlocutor's attention and higher cognitive processes in HHI.

2.1.1 *Theory of Attention*

I pay attention to you. What exactly is the meaning of this sentence? Does it mean, that the speaker looks at the interlocutor, listens to them, or is it a "condition of readiness for such attention involving especially a selective narrowing or focusing of consciousness and receptivity" as described in the dictionary [Mer19a]? In a famous quote, the psychologist and philosopher James says

> "Everyone knows what attention is. It is taking possession of the mind, in clear and vivid form, of one out of what seems several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence. It implies a withdrawal from some things in order to deal effectively with others."[Jam90]

He aims that everyone has a clear concept of what attention is—but is that true? Researchers from various fields, ranging from philosophy and psychology through to cognitive neuroscience and computer science, try to define the concept of *attention*. According to Styles, the term *attention* has no single definition but is rather representative for various concepts and phenomena, beginning from spending visual attention towards stimuli, up to complex processing systems which are intimately related to memory [Sty06].

Attentional Concept	**Theory**	**Representative**
Selective Attention: Focus on a specific stimulus by ignoring distractors.	Early selection filter theory	Broadbent [Bro58]
	Late selection filter theory	Deutsch and Deutsch [DD63]
	Attenuation theory	Treisman [Tre64]
Divided Attention: Pay attention to two tasks simultaneously.	Capacity theory of attention	[Kah73]

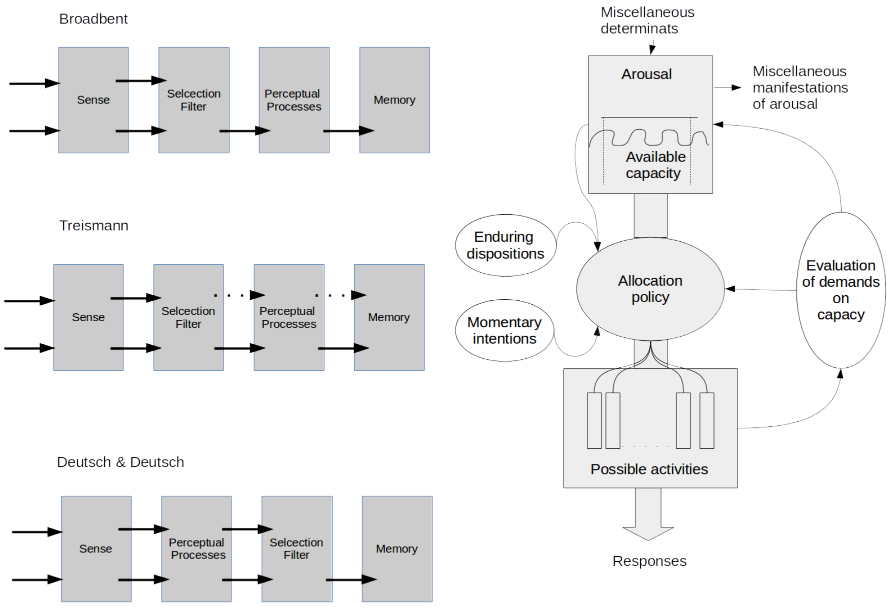
Table 2.1: Attentional concepts, corresponding theories, and representatives.

In the 1950-60's, a debate emerged on different models of the functionality of attention. Table 2.1 presents a short overview of various attentional concepts, corresponding theories, and their representatives. The concept of *selective attention* describes the phenomenon to focus on a specific stimulus by ignoring distractors. Various theories about the functionality of attention try to explain these phenomena. The first group is the *filter theory of attention*. Broadbent developed his filter model of *selective attention* [Bro58] based on the findings of Cherry, who found out that people can separate two different signals heard at the same time [Che53]. When confronted with a different speech for each ear, we can listen to only one of them. We are able to focus on a specific stimulus by each ear and can ignore distractors [Che53]. This phenomenon is also known as "cocktail party effect"[1]. Based on this so-called dichotic listening task, Broadbent describes this selective attention as a bottleneck in information processing capacity in the connection of two separate perceptual inputs. This limited capacity for the processing of the inputs needs a filter to protect against overload. In Broadbent's model, this filtering is based on physical characteristics and happens early in the processing: Before semantic processing is carried out, only one channel of information proceeds through the bottleneck.

A few years later, and in contrast to Broadbent's *early selection* filter model, Deutsch and Deutsch proposed the *late selection* theory [DD63]. They agree that there is some kind of filter, but argue that the attentional filter is located later in the processing pipeline. Based on evidence from different studies, they suggest that some information

1 Cherry introduced it as "cocktail party problem" [Che53]

reaches the short-term memory, e.g., the reference of the person's own name [DD63].



(a) Filter theories.

(b) Capacity theory.

Figure 2.1: On the left: attention filter theories (after [Bro58; DD63; Tre64]). On the right: Capacity model of attention (after [Kah73]).

Treisman developed an *attenuation theory*, where the filter determines how much information from each channel is being processed. This could be seen as a revision of Broadbent's *early selection* filter model. Figure 2.1 shows an overview of the three different filter theories. The main difference between them is the location and the principle function of the filter. They differ in their assumption of how far unattended information is processed. Whereas Broadbent locates the filter directly after the sensory input before the perceptual processing, in the model of Deutsch and Deutsch the filter is located after the perceptual processing. Treisman's model chooses another approach, the filter is located before the perceptual processing and determines how much information from each channel is being processed.

Sometimes we can attend to more than one input at the same time. This phenomenon is described by the concept of *divided attention*. Beside the previously mentioned filter theories, Kahneman argues that attention is not a bottleneck in information processing, but rather based on limited resources [Kah73]. Kahneman's so-called *capacity theory of attention* is an alternative model to these filter theories. Whereas the attention filter models in fig. 2.1 are a sequence of information processing, Kahneman suggested that we only have a limited amount of attention, which is allocated to tasks by a central processor. Beside the limited amount of attention, he assumed that this limit depends on the current arousal state, a higher level of arousal leads to a greater attention capacity. Furthermore, we have a set of possible activities, which require an allocation of attention. Figure 2.1b illustrates the main principle of this mental model. The allocation policy decides how much attention is paid to each activity or task, meaning how much cognitive effort is exerted. Several factors have an impact on this decision. On the one hand, the expected demand on capacity is evaluated and on the other hand, the arousal state has a direct influence on the distribution of mental capacity. In addition, the model incorporates bottom-up and top-down attention management. Enduring dispositions can grab the attention (bottom-up), e.g., novel stimuli or hearing their own name. Also, momentary intentions influence the policy, meaning that we can consciously decide to allocate attention to a certain task (top-down). Of course, several other theories of attention are developed, such as the *Theory of Visual Attention (TVA)* by Bundesen [Bun90]. The *TVA* explains attention through two successive processes, filtering and categorizing. In doing so, it incorporates both top-down and button-up precesses.

These different models of attention reveal some **considerations for the development of human-agent interactions**:

1. The human attention capacity is limited. Consequently, the agent needs to be careful with the allocation of the human's attention. It is a resource and should be treated as such. If the agent loses the attention of the human interaction partner, the agent should not try to regain it by all available means.

2. Changes in the environment can influence the attention policy of the human. The agent needs to be aware of it and consider environmental changes, to understand human responses.

3. Finally, the process of attention allocation can be top-down or bottom-up. For both cases, different inputs are responsible. To attract the human's attention, there are various approaches. The agent needs to decide which one is suitable in the current situation. In some situations, a short attention grabber may be enough, e.g., in the form of a novel stimulus. In other situations, a change of the intention should be preferred, e.g., by increasing the engagement into the interaction or the interest in a special object.

2.1.2 *Concepts and Definitions*

The overview of attention theories in the last section reveals that there is no general or joint concept of attention. Just as diverse are the terms and definitions used associated with attention. This section gives a broad overview about different definitions of attention and correlated concepts and classifies these into observable features or concepts for mental states.

Visual Attention

In the literature, there are different models of how humans control their *VFoA*, mainly the button-up and the top-down model. The button-up approach is based on saliency stimuli and is well-developed as a computational model. It is based on the concept of a saliency map, which combines information from various feature dimensions (e.g., color or motion) "into one global measure of conspicuity" [KU87]. However, attention shifts can also be produced top-down, based on the knowledge about the current task [DD95]. To build a computational attention model, both approaches should be considered [IK01]. Therefore, the coordination of a dialogue must keep in mind that the *VFoA* reflects both conscious and unconscious attention shifts. The current *FoD* can have a big impact on the visual attention of the interaction partners. Talking about specific objects or embodied entities leads to specific gazing behavior. I distinguish between *mutual* and *directed* gaze. Mutual gaze occurs when two people look into each other's eyes, whereas directed gaze occurs whenever the interlocutors look at the same object.

Joint Attention

The role of joint attention was well discussed in the last years, particularly it's importance for human development (e.g., [TF86; Bal95; MDD97]). Infants follow the eyes of their parents and share attention to the same objects. In particular, the temporal patterns of eye-gaze coordination between interacting humans, play a critical role in establishment of mutual rapport and understanding, which is generally referred to as "joint attention". These patterns include eye fixations as well as following gaze shifts to perceivable objects in the environment.

Cognitive Attention and Correlated Concepts

The visual attention is often used in human-human communication as an indicator for more high-level concepts, like the *cognitive attention*. Sharma et al. define a gaze-based measurement of student's attention during lectures as a concept of "*with-me-ness*" and "tackle this question from a teacher's perspective: 'How much the student is with me?'" [SJD14]. They measured and evaluated students' gaze patterns with an eye-tracker during online courses and define two different kinds of with-me-ness, perceptual and conceptual. Perceptual with-me-ness is measured by the fixation duration and revisits of objects, referred by the teacher through a pointer. In contrast, the conceptual with-me-ness measures how often the students look at the object verbally referred by the teacher. In a user study (n=40) Sharma et al.

found a relationship between the gaze patterns and post-test scores (performance tests) [SJD14]. The students which perform better in the post-test have more perceptual and cognitive with-me-ness. More precisely, they looked more at the objects referred by the teacher—both verbally or via pointing.

A similar concept often used in the *HRI* community is *engagement*. Sidner et al. define engagement as

> "[...] the process by which two (or more) participants establish, maintain and end their perceived connection. This process includes: initial contact, negotiating a collaboration, checking that other is still taking part in interaction, evaluating whether to stay involved, and deciding when to end the connection." [Sid+04]

They characterize engagement as a process and define engagement rules for their robot to control this process. In contrast, Peters et al. define engagement as

> "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction." [Pet+05]

They comprehend the concept not as a process but rather a value which is measurable at each point of time in the interaction. Glas and Pelachaud's literature review on the term engagement in *HAI* shows that several definitions exists in the community, often depending on the focus of the studied interactions [GP15]. They all have one thing in common: the interaction between the human and the agent requires a "minimum level of connection and cooperation". Other characteristics may differ between the definitions, e.g., regarding the used perspective. The variation in the definitions leads to different measurements of engagement.

Sidner et al. model engagement as a process with three parts (i) initiating the interaction (ii) maintaining the interaction and (iii) disengaging [SLL03] They formulate maintaining engagement as doing what the speaking communication partner does. More precisely, the rules for the looking behavior are:

- look wherever the interaction partner looks
- look at him/her if the interaction partner looks at you
- look at whatever objects are relevant to the discussion when the interaction partner does

When the roles are switched, they expect the same looking behavior from the other interaction partner while they are speaking. Sidner et al. argue these rules based on a *HHI* case study [SLL03]. They annotate head movements in a hosting situation, where a visitor tracked the head motion of the host. They classified the (female) visitor's failures to track changes in the (male) host's looking behavior into three classes: *quick looks*, *nods*, and *uncategorized*. The *quick looks* are failures, where the visitor fails to track a look that lasts for less than a second. In the *nod* category, the visitor reacts with nodding instead of gaze following (towards the speaker or an object) and indicates that she still follows the conversation. The last category *uncategorized* describes cases in which the visitor fails to track for more than two seconds, because of other actions or goals. Based on this data, Sidner et al. state their *principles of conversational tracking*, which says that interaction partners usually track the other's face during the conversation. They only look away because of an action relevant to the collaboration or an unrelated action [Sid+05].

Peters et al. agree with Sidner et al.'s three levels of interaction (establish, maintain, close) but define engagement as a measurable value which should be assessed from the interaction partner [Pet+05]. They distinguish two different phases of interaction. The first, the *establish phase*, is before the interaction and the participants need to decide if the other potential interaction partner wants to start an interaction. In the second phase, the *maintain phase*, the speaker needs to monitor the other's level of engagement to see the effectiveness of the interaction. Peters et al. closely connect engagement with the *interest* of the interaction partner in the interaction and the possibility to learn. They postulate that if an interaction partner is interested in the interaction, they engage and pay attention. To detect the engagement, a perception of the attention of the interaction partner is needed [Pet+05].

Whereas the concept of engagement manly evolved in the *HRI* community, the psycholinguist Clark coined the term *common ground*. According to him, the common ground is "the set of knowledge, beliefs and suppositions that the participants believe they share" [Cla96]. To have a conversation is not just to speak and hear utterances. Each participant needs to coordinate the conversation. The speaker must be aware if the listener is attentive, hears, and tries to understand the speaker. On the other hand, the listener has to display this information to support the communication. The interaction partners need to coordinate their communication. Thus, they need to keep track of their common ground, which requires them to determine what has been said and what has been understood. Clark, Brennan, et al. describe this process as *grounding* [CB+91].

Sidner et al. argue that grounding is a part of engagement [SLL03]. More precisely, that successful grounding is evidence for continuous interaction. Therefore, failures offer evidence that one of the interaction partners may wish to disengage. Although, I agree that grounding and engagement are closely coupled, I doubt that grounding errors always offer evidence for disengagement. In contrast, they can also reveal difficulties in understanding. Clark and Schaefer formulate the grounding criterion that the "speaker and addressees mutually believe that the addressees have understood what the speaker meant to a criterion sufficient for current purposes." [CS87]. Therefore, the speaker monitors the current state of understanding of the listener and, when necessary—more precisely when the understanding is not high enough for the current purpose—changes their utterances during the interaction [CK04]. Understanding thereby has different degrees. For a deeper insight into the different degrees between non-understanding and full (or strong) understanding, see [Bus18, pp. 13–19]. This concept of a spectrum of understanding is futher developed in a conceptual framework for "explanation as a social practice in which explainer and explainee co-construct understanding on the microlevel" [Roh+20].

It can be concluded, that in the *HAI* community several high-level concepts exist, but not always clearly defined. Often the *VFoA* is used to measure these concepts.

Restructuring Definitions and Terms

There are various definitions of attention and correlated concepts, which have similar characteristics and describe similar phenomena, or are subsets of each other. As a communication partner or from outside the interaction, it is only possible to observe the interaction partner and to draw conclusions to its mental states. Thus, I split these terms into two subgroups—*observable* and *concept*. Figure 2.2 sketches this categorization. At the bottom, observable features are presented,

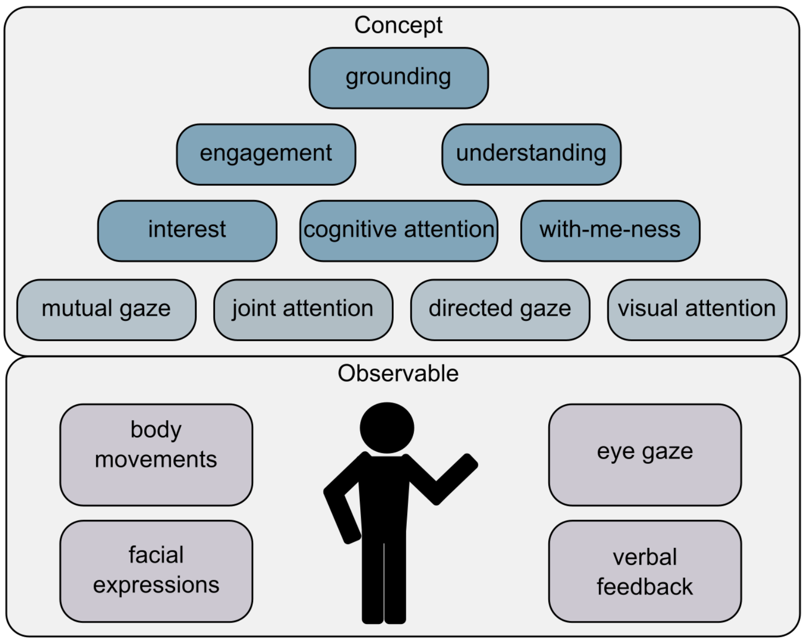


Figure 2.2: Categorization of observable features and mental concepts involving *attention*.

starting from visual cues, e.g., body movements. The orientation of the body itself or the head position is often used to measure attention. Another important observable feature is the eye gaze. In addition to the eye gaze, other facial features can give feedback to the other interaction partner. This not only includes emotional expressions, such as smiling, but also back-channels like head nodding or shaking. Beside visual cues, auditive cues can also be used, e.g., verbal back-channels All these observable cues allow conclusions about different mental states. The bottom row of the concepts consists of concepts which can be estimated by the observable features, such as the eye gaze. This includes the current *VFoA* of the interaction partner, gaze patterns like mutual or directed gaze, and resulting joint attention. These features allow conclusions to concepts above—the inner states—which are more complex, e.g., to the cognitive attention of the interaction partner, in the following referred as high-level concepts. The term *attention* plays a role in all of these concepts. I define attention in this thesis as follows:

Working definition: Attention

Attention is a cognitive concept. Being attentive in an interaction means to focus their own senses on the interaction partner and the current focus of discourse. As an inner state of a person, it cannot be directly observed. However, by observing a person's behavior, e.g., their VFoA, conclusions about this inner state can be drawn.

Furthermore, being attentive is a precondition for engagement and understanding the information provided by the interaction partner.

Working definition: Engagement

Engagement is a cognitive concept. Beeing engaged in an interaction means that the interaction partner is interested in the interaction with the other participant(s) and wants to continue it. They pay attention, meaning they focus their own senses on the interaction partner and the current focus of discourse. As an inner state of a person, it cannot be directly observed. However, by observing a person's behavior, e.g., their VFoA, conclusions about this inner state can be drawn.

Understanding, in turn, is a prerequisite for successful grounding. Clark and Schaefer formulate the grounding criterion that the "speaker and addressees mutually believe that the addressees have understood what the speaker meant to a criterion sufficient for current purposes." [CS87]. In line with this, I define understanding as a prerequisite for successful grounding.

Working definition: Understanding

Understanding is a cognitive concept. Understanding in interaction means the successful processing of the relevant information provided by the interaction partner to get common knowledge. There are different degrees between non-understanding and full understanding. As all-encompassing understanding of the interaction partner's intentions, ideas and thoughts is hardly possible, understanding is related to the expected action or goal. Successful understanding is reached, if the interaction partner understood what the speaker meant to a criterion sufficient for current purposes, e.g., performing the current action. Furthermore, as an inner state of a person, it cannot be observed directly, but the speaker can observe the behavior of the listener to infer whether the listener's understanding is sufficient to achieve the action or goal of the interaction.

Understanding is a precondition for successful grounding, which I define in this thesis as follows:

Working definition: Grounding

Grounding is a cognitive concept. It is a backward function to indicate that what has just been said has been understood. This feedback can be binary, by signaling understanding or non-understanding, or more nuanced, by communicating what part of the information was or was not understood. Information in interaction is grounded, if both interaction partners believe they have a common understanding of the information. As an inner state, it cannot be directly observed. However, by observing a person's behavior, e.g., their VFoA, conclusions about this inner state can be drawn.

Context information, like the current situation, *VFoA* of the interaction partner, or the *FoD* can be used to further enhance the recognition of a persons attention or understanding and respectively the common ground. Furthermore, an inattentive interaction partner is a signal for a disturbed interaction. This is elaborated in more detail in chapter 4. In the following, the role of eye-gaze and the corresponding visual attention for the *HHI* is presented.

2.1.3 *The Role of Eye Gaze*

In human-human interaction, the visual attention and eye gaze plays an important role during the conversation: at the beginning of the interaction, the turn-talking, and as a referencing signal. We use it as a synchronization mechanism, as an attention signal, and to communicate about the environment.

There is various research about different eye gaze patterns in interaction. Vertegaal et al. evaluated eye-gaze patterns in multiparty conversations [Ver+01]. They recorded 7 four-person discussion-groups and analyzed their looking behavior while speaking and listening. They suggest that the "user's eye gaze can form a reliable source of input for conversational systems that need to establish whether the user is speaking or listening to them". In particular, they found that the listeners gaze more to the speaker than to others. This is in line with the research of Argyle et al., who investigated the important role of eye gaze and especially *mutual gaze* in human-human conversations between two people in the seventies [ACC94]. Although mutual gaze implies eye contact—looking into the other's eyes—in fact humans not always fixate the other's eyes, but rather different points in the face of their interaction partner [Coo77]. Several researches investigated the distribution of gaze-direction in human-human conversation and found that, on average, people look at each other about 60%-80% of the time [Ken67; Ver+01; ACC94] (see also the review of Admoni and Scassellati [AS17]). This distribution has a very high variance and is influenced by many factors, such as differences in personality, the relation to the other interaction partner, or the topic of the conversation. Moreover, the conversational role has an important impact on the gaze behavior. Speakers usually look at the beginning and the end of utterances towards the listener. During longer utterances, they perform short glances. In general, listeners look more at the speaker, than the other way around. Cook reported that people "look more while listening than while talking, it being usual to look at someone more or less continuously while he is talking" [Coo77]. They could demonstrate that the human gaze is a reliable source for humans attention.

Kendon examined the gaze as a turn-taking signal. With his work, he could show that speakers mostly look at the listener at the end of their utterances to hand over the speaking turn [Ken67]. Similar results are reported by Goodwin. He investigated the gazing behavior, especially at turn-beginning. One finding was that the speaker gazes towards the hearer, particularly at the beginning of a turn to achieve mutual orientation [Goo81]. To receive the attention of the listener, the speakers undertake different techniques, which are explained in section 2.2.3 in more detail.

Furthermore, the speaker perform gaze aversions. These signal cognitive effort, at the beginning of a response to a difficult question [Bea81; DP05]. Andrist et al. analyzed gaze aversions in dyadic human-human conversations and grouped these into three different categories [And+14]. Gaze aversions of the speaker during an utterance are labeled as *turn-taking*, aversions near cognitive events (e.g., thinking about a response) are labeled as *cognitive*, and the remaining gaze aversions are labeled as *intimacy*. They analyzed the length of the gaze aversion for these categories and found out that on average the gaze aversions for *intimacy* regulation (while speaking M=1.96s) are shorter than the others (*cognitive* M=3.54s; *turn-taking* M=2.30s). Gaze aversion for *intimacy* regulation while listening are even shorter (M=1.14s) [And+14]. Based on these results, it can be assumed that gaze aversions with different functions vary in their length. Whereas gaze aversions for intimacy regulation tend to be shorter, gaze aversions for turn-taking signal (or rather turn-holding signal) are much longer.

Heylen explains the different functions of gaze based on Clark's multi-level organization of joint action in communicative acts. These consisting of (1) attend to sound/gestures (2) identify the signal (3) understand the meaning (4) consider answering [Cla96]. The first level is perception. The listener pays attention to the speaker and vice versa to be able to observe their signals. In this stage, gaze is a symptom resulting from the monitoring of the interaction partner. However, gaze has not only the function of perceiving, it can also be a communication signal. This fits into the second level of presenting/identifying signals. Heylen argues that the fact that people are aware of the fact that the interaction partner may observe their looking behavior can also lead them to use gaze as a signal for showing attention [Hey05]. In this second stage, the listener is not only looking at the speaker to perceive what the speaker is doing, but rather with the intention to signal attention to the speaker. This idea is not new. Bavelas et al. describe gaze behavior in conversation as a coordination mechanism for collaboration [BCJ02]. In fact, they argue that the speaker gaze towards the listener requests feedback signals from the listener and frequently leads to back-channel signals from the listener [BCJ02]. Furthermore, gaze is not only a signal to present the own attentiveness. Human also use gaze as a reference signal to communicate about their environment. Several researchers investigated the correlation between eye-gaze and objects referred to in an utterance.

In their comprehensive literature review, Admoni and Scassellati summarized findings of several studies which investigate gaze behavior in combination with object reference [AS17]. One of these findings is that speakers look at objects before they refer to them (e.g. [GB00]). This gazing behavior of the speaker can be used to better understand the meaning of the spoken utterance [HB07] or even predict what the speaker will talk about next [Bou+12]. Similar findings are presented from the field of object manipulation and handover scenarios. Land and Hayhoe evaluated the relation of eye gaze and hand movements in a food preparing scenario. They could show, that the eye gaze often shifts to the next object before the hands follow [LH01]. Strabala et al. analyzed multi-modal cues in human-human handover scenarios. Surprisingly, they discovered that mutual gaze is not one of the key features to predict the intent to handover. However, they found that gaze in general plays an important role, but rather asynchronically [Str+12].

2.2 HESITATIONS AND THEIR ROLE IN HHI

A phenomenon often observed in *HHI* is disfluent speech. This section reviews evidence for my claim that *hesitations* in speech can improve the listener's comprehension in *HHI* and are often produced as a reaction to the listener's inattentiveness. To this end, the occurrence of disfluent speech is discussed and the working definition of hesitations is developed in section 2.2.1. Afterwards, the debate about the reason why speakers are disfluent is presented in section 2.2.2. In section 2.2.3, research on the effect of hesitations on the listener is presented which support my claim.

2.2.1 *Disfluent Speech and Hesitation Definition*

The human speech is not fluent, it is full of pauses, corrections, and hesitations vowels, such as "uh" and "uhm"[2]. The following Example 2.2.1 shows a cascade of different disfluencies.

Example 2.2.1: A cascade of different types of disfluencies.

Give me the: (- - -) uhm (- - -) the green bottle.

At first, the word "the" is lengthened, followed by an unfilled pause. Afterwards, a filled pause is produced (the hesitation vowel "uhm") again followed by a second pause. The speaker then continues with a repetition of the last previously produced word.

Fox Tree estimates that about 6% of words in spontaneous speech are disfluent, excluding silent pauses [Fox95]. Bortfeld et al. analyzed different factors for disfluencies such as gender, age, and the difficulty of the topic in a corpus analysis of 40 hours of spontaneous speech [Bor+01]. They could support Tree's estimation and found 5.97 disfluencies every 100 words on average, whereas older speakers have a slightly higher disfluency rate (6.65) than younger speakers (5.55). In this corpus, men have a higher disfluency rate (6.80) than women (5.12) and the analysis showed that people in general produces more fillers in difficult tasks [Bor+01].

Disfluencies in spontaneous speech are investigated in linguistics to get insights into the human speech production. In the following, I list some possible examples of disfluencies, based on the taxonomy of Eklund [Ekl04]

2 for a detailed overview of different types of disfluencies see [Shr94; Ekl04]

- **Unfilled pauses**: are periods of silence produced by the speaker, which can occur inside a word, phrase or between grammatical complete sections.
- **Filled pause**: are vocalized hesitations, such as "uh" and "uhm".
- **Prolongations**: are created by cutting-off words or lengthening a phoneme.
- **Repairs**: can be repeated, inserted, deleted, or substituted items.
- **Editing terms**: are words like "ups" or "no wait" wait", which can be used before a self-correction.

However, no common taxonomy exists. As Eklund noted, the phenomena of disfluencies are referred to a variety of different terms [Ekl04]. Hartsuiker and Notebaert presented a slightly different categorization. They divided disfluencies into tree broad categories: *self-corrections*, *repetitions*, and *pauses*. The category *self-corrections* consists of substitutions, additions and deletions, whereby the *pauses* could be silent, filled - like "uh" and "uhm" (so-called *fillers*), or prolongations [HN09].

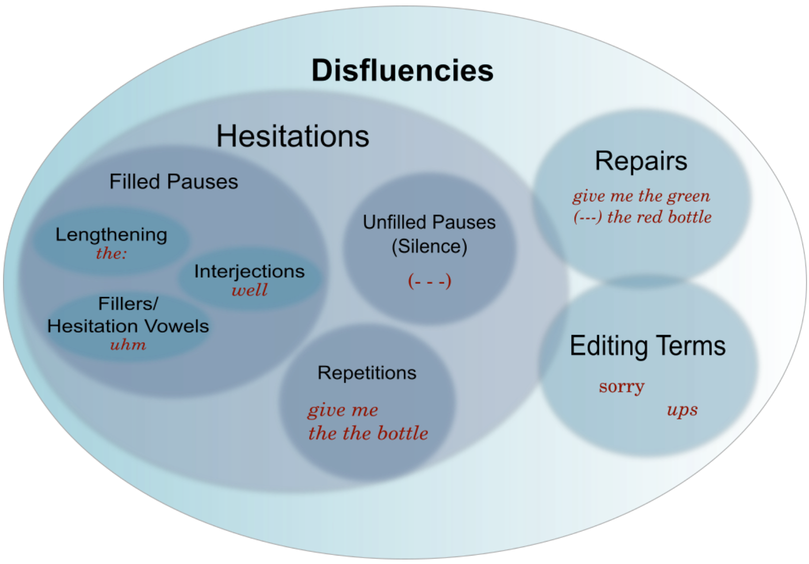


Figure 2.3: Definition of hesitations used in this thesis.

What Hartsuiker and Notebaert describe as pauses, are usually called **hesitations** in the literature (see e.g., [CS08]). The concept behind this, is that the speaker's hesitate not to correct itself but to "bye time" for itself. Reasons for this hesitating behavior are diverse and are more discussed in section 2.2.2. In this thesis, I focus on *hesitations* as a subgroup of disfluencies as depicted in fig. 2.3.

Working definition: Hesitation

Hesitations are a subgroup of disfluencies which are produced to delay the current speech content, consisting of silence, lengthening, fillers, other injections, or repetitions without self-corrections.

Thereby, I distinguish between *unfilled pauses*, *filled pauses*, and *repetitions*. *Unfilled pauses* are simply silence, whereas *filled pauses* could be the lengthening of a phoneme or a hesitation vowel also called filler, such as "uhm". Interjections are words such as "well" which I also count as filled pauses, but which are not further considered in this thesis. Furthermore, repetitions of previously produced words without any self-correction are defined as hesitations in this work. Other disfluencies, such as repairs, which of course could also include phenomena like pauses and repetition are not defined as hesitations. Also, editing terms are to be distinguished from it. Of course, this definition is incomplete, but should narrow down the term of *hesitations*.

Even though hesitations occur frequently, their role in communication is not well-defined and has been discussed intensively in the last decades.

2.2.2 *Hesitations: Symptom or Signal?*

According to common opinion, speech disfluencies are often associated with speech errors and these phenomena should be avoided to have a fluent conversation. In theoretical models of language generation and understanding, they are normally not considered (e.g [Cho76; Fer00]) and are not even part of the general language. However, in the last years, a debate has emerged about why speakers are disfluent[3]. Most researchers describe hesitations as a symptom or side effects of difficulties of the planning process itself. According to Levelt's speech production model, the reason for the delay is that the speaker has problems in speech planning and is unable to proceed the current speech production [Lev89]. Therefore, hesitations often occur when the speech planning process is difficult, e.g., in difficult or unfamiliar topics [HN09; Bor+01]. However, it is unclear, why this disfluencies occur.

Some researchers postulate that hesitations are used as a communicative act to inform the listener that the speaker has difficulties. Clark and Fox Tree claim that some hesitations have a communicative function, more precisely, that speakers use "uh" and "uhm" to announce a delay in speaking [CF02]. They go even further and postulate that speakers are aware of the expected delay and use "uh" for minor and "uhm" for a major delay. Therefore, these hesitations are signals which are produced top-down [CF02].

3 See [CS08; FC12] for an overview of this debate.

While several researchers describe hesitations as symptoms or signals for intrinsic issues, a few researches also point out the possibility to use it as a communication mechanism to express issues with the interaction partner. In the following section, I will present research investigating the effect of hesitations on the listener.

2.2.3 *Effects on the Listener*

Few researchers investigated the effect of hesitations on the listener. In particular the effect of hesitation vowels is investigated, see the work of Corley and Stewart for a broad overview [CS08]. This research shows that they can be beneficial for the listeners' comprehension. Fox Tree, for example, studied the effect of hesitations, more precisely "uh" and "uhm" on the listeners' ability to recognize words in the upcoming speech in English and Dutch [Fox01]. She could show that the hesitation vowel "uh" increases the reaction time of listeners in the recognition of upcoming words. Other researchers found positive effects in comprehension or memory performance [FW11].

Fraundorf and Watson organized the existing studies into three different hypotheses [FW11]:

H1 *predictive processing hypothesis*: The listener can use disfluencies to predict what they will hear next.

H2 *attentional orienting hypothesis*: Disfluencies (re-) orient the listener's attention towards upcoming speech.

H3 *processing-time hypothesis*: The listener has more time to process the information.

Several researchers investigate when speaker produce disfluent speech. As already explained in 2.2.2, hesitations often occur when the speech planning is difficult, e.g., before topic changes or before difficult names. The predictive *processing hypothesis* assumes that the observation of this behavior leads the listener to expect exactly these difficulties in the speech planning process of the speaker.

The attentional *orienting hypothesis* expects that disfluencies (re-) orient the listener's attention towards the upcoming speech, regardless of whether the upcoming speech material is more difficult to the speaker or not. Evidence for this hypothesis comes from Collard, who studied immediate and lasting effects of hesitations on the listener's attention. Based on several experiments, he tested the effect of fillers on event-related potentials (ERPs) [Col09].

The basic idea of the last *processing-time hypothesis* is that the listener simply has more time to process the information. Several researchers found similar positive effects for fillers and silent pauses on the reaction time of listeners (e.g., [BS01]). Counterarguments for this hypothesis come from Barr and Seyfeddinipur, who compared the effects of fillers and noise (a cough or a sniffle). They found a faster reaction time of the listener for fillers than for the noise condition [BS10].

Fraundorf and Watson themselves argue for the attentional orienting hypothesis and underline the existing evidence with their own experiments [FW11]. In their first experiment, they assess the effect of fillers on the recall of a discourse. They compared linguistic interruptions (fillers) with non-linguistic interruptions and the baseline condition (fluent speech). They found that fillers have a beneficial effect on the recall of the entire discourse, not only on the manipulated plot points [FW11]. Interestingly, the non-linguistic fillers impaired the memory performance compared to the fluent condition. These result contradicts the third hypothesis, as both interruptions have the same length. In a second experiment, Fraundorf and Watson compared different filler locations more precisely the effect on recall between fillers which occur at typical locations (e.g., before a new plot point) and atypical locations (e.g., within a plot point). They could not show a significant effect on recall between the two conditions. However, both conditions have a beneficial effect compared to the baseline condition [FW11]. Based on these experiments, Fraundorf and Watson conclude:

> "These results are most consistent with an attentional orienting account in which fillers direct attention to the speech stream but do not always result in specific predictions about the nature of upcoming material. These results also generalize past experimental findings on fillers to the level of the discourse and to later recall, demonstrating that fillers can facilitate recall even of complex discourses." [FW11]

Thus, there is evidence that fillers are beneficial for the listener's comprehension. The reason for this is still unclear. Different hypotheses explain these effects. It may be because of extra processing time, the inherent nature of fillers to orient attention, or the humans' experience which leads us to predictions about the upcoming speech. Still questionable is, however, if this effect is transferable to other hesitations, e.g., silent pauses or lengthening.

As already mentioned in section 2.2.2, it is still unclear if disfluent speech is a symptom resulting from speech planning problems or used as a communicative act. Evidence for the later hypothesis comes from Goodwin. As discussed in section 2.1.3, he investigated speaker and listener behavior within a turn [Goo81]. His data set contains approximately 50 hours of natural conversations, e.g., family get-togethers or dinners with friends. Special attention is paid to the correlations between (mutual) gaze and disfluent speech. He could show that disfluent speech often occurs with missing mutual gaze [Goo81]. In the following, we will give some examples for such synchronization mechanisms. In example 2.2.2, the speaker Eileen is not convinced that she has the attention of the listener (Debbie). She starts speaking and restarts until she has Debbie's attention.

Example 2.2.2: Restarts until mutual gaze is achieved (adapted from [Goo81, p. 63]).

(First Restart)
↓

Eileen:	*I ask him, (0.1) I ask him if he- (0.4)*
Debbie:	....

(Second Restart)
↓

Eileen:	*could- If you could call me when you got in.*
Debbie:	.... *X* ————————————

↑
(Mutual Gaze)

Goodwin presented several examples of such restarts and interpreted this behavior as a speaker's request for a listener's attention (mutual gaze). Beside this procedure to obtain the gaze of the recipient, he found an alternative strategy of speakers: the unfilled pause. Example 2.2.3 shows such a situation. Ethyl starts speaking and after a few words she makes a pause until mutual gaze with Jim is achieved.

Example 2.2.3: Unfilled pause until mutual gaze is achieved (adapted from [Goo81, p. 66]).

Ethyl:	*I had a who::le:: (- - - - - - - - - - + - -) paul full of*
Jim:	.... *X* ——————

↑
(Mutual Gaze)

Other examples show that, in addition to the pause, the speaker often produces a filler (like in example 2.2.4). Goodwin argues that the pause in the middle of a turn is a noticeable disruption in the speech stream and may be used—like the restart—to signal that the listener's attention is requested.

Example 2.2.4: Hesitation with filler (adapted from [Goo81, p. 66]).

Dianne:	*He pu:t uhm, (- - - - - - -) Tch! Put crab meat on*
Marsha:	.... *X* ——————

↑
(Mutual Gaze)

While these results show how mutual gaze can be achieved at turn beginning, Goodwin also observed similar behavior if the listener's gaze is withdrawn in the middle of a turn.

Example 2.2.5: Hesitations after listener's attention shift in the middle of the turn (adapted from [Goo81, p. 86]).

Margie:	*And get out it a:ll the way up my ba:ck which was a*
Ross:	————————————————————
Margie:	*big uh (- - - - - - - -) help on that*
Ross:	—— , , , , , ... X ——————
	↑ (Mutual Gaze)

Example 2.2.5 demonstrates such a behavior. Margie starts speaking, meanwhile Ross is looking at her. In the middle of the turn, Ross' gaze moves away from Margie. At this point, she hesitates, more precisely she produces a filler and pauses afterwards. After mutual gaze is established again, Margie continues speaking.

2.3 SUMMARY OF RESEARCH ON ATTENTION AND HESITATIONS IN HHI

In the last two sections, I presented two main theoretical foundations of this thesis from *HHI*: attention and hesitations in speech. I first introduced different attention concepts and resulting theories and revealed considerations from it for the development of human-agent interactions in section 2.1.1. In addition, I discussed attention and similar terms in section 2.1.2 and classified them into observable features or (high-level) concepts for mental states. Furthermore, I outlined that the visual attention is a reliable indicator for these high-level concepts (see section 2.1.3). Based on these findings from human-human interaction studies, I found evidence for the statement, that the human gaze is a reliable indicator for their attention and higher cognitive processes in *HHI*. Especially the gaze of listeners gives insights to their mental states. With their gaze, people signal their attention. Except of brief gaze aversions for intimacy regulations (1̃sec), they look at the speaker or the current *FoD*.

In the second section, I presented the phenomena of speech hesitations in *HHI* (see section 2.2). Whether disfluent speech is a symptom or signal cannot be ultimately answered in this thesis. Nevertheless, they occur regularly in spontaneous *HHI* (see section 2.2.1). I presented evidence in section 2.2.2 for the hypothesis, that they may have a communicative function and that they influence the interaction partner. Furthermore, I have shown that especially hesitations could have a beneficial effect on the listeners' comprehension of utterances in section 2.2.3. Additionally, I presented evidence that hesitations orient the listener's attention towards the upcoming speech. Speakers in *HHI* use this effect to achieve mutual orientation. Based on these research, I found evidences for the statement that hesitations can improve the listener's comprehension in *HHI*.

Regardless of whether humans are using hesitations intentionally as a communicative act or not, the question rises if it is possible to use these speech phenomena in *HAI* as a communicative act to coordinate the dialogue between the human and agent.

3 HUMAN-AGENT INTERACTION

In the last chapter, evidence was presented on two claims. At first, in section 2.1 research shows that the human eye-gaze is a reliable source for their attention and higher cognitive processes in *HHI*. Furthermore, evidences are presented for my claim that *hesitations* in speech can improve the listener's comprehension in *HHI* and are often produced as a reaction to the listener's inattentiveness. However, it is questionable if these findings are transferable to *HAI*. To this end, the concept of attention and the phenomena of hesitations are further examined in this chapter for *HAI*. Different approaches for modeling the human attention (and to react on it) are discussed in section 3.1. In addition, the use of hesitations as intervention strategy in *HAI* is examined (section 3.2) and the finding for my research question is summarized in section 3.3.

3.1 ATTENTION IN HUMAN-AGENT DIALOGUE

This section set out to describe whether the statement that the human eye-gaze is a reliable source for their attention and higher cognitive processes in *HHI* is transferable to *HAI*. Therefore, the use of the visual attention—especially the eye-gaze—in *HAI* in general and several findings from interaction studies (section 3.1.1) are presented. In doing so, some interaction strategies to incorporate the human attention into a dialogue system are presented. Afterwards, different approaches for modeling the human attention to react on it are discussed in section 3.1.2.

3.1.1 *Visual Attention in Interaction*

Expressing the agent's attention through eye gaze behavior is a big research area in the *HRI* community, as well as the research area of virtual agents [AS17; Ruh+15]. Several findings from *HHI* (discussed in section 2.1.3) have already been replicated in *HAI*. Researchers investigated the effect of gaze, both artificial gaze generation and the detection of human's gaze. To this end, after a short summary of research investigating generating gaze behavior (section 3.1.1), research on investigating incorporating the human visual attention into dialogue systems is presented. This contains both reacting on gaze (section 3.1.1) and head movement (section 3.1.1).

Systems Generating Gaze Behavior

There are various models for realizing gaze behavior, from biologically inspired models for low-level eye animation to more abstract realizations with head movements. The review by Ruhland et al. provides an excellent overview of this research [Ruh+15].

The general realization of gaze strongly depends on the appearance if the agent itself. Admoni and Scassellati sorted robots and virtual agents based on their appearance and capabilities dimensions, which should reflect the range of behavioral realism [AS17]. While virtual agents can simulate human eye-gaze, social robots do not reach such realistic behaviors. Robots such as Nao or Keepon can not even move their eyes and therefore move their whole head. On the other side of the range, there is the android robot FACE, which is equipped with a "social gaze-control system" [Zar+14]. It should enable the robot to show attention to relevant target points in a social interaction. It uses social ques to manage the robot's attention and to coordinate its gaze behavior. Zaraki et al. presented an extensive literature research on attention and gaze modelling and summarized four areas of social cues for attention elicitation: nonverbal/verbal cues, proxemics, effective visual field of view, and habituation effect [Zar+14]. The authors integrated different areas into their attention module and evaluated it in a proof of concept evaluation. They recorded the gaze behavior of eleven participants watching a video. In this video, two people entered and left a room independently and had a discussion in between. During the discussion, they interacted with the camera from time to time, as if someone were there. In each scene, only one person spoke at a time, while the other person performed gestures to attract the attention of the imagined viewer. Afterwards, they tried to replicate the gaze behavior of the participants watching the video with their android robot FACE. They achieve a replication factor of 89.4% throughout the video [Zar+14]. This evaluation shows impressive results, but has also some weaknesses. First, the dataset is not recorded in a real interaction. The recorded interaction itself simulated the position of the viewer/robot with a camera. It is probable that the two persons would perform differently in front of another person instead of the camera, which can not give any kind of feedback. Secondly, the recording of human gaze behavior was not in a real interaction, but rather of people watching a video. It is understandable that the authors choose this methodology, because of its reproducibility. However, this gaze behavior does not necessarily inform about behavior in a real interaction, in which behavior is affected by all co-present interaction partners. Lastly, the generated behavior of the robot should be evaluated in a real human-robot interaction as well. The effect on the user, side effects in the interaction, and the general appearance of the robot are important aspects, which need to be examined too.

Admoni and Scassellati classified the FACE android robot as one of robots with the most realistic behavior capabilities [AS17] While I agree that the appearance is realistic, a proof that the behavior generation is realistic as well is not given. Even the hardware possibilities—the FACE robot has four *Degrees of Freedom (DoF)* for the head and two *DoF* for the eyes—do not match the possibilities humans have for gaze generation.

A robot which is not mentioned by Admoni and Scassellati is the anthropomorphic robot head Flobi, which can be used with a motion capture system for the head and eye control and already has three *DoF* for the eye movements [Lue+10; Sch+13]. Besides the appearance of the agent, several other aspects influence the agent's possibilities to use and interpret gaze in *HAI*. Various sensors can be used and different kinds of information processed for the attentive system.

Similar to Zaraki et al.'s so-called "social gaze-control system" [Zar+14], other attentive systems are capable of interactively directing the robot's attention towards the human and vice versa (cf. [Lan+03; BS99; Dau07]). Breazeal and Scassellati use the concept of saliency maps (see 2.1.2) for their attention system for the social robot Kismet [BS99]. The system creates saliency maps from color and motion and uses them as button-up information. This is further enriched with context information, in this case face detection results. Additionally, the attention system has a habituation model. Based on the output of the system, the social robot Kismet looks at specific points of interest.

The attention system by Lang et al. fuses multimodal information on a symbolic level [Lan+03]. It anchors the abstract representation of the face detection, leg detection, and sound source localization to a person model. The robot shifts its attention to the resulting person of interest, by turning the camera on top of the robot. Such kind of attention systems can shift the attention of the robot to specific points of interest, but not incorporate high-level information such as the current *FoD*.

Especially in collaboration tasks, the robot gaze could have a big impact on the *HAI*. Vollmer et al. investigated robot gaze feedback on human tutors' demonstrations [Vol+14]. They examined three different gaze behaviors for online feedback: social, random, or static gaze. As a platform, ASIMO was used. The authors found that in the social gaze condition, tutors demonstrated the actions slower than in the static gaze condition, which indicates that people use this behavior as a feedback signals [Vol+14]. The eye gaze to a target position indicates that the robot understood what the goal of an action is. Vollmer et al. propose to consider an interactional loop for robot learning [Vol+14].

In another cooperation task, Moon et al. investigated the effect of robot gaze on object handover between a PR2 humanoid robot and a human [Moo+14]. They investigated three different gaze patterns to indicate the beginning of the handover:

- *no gaze*: only tracking the gripper
- *shared attention*: tracking the gripper and shifting the gaze to the handover location
- *turn-taking*: tracking the gripper, shifting the gaze to the handover location, and looking at the human.

Moon et al. found out that gaze expressions in this *HRI* object handover can affect timing, i.e., the mean reach time for the *shared attention* condition is significantly earlier than in the *no gaze* condition [Moo+14]. The effect of gaze as synchronization mechanism during human-robot object handover is further investigated by [Mey20].

Robot's eye gaze behavior does not only impact cooperative tasks, but also the conversational role and can be used to manage turn-taking. Mutlu et al. showed that the gaze behavior of a robot has a significant effect on conversational roles in a multi-party question-answer scenario with one robot and two participants in a between-subject interaction study [Mut+09]. They found robot's gaze behaviors, which cue three different conversational roles:

- *addressees*: gaze at the participants all the time during greeting, conversation and turn-taking phases
- *bystanders*: gaze at the participants only at greeting and short glaces during the conversation
- *overhears*: no gaze at all

Mutlu et al. showed, that participants in the *addressees* condition took significantly more speaking turn and spoke significantly longer than *bystanders* and *overhearers*. Interestingly, the gaze behavior did not affect subjects' recall, but *addressees* and *bystanders* liked the robot significantly more than *overhearers* [Mut+09].

These results are confirmed by Skantze et al., who investigated turn-taking cues in a multi-party *HRI* scenario with two humans and the Furhat robot [SJB14]. They evaluated the effect of Furhat's gaze behavior on the users' turn-taking behavior and found out that the gaze of the robot strongly influences which user will speak next. Additionally, different multi-modal cues for claiming the floor are evaluated. Looking away while smiling or producing a filled pause ("eh") has the biggest effect as a turn-taking cue [SJB14].

Sidner et al. implemented gazing behavior to demonstrate face-tracking in their robot to control the engagement between their robot and the user and tested it in a human-robot interacting study [Sid+05]. They found that participants direct their attention to the robot more often in interactions where these engagement gestures are present.

Other research was conducted to evaluate the effect of robotic or virtual agent's gaze. Mutlu et al. evaluated the effect of different frequencies of the robot's gaze. They found that looking at a person affects their recall in a storytelling scenario [MFH06]. Other researchers investigate the effect of a robot's or virtual agent's (e.g., [Gar+03; BAT09; And+12b; Mur+07]), or try to replicate humans gaze behavior on a (virtual) agent (e.g., [Ver+01; Lee+07; LMD12; And+12a]).

Systems Reacting on the Gaze Behavior of the User:

There are various systems which directly react to the gaze behavior of the human. I present two which react on it during system speech. Eichner et al. developed an infotainment application where life-like characters present two MP3 players in a virtual showroom and—based on the user's gaze behavior—adapt the information presentation [Eic+07]. They defined different areas of interest in the virtual environment and measured the gazing behavior of the participants during the presentation performed by the agents. The system checks two different kinds of "grounding situations". A *short grounding situation* persists for less than 200ms (e.g., during deictic gestures) where the user is expected to look at a referred object "during the utterance or within one second after the utterance or gesture terminated for at least 150 ms". Situations longer than two seconds are defined as *long grounding situation* where "the user is supposed to look at the grounding object for 45% of the time of the duration of the utterance". The agents respond to failures in these grounding situations by interruption of the demonstration. Their intervention strategy consists of explicit statements and comments about the attention of the user or waving a hand to attract attention. Eichner et al. evaluated this behavior in an interaction study (N=35). In their control condition (*pseudo interactive*), the agents act at seven predefined points of the interaction with this explicit re-attention strategies, regardless of the gaze behavior of the user. They found that in the *interactive* condition the grounding behavior was more successful (77%in the *interactive* vs. 56.67% in the *pseudo interactive* condition) and the users reported a higher believe that the agents are aware of them [Eic+07]. This study shows the positive effect of re-attention strategies on the groun-ding. Nevertheless, the choice of the control condition is not ideal. It leads to incomprehensible behavior because the agent performs a re-attention strategy even though the user is attentive. It also raises the question whether a less explicit re-attention strategy could achieve the same results.

Palinko et al. evaluated the role of gaze as an implicit signal to turn-taking in a dictation scenario [Pal+15]. They used an iCub robot, who played the role of a teacher and dedicated English or Italian sentences to eight participants. The participants interacted with two versions of the system: *contingent* and *rhythmic*. In the *rhythmic* condition, the robot waited a fixed amount of time after each sentence, while in the *contingent* condition the robot continued dictating a new sentence whenever the participant gazed at it. Even though, they could not find any difference in the task performance of the participants (number of errors, writing speed) they found some benefits of this adaptive behavior. First, although the participants were unable to recognize the difference between the two conditions, 5 of 8 participants rated the *contingent* conditions as less difficult. Additionally, they found some benefits for the individuals. Especially very fast and very slow participants could benefit from the adaptive behavior of the robot. They could perform their task execution in their speed, which leads to differences in the wait time between the sentences. For easier tasks, most participants reduced the wait time, whereas for difficult tasks the opposite was the case [Pal+15].

Buschmeier developed conceptual and computational models of "attentive speaking" for a virtual agent [Bus18]. By incorporating gaze as well as other multi-modal feedback by the human interlocutor, his model estimates the human's understanding of the agent's utterances. Based on the estimated mental state, the dialogue system reacts with different utterances describing the estimated grounding state, such as "I'm not really sure whether you understood me. Should I repeat it or should we continue?" In a semi-autonomous *Wizard-of-Oz (WoZ)* study he investigated the effect of the attentive agent in a *HAI* (N=36) The agent was perceived by the participants as "interested in the feedback that they provided" and "helpful in resolving their difficulties in understanding". However, the agent could not outperform the baseline (without ensuring the understanding of the listener) in terms of dialogue efficiency. This research raises the question whether it is possible to react in other ways to not understanding listeners, which may be less costly and not directly addressing the grounding errors.

Besides reacting to gaze, several research uses head movements as attention signal.

Systems Reacting to Head Movements as an Attention Signal

Using head movements instead of gaze is a frequently used abstraction of the users' attention. Skantze and Gustafson evaluated attention as an interaction control mechanism in a multi-party human-human-computer dialogue setting [SG09]. In this scenario, a virtual agent assisted people in the organization of daily activities. The agent was able to answer questions about events by using a Google Calendar as backend. Skantze and Gustafson presented an attention and interaction model that allowed the users to switch attention between the system and another human. The system monitors the users' attention by tracking their head movements, and adapts the speech production, i.e., the system only starts speaking, if it has the attention of the user [SG09]. Furthermore, the system can interrupt itself, when the user states an utterance while the system is speaking. Even though they present a very interesting and detailed attention model (which will be discussed in more detail in section 3.1.2) they primarily evaluated the addressee recognition of the system. In fact, they compared their approach "look-to-talk" with a "push-to-talk" scenario in an interaction study with seven participants. Each participant interacted in both conditions (within-subjects-design). In the human-human-computer dialogue setting, the study assistant first explained both conditions of the system. After the introduction, the participant had to find a suitable slot for a dinner appointment with the other human interaction partner (the study assistant). The study assistant used a paper calendar, while the participant had to interact with the agent. Skantze and Gustafson evaluated the number of misdirected utterances of the user. They found that the participants always looked at the system when they talked to it [SG09]. By contrast, 5.1% of the utterances of the participants—while they look at the system—were addressed at the other human. In the push-to-talk condition, 24.8% off all utterances addressed to the study assistant were interpreted by the system because the participants forgot to deactivate it. After the interaction, the participants gave feedback about the two different systems, but there were no clear consensus which system perform best.

Yu et al. use head movements of the user as attention signal to coordinate the robot's speech production [YBH15]. They introduce their model in a direction giving system. At predefined points in the interaction, more precisely at the beginning of a phrase, the robot checks if the user is still attentive based on their current head position. As an intervention strategy, the robot make a restart, more precisely it repeated the first two words of the phrase. The effect of this strategy was not evaluated in any form. The attention state is inferred based on the head movements of the user. This approach is well-developed in different projects. In the *HUMAVIPS* project, the head position of the human is used to infer their current *VFoA* [She+13]. In this project, the humanoid robot Nao explains different paintings surrounding him to visitors at a vernissage. Similarly to the direction giving system, the robot tries to shift the attention of the user to different locations via speech and deictic gestures [She+13]. This information is used by the robot to decide automatically if it "'should' or 'should not' respond".[JO13] Furthermore, the *VFoA* serves as a basis for engagement detection in this scenario [Klo+11]. In this multi-party interaction scenario, the head movements function as engagement cues for the dialogue. The dialogue manager uses this engagement state, e.g., to determine if there is a new user which should be integrated into the interaction.

In a similar scenario, Dankert et al. also use head movements to detect engagement [Dan+16] and perform re-attention strategies on it [Pit+16]. These strategies are in particular a second reference to the object. Pitsch et al. designed this repair very explicit. The verbalization

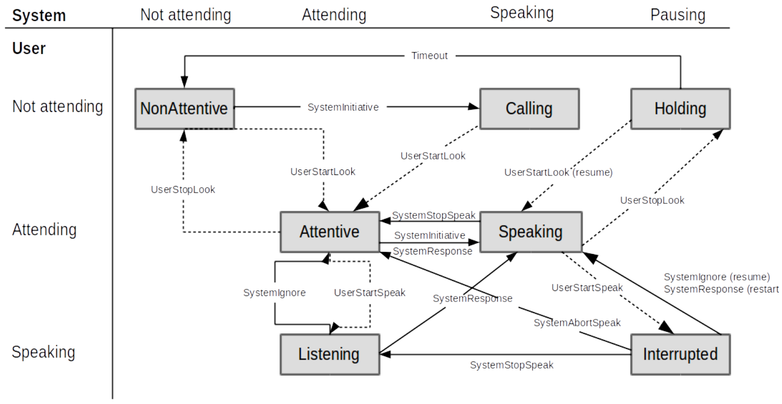


Figure 3.1: Attention model after Skantze and Gustafson [SG09]

uses more details for disambiguation. In addition, deictic gesture and robot's head moves toward the specific object are used [Pit+16]. Similar to the work of Yu et al., the robot decided at specific points of the interaction, especially after deictic gestures, if such a re-attention strategy is required. However, this re-attention strategy requires additional information about the current referred object, but serves as an appropriate strategy to deal with disambiguation.

3.1.2 *Attention Models and Measurements*

Several existing models are related to high-level concepts, such as attention or engagement. In the related work, abstractions for attention could already be found. Yu et al. measure attention by analyzing the head movements of the user and comparing it with the desired head movement [YBH15]. Kousidis et al. use complex task procedures as features, estimating the time of user's attention is required somewhere else [Kou+14]. They do not measure the users' behavior directly to find these special times of interest, instead the driving task itself delivers the exact moment. Skantze and Gustafson presented a model of attention, which incorporates attention states of both, the user and the system (see fig. 3.1) [SG09]. The model uses information of whether the user is looking, speaking, and a set of system events as input. In this system, user's looking behavior is again estimated using head tracking.

Lemaignan et al. try to measure cognitive attention, i.g., the concept of with-me-ness (see section 2.1.2) on a teaching robot [Lem+16]. They carry out a small user study with six children interacting with a nao robot in a teaching scenario. Instead of an eye-tracker, they use an RGB head pose estimation and approximate the *VFoA* of the human. The objects lie inside a central region of the students' field of view. To calculate a with-me-ness level, the authors measure the

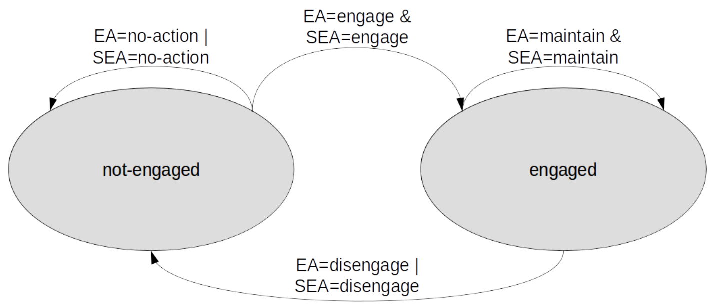


Figure 3.2: Engagement model after Bohus and Horvitz [BH09]: EA is the human's engagement action; SEA is the system's action.

time participants look to the attentional target addressed by the robot. Lemaignan et al. argue that this measurement of the with-me-ness concept is more specific and well-defined than the in *HRI* often used concept of engagement [Lem+16].

In addition to the attention models, a set of engagement models exists (e.g.,[BH09; Klo+11; VJC16; MSS06] Bohus et al. presented an engagement model which consists of two states: *engaged* and *not-engaged*. This model is based on the engagement definition by Peters et al. presented in section 2.1.2. Figure 3.2 visualizes the states and their transitions. Klotz et al. presented an implementation of an engagement model based on the idea of Bohus and Horvitz [Klo+11]. In their model, the dialogue manager receives engagement cues from the agent's perception of the user, i.e., an estimation of the user's *VFoA* based on head tracking and initiated explicit engagement/disengagement actions, such as "Excuse me, would you like to join in?". In a later publication, Bohus and Horvitz present a continuous specification of disengagement with four thresholds to bin the continuum back into four states [BH14]. In this work, they train a model to forecast the moment of disengagement based on several features, including head position, head velocity, and the current state of the ongoing dialogue.

As manifold as the definitions of engagement (see section 2.1.2) are the models for its detection. Most systems use visual features for estimating the engagement state of the user (e.g., [BH09; VJC16; Klo+11]). Vaufreydaz et al. use multimodal information and fuse skeleton information with features from the face recognition to estimate different engagement states [VJC16]. Their model consists of four different states: *someone around*, *will interact*, *interact*, *leave interact*. Other approaches describe and detect similar social engagement states based on spatial relationships [MSS06] or integrate context information into their model and define engagement as a function of context [SC15]. Rich et al. use different types of connection events, including directed gaze, mutual face gaze, and back-channels and build their model using more high-level representations [Ric+10]. Similarly, Pitsch et al. mention that, at the social level, joint attention indicates engagement in an interaction [Pit+09]. Beside these mostly visual approaches, Yu et al. estimate users' engagement in continuous speech using low-level prosodic features and a *Hidden Markov Model (HMM)* that encodes the inherently continuous dynamics of users' engagement states [YAW04].

In addition to these single-user engagement models, some concepts for the detection of group engagement are found in the literature. For example, Salam et al. detect group engagement based on head pose and skeleton joint estimation of each participant [Sal+17]. Further, they examine the impact of personality on individual engagement as well as group engagement. One of their findings is that for detecting individual engagement, participants' personalities play an important role.

Furthermore, a wide range of attention systems, which are mainly bottom-up with low-level automatic attention-related mechanisms can be found. For more information in this topic, look at the survey by [FD14]. Besides the amount of research investigating the effects of incorporating attention, there is a rising interest in the use of hesitations in *HAI*.

3.2 HESITATIONS IN HUMAN-AGENT INTERACTION

This section set out to describe whether the statement that hesitating speech can improve the listener's comprehension in *HHI* presented in section 2.2 is transferable to *HAI*. The research on disfluent speech from a phonetic perspective receives an increasing interest in recent years. The phonetic consequences are discussed (e.g., [Shr99]), models for synthesizing hesitations are presented (e.g., see [ABE07; AEB12; BWS15; DTW16], and the detection of hesitation in human speech was investigated (e.g., see [KT13; SPV13; ABR13]). This section presents the results of several interaction studies with systems detecting (section 3.2.1) or producing hesitations for dialogue coordination (section 3.2.2). Especially, the use of hesitations as intervention strategy and their effects on the interaction are discussed.

3.2.1 *Detecting Hesitations*

While several researchers investigated the detection of hesitations in human speech, (e.g., see [KT13; SPV13; ABR13]) only few evaluated the benefits in *HAI*. Initially, hesitations were only detected to filter them from spontaneous speech and improve speech recognition accuracy (e.g., [KTH10; Can+10; Sch10]).

Later on, it became clear that it is possible to use this information. Bilac et al. presented one of the first real-time systems which uses filled pause detection and gaze to manage conversational turn-taking with robots [BCL17]. They compared two different turn-taking strategies:

- *Gap-turn System*: A silence for longer than 200ms is interpreted as the human's turn release.
- *HOMAGE System*: At the end of an utterance, the system waits for 1.5s. If the utterances ended with a filler, the system interprets this as a turn-keeping signal. Gazing away is also interpreted as turn-keeping signal. Otherwise, the robot takes the turn if the user looks at the robot.

They evaluated their system in an interaction study using within-subject design with 28 participants. As a platform, the Pepper robot is used. Each participant interacted with both systems, by answering five open-ended questions. Each response was classified as success or failure. The failures were been further subdivided into overlap (robot and human spoke simultaneously) and repetition (the user repeated its answer). Table 3.1 illustrates the results by [BCL17]. The HOMAGE system received a better success rate. Additionally, combining both gaze and filled pause detection led to a decrease in robot interruptions, in the case of the robot asking open-ended questions. However, this system received a much higher repetition rate. Presumably, this is rooted in the fact that the HOMAGE system has a longer reaction time by design of 1.3 seconds. Equal reaction times between the conditions would be preferable for the comparison. It is not possible to conclude whether the different effects result from the detection of filled pauses and gaze or from the longer reaction time. Also, an evaluation in other scenarios would be interesting, e.g., with faster question-answer patterns. Nevertheless, this research is a milestone in the design of HAI dialogues, by incorporate the user's hesitations as a communicative act in interaction.

Turn-taking Strategy	Success	Overlap	Repetition
Gap-turn System	50.5%	38.6%	10.9%
HOMAGE System	63.5%	13.5%	26.0%

Table 3.1: Results of turn-taking strategies (after [BCL17]).

3.2.2 *Hesitations as an Intervention Strategy*

Besides the recognition of human hesitations, little research exist investigating hesitations as intervention strategy. In the following, systems using different kind of hesitations are presented and their effect on the interaction, if it was evaluated.

Producing Unfilled Pauses

As mentioned in section 3.1.2, Kousidis et al. suggested a dedicated attention strategy, that allows the user's attention to focus on a distracting difficult driving maneuver [Kou+14]. This is achieved with a self-interrupting dialogue. In an interaction study with 17 participants, Kousidis et al. compared a situated in-car dialogue system with a non-adaptive strategy. The task procedure was as follows: the participants had to drive 30 minutes in a driving simulator. From time to time, the driver received a signal to change the lane for a short time and return to the middle lane after a second signal. During the driving, a dialogue system presented information about calendar entries. After each information, the participant had to answer a short false-true question about the presented information to assess their memory recall (by pressing a button on the steering wheel). The study used a within-subject design. Thus, every participant was faced with both versions of information presentation strategy. For each information, either the non-adaptive or the interrupting version was randomly chosen. In the interrupting condition, the system paused its information presentation during the lane change, whereas in the non-adaptive condition, the system ignored these external events. Each participant received 44 information presentations in total. Kousidis et al. found, that both tasks—lane change maneuver and memory task—benefit from the adaptive version. The participants had a better information recall in the adaptive version and performed the lane change better when the system was silent [Kou+14]. After the lane change, the system repeated the last information chunk and continues speaking. This attention strategy presents an interesting approach for a dialogue system to react in a situated manner with disfluent speech to the attention state of the user. This raises the question, if such an attention strategy is transferable to a *HAI* in a smart home. Kousidis et al. presented a well controlled interaction. The interruption always started at predefined points in the interaction, only the resumption depended on the behavior of the user [Kou+14]. The attention state of the user was inferred based on the difficulty of the task. It would be interesting, if the user can directly influence the beginning of the interruption by itself. In this scenario, the system gives the user time to perform a distracting difficult driving maneuver. It would also be interesting if such a behavior has positive effects in a conversation without a parallel task.

As mentioned in section 3.1.1, Palinko et al. investigated eye-gaze as an implicit signal for turn-taking in a dictation scenario [Pal+15]. The robot produces an unfilled pause as long as the human has not finished the last sentence of the dictate. Only when the participant gazes at the robot, it continues speaking. Surprisingly, they could not find any effects on the task performance itself.

Producing Filled Pauses

As mentioned in section 3.1.1, Yu et al. presented a system that used restarts as intervention strategy, whenever the human attention is missing [YBH15]. However, the effect of this strategy was not evaluated

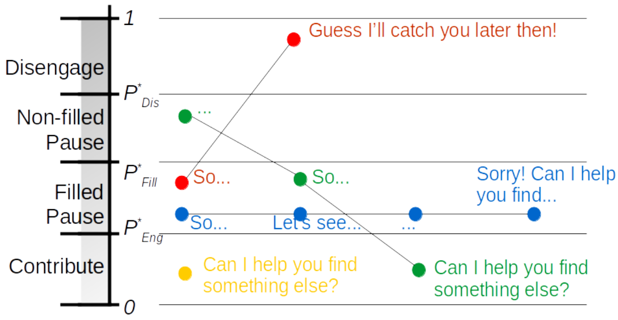


Figure 3.3: Disengagement strategy after Bohus and Horvitz [BH14]

in any form. Bohus and Horvitz presented a disengagement policy which used hesitation actions in a direction-giving robot [BH14]. Figure 3.3 illustrates this policy. Depending on the estimated engagement level of the user, the robot generated different behaviors: the next dialogue act, a filled pause, an unfilled pause, or a disengagement action. This policy was triggered every time the robot had to take the turn and start a new dialogue act. Bohus and Horvitz combined this policy with a forecasting model which anticipated when a user wanted to terminate the interaction [BH14]. Although, the title of the paper suggests something else, the usage of hesitations vowels such as "um", the agent uses short interjections. The phrases are "So..." or "Let's see...", which seem reasonable in this situation. The authors deployed this policy on a Nao robot and ran the system in their building for 5 days. During this time, the robot initiated 158 interactions. The found that a combination of the forecasting model for disengagement and the and the hesitation-action policy to buy time for the system to detect the disengagement perform best. This is measured by less so-called *costly disengagement* actions by the agent. These consist, e.g., of disengagements situations when the agent stops the conversation to early or to late, based on annotations after the interaction. However, they neither performed an evaluation of the effects on the task performance nor on the influence of the appearance of the robot, but rather the ability of their model to detect the right moment for disengaging.

[BWV16] analyzed hesitations in a *HHI* corpus study [BWV16]. Based on their findings, they propose a disfluency insertion strategy for synthetic speech to *buy time* for the system. It consists of the following cascade:

1. *Lengthening:* add lengthening at the next appropriate syllable
2. *Silence:* insert a silence for a maximum of 1000ms
3. *Filler:* insert a filler (e.g., "uhm")

4. *Silence:* insert an additional silence.

However, they neither implement this strategy nor evaluated it on the human interaction partner.

Shiwa et al. implemented one of the first robotic systems that could produce conversational fillers [Shi+08]. In this work, they investigated how quickly a social robot should respond. They could find evidence supporting findings from [Mil68] which states that a system should not take more than two seconds after input to respond. In their interaction study, they found that the users' impression of the system respond time rapidly worsens at two seconds, whereas a system respond time of one second is preferred by the users. Moreover, in a *WoZ* study they could show out, that the Japanese conversational filler "*etto*" improves the users' impression of the robot, if it cannot produce an answer within two seconds. Interestingly, Shiwa et al. only tested these fillers from two seconds up to a maximum of nine seconds [Shi+08]. It would be interesting to see, if it also shows a benefit in faster responses. For long reaction times, the robot simply repeated the conversational filler. This simple strategy of *buying time* was already effective in terms of better user judgments.

The concept of *buying time* for the system has been used several times replicated in different scenarios. Galle et al. for example, propose a model to automatically generate different conversational fillers depending on the estimated response time of their processing modules, such as *Automatic Speech Recognition (ASR)* and *Dialogue Management (DM)* [Gal+17].

Ohta et al. evaluated the effect of filled and unfilled pauses in the users' comprehension, and ratings of naturalness and listenability [OKN14]. They carried out a *WoZ* study with a virtual agent. The virtual agent took the role of a tourist guide, explaining how to travel from the departure point to different destinations. Ohta et al. reported that filled pauses can enhance the users' comprehension and improve the reported naturalness of the spoken dialogue system [OKN14]. In this system, the agent's utterances—including the Japanese filled pauses—are predefined speech recordings of a female member of a drama group.

Skantze and Hjalmarsson presented a model of incremental speech generation based on the *IU*-model (ref. section 6.1.4) [SH13]. It allows generating responses while interpreting speech results and automatically producing hesitations to retrain the floor. The authors tested their model in a *WoZ* study, in which participants negotiated the price of example objects in a virtual flea market. The incremental version of the system starts speaking while the wizard is transcribed the human's speech. However, it waits until the hole utterance is produced by the user before answering the question. The system produces a filled pause, mainly at the beginning of the dialogue act to communicate delays in processing and to retrain the floor. Skantze and Hjalmarsson found, that the incremental system leads to faster responses and better subjective ratings of the participants in terms of efficiency and politeness [SH13].

While the benefit of hesitations in speech does not seem to be sufficiently convincing, the use of hesitations in other modalities is more common. Moon et al.; Hart et al. evaluated hesitation gestures and found out that hesitation gestures can resolve resource conflicts during handovers [Moo+13; Har+14]. Kwon et al. produce hesitation gestures to express their incapability, i.e., to communicate what the robot is trying to do and why it will fail [KHD18] and Dondrup et al. investigated hesitation signals in human-robot head-on encounters as a form of implicit feedback signals [DLH14].

To sum up, researchers propose different models to integrate system-hesitations (e.g., [Gal+17; BH14; SH13; YBH15]) and thoroughly explain the technical requirements and implementation. However, their effect has rarely been studied. In [YBH15; BH14] the models were not evaluated in a user study at all. Instead, the authors deployed their system "in the wild". Yu et al. only drew qualitative observations from the interaction. With this method, they could examine the performance of their autonomous systems, but could not make conclusions on the effects of the interaction as can be done in a classical interaction study. Other researchers evaluated the resulting behavior with an online video study (e.g., [Gal+17]). Participants looked to videos of the agent's behavior and rated them afterwards. This method works well to create a first impression of the resulting behavior. However, it does not provide sufficient data about a real human-agent interaction.

Few researchers investigated their models in interaction. Interestingly, these studies are mostly *WoZ* experiments in narrow domains, concentrating on specific aspects of the interaction (e.g., [Shi+08; SH13]). This approach ensures that the results are not influenced by other side effects and permits a clearer interpretation. However, it tells nothing about the effect in the "real world" or the possibility of autonomous interaction. For example, Shiwa et al. even use prerecorded utterances for their study [Shi+08]. It is highly questionable whether their results are reproducible with state-of-the-art *Text-to-Speech Synthesis (TTS)* and if the implemented decision processes performs as good as human wizard.

However, some researchers present positive effects—such as better task performance—of systems-hesitation on *HAI* (e.g., [Kou+14; Pal+15; OKN14]), but not as a reaction on an inattentive interaction partner.

3.3 SUMMARY OF RESEARCH ON ATTENTION AND HESITATIONS IN HAI

The amount of research investigating the effects of the incorporation of attention in *HAI* is large, and the observed advantages are manifold (see section 3.1.1). It shows that findings and concepts from *HHI* research are transferable to *HAI*. For *HAI*, both the attention of the agent (see section 3.1.1) and the user and are important (see section 3.1.1). Also in *HAI*, the human eye-gaze is a reliable source for their attention—and other high-level concepts. Often this is approximated using head tracking (see section 3.1.1).

I extracted the following insights from models recognizing attention and similar high-level concepts in section 3.1.2. Engagement detection is usually based on visual features, especially the proxemics and head position plays an important role and provides information about the state of the interaction (e.g., *someone is around*, *will interact*, *is in interaction*, *leaves the interaction*). For more precise states during an interaction, context information (e.g., dialogue state or speaking state) and information about the gaze behavior of the human interaction partner are useful. During the interaction, the visual attention is used as turn-talk signal and when reacting to on grounding errors. Even though various effects of the incorporation of visual attention in *HAI* are examined, only little research investigates how to deal with an inattentive interaction partner or how to reacquire their attention. Furthermore, the intervention strategies applied in *HAI* are explicit. The agent uses hand waves, statements about the inattention, or attention-grabbing phrases such as "Excuse me!".

Besides the amount of research investigating the effects of incorporating attention, there is a rising interest in the use of hesitations in *HAI*. The change of the attitude towards hesitations hat can be observed in recent linguistics literature, can also be found in the *HRI*/*HAI* community (see section 3.2). While hesitations were previously only detected to remove them from the speech stream, the additional information is now used in interaction, e.g., for turn-taking (see section 3.2.1). Furthermore, system-hesitations are used deliberately to achieve a specific effect—to improve the quality of the interaction (see section 3.2.2). Some researchers present positive effects—such as better task performance—of systems-hesitation on *HAI*, but not as a reaction on an inattentive interaction partner. However, the research field is new and more research on the effects of hesitations on the interaction between humans and robotic or virtual agents needs to be conducted. The presented works provide important insights into the usefulness of hesitations, but also show several drawbacks, especially in the evaluation of the models. While the presented studies allow the conclusion that disfluent agent speech can improve the human listener's comprehension, it is questionable whether this can be achieved with an autonomous system and the current state of technology.

Research that combines the following points, is still rare: (1) a **model** that is based on *HHI* research, (2) its **implementation** in a *HAI* scenario in which the agent performs as autonomous as possible, and (3) an **evaluation** in a real *HAI* interaction study.

4 THE ATTENTION-HESITATION MODEL (AHM)

The literature review in chapter 3 revealed a gap in the interaction research field. Even though various effects of the incorporation of visual attention in *HAI* are examined, only little research investigates how to deal with an inattentive interaction partner or how to reacquire the attention of the human interaction partner. Furthermore, the used intervention strategies are explicitly, by using hand waves, explicit statements on the inattention, or attention-grabbing phrases.

Additionally, using hesitations intentionally as a mechanism—as a conversational act—has rarely been studied, as the research field of hesitations in interaction is fairly new. The influence of hesitations on the interaction of some models presented is not assessed. The method of interaction studies provide the only opportunity to have the human "in the loop". Furthermore, the effect on the interaction of such models should not only be evaluated in *WoZ* study. It is questionable whether such a system can be implemented autonomously with the current state of technology. It is still questionable whether hesitations can be used as a conversational act for the dialogue management to deal with an inattentive interaction partner.

As a consequence of the research gaps described above, my research will be guided by the following three aspects to investigate my research hypothesis: To investigate my research hypothesis three aspects are especially important: (1) a **model** that is based on *HHI* research, (2) its **implementation** in a *HAI* scenario in which the agent performs as autonomous as possible, and (3) an **evaluation** in a *HAI* interaction study. In this chapter, I develop my **model** to coordinate the human attention and system speech based on the literature review on research from *HHI* (see chapter 2) and the findings from *HAI* (see chapter 3). Thereby, I lay special interest in the information gained from the observation of the *VFoA* of the human interaction partner.

4.1 INTERACTION PHASES AND DISTURBANCES

Figure 4.1 visualizes different interaction phases. Similarly to the work of Sidner et al. [SLL03] and Peters et al. [Pet+05], in regard to the topic of engagement, I distinguish different phases of the interaction (i) *establish* (ii) *maintain* and (iii) *close*. Human behavior needs to be interpreted differently in these phases. According to Peters et al., especially in the first two phases, the monitoring of the human is essential. While before the actual interaction, attention provides insights about the willingness to start the interaction, the agent should

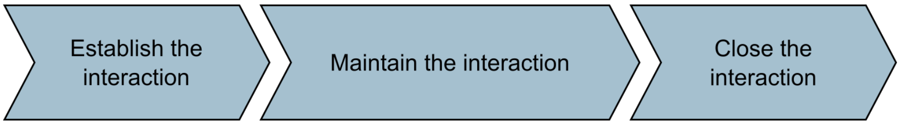


Figure 4.1: Interaction phases.

monitor the potential interaction partner to decide if s/he wants to start an interaction. In the *maintain* phase, the attention gives additional information about the effectiveness of the interaction and displays the level of engagement. However, as Bohus and Horvitz showed, the attention of the human interaction partner can also be used to find a suitable ending point of the interaction [BH14]. In the following, I mainly focus on the *maintain* phase.

I distinguish two mental states of the human interaction partner and discuss the importance of visual attention to detect them. I lay special interest in the states that lead to a disturbed interaction, i.e. *not engaged* and *not understanding* in terms of grounding errors. Of course these two mental states are not independent but influence each other. Figure 4.2 visualizes some of these disturbed interactions, in a two-dimensional space with the dimensions *Engagement* and *Understanding*. I define four different states:

A (**Engagement** ↑, **Understanding:** ↑): The interaction works and the communication between the human and the agent is not disturbed. The human is fully engaged in the interaction and both have a joint understanding of the task.

B (**Engagement:** ↓, **Understanding:** ↑): The human is not fully engaged in the interaction, e.g. because she/he is distracted by an external disruption or not motivated or interested in the interaction, but both have a joint understanding of the task.

C (**Engagement:** ↑, **Understanding:** ↓): A communication or comprehension problem has occurred, while the human is fully engaged in the interaction.

D (**Engagement:** ↓, **Understanding:** ↓): The human is disengaged in the interaction and there is no joint understanding of the task. Communication is no longer possible.

The question arises whether and how the agent can distinguish these different states, or more simply, recognize a disturbed communication (State B-D) that requires a repair action, e.g., an intervention strategy to deal with the inattentiveness of the interaction partner. Based on the literature, we know that the *VFoA* plays an important role in *HHI* (see subsection 2.1.3) as well as in *HAI* (see section 3.1). On this basis, I developed the Attention-Hesitation Model (AHM) to coordinate the human attention with system speech (see fig. 4.3), It draws conclusions about the inner mental state of attention of the human interaction

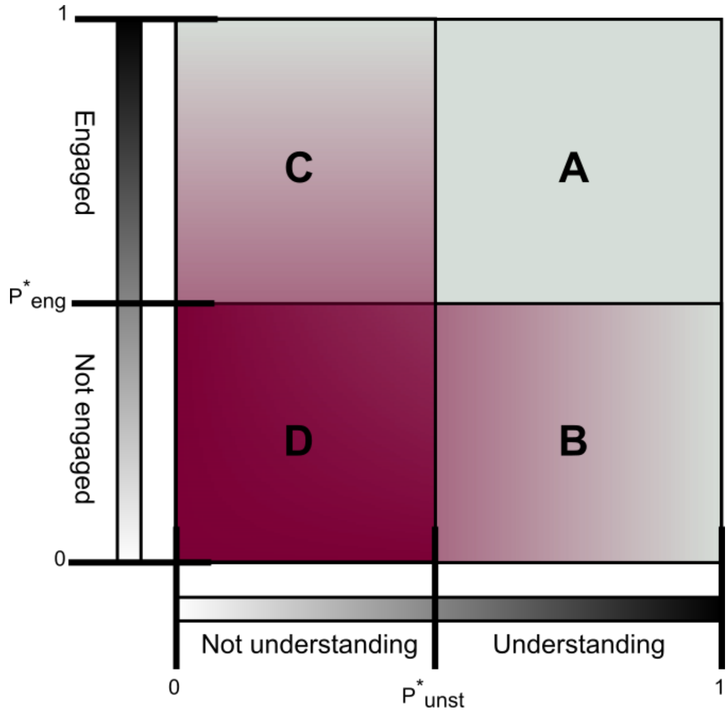


Figure 4.2: Disturbed interactions in different understanding and engagement states of the interaction partner.

partner and react with different hesitation intervention strategies. In doing so, it can be used as a tool to incorporating the human attention into the dialogue management component.

4.2 DIALOGUE MANAGEMENT RESPONSIBILITIES

Based on the previous research, I propose the Attention-Hesitation Model (AHM) to incorporate the human attention into a dialogue management system. Figure 4.3 depicts this model for the coordination of human attention and system speech during a speech act in the *maintain* phase of the interaction. It consists of two state graphs, which influence each other in an interaction loop. On the left is the human state of attention. During the speech act, the human can be attentive or (repairable) inattentive, corresponding to the states A and B-C in fig. 4.2. If the speech act is completed and the human attentive, then the state graph is completed successfully. Based on the attention state of the interaction partner, hesitations are used as repair intervention strategies, depicted in the state graph on the right of the *AHM* in fig. 4.3. The agent is in a speaking state until the speech act is finished. When the human is inattentive, the agent hesitates as an intervention strategy. Thereby, it uses global context information as well as observations from the behavior of the interaction partner to draw conclusions on the inner states of the interlocutor and deal with or repair disturbed interaction states. It interprets missing human's

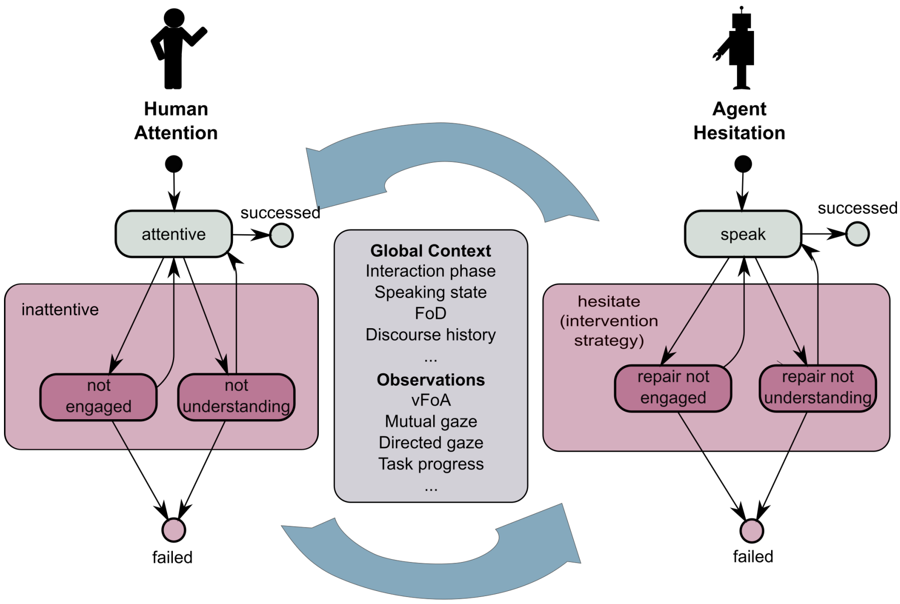


Figure 4.3: The Attention-Hesitation Model (AHM) to coordinate the human attention with system speech.

attention to identify two disturbed interaction states: (missing engagement) the human is not fully engaged in the interaction and (missing understanding) a communication or a comprehension problem has occurred, in terms of grounding errors. Both states require different intervention strategies as repair mechanisms. When the human is unrepairable inattentive, the agent's speech act fails because the agent cannot leave the hesitation state. This corresponds to the disturbed interaction state D in fig. 4.2 and can happen when, e.g., the human leave the interaction during a system's speech act. This model can be further configured with different features for both the attention concept and the hesitation intervention strategies. In the third part of this thesis, I evaluate five different configurations of the model, depending on the scenario and the possible features in the current implementation. I discuss the features used in these models in more detail in the following. In doing so, a large search area for features is described and variables for the implementation are shown.

4.2.1 *When to (re-)act: From Visual Attention to a Cognitive Model*

The model uses button-up information—the *VFoA* from the perception—as well as top-down information from the dialogue manager. Besides the *VFoA*, it incorporates additional information, about the interaction phase and the current speaking state of the agent. My model uses this additional context information, similarly to the attention model by Skantze and Gustafson presented in fig. 3.1. However, the *VFoA* permits conclusion to the cognitive attention of the human interaction partner. In my model, the visual attention is the first requirement for both, engagement and understanding or the other way around, visual inattentiveness can be because of engagement or understanding problems. With task related information from the *DM*, it can interpret the *VFoA* in different ways.

VFOA FOR MUTUAL GAZE: On the one side, paying attention to someone is directly linked to interest (e.g., with mutual gaze we pay attention towards our interaction partner) and is therefore one possibility for measuring engagement, as depicted in section 3.1.2. On the other side, being attentive is the first step towards a common understanding. However, only with the information of the current *VFoA*, the model cannot distinguish between the reason of inattentiveness and can therefore only react with the same hesitation intervention strategy for engagement and understanding problems.

FOD FOR DIRECTED GAZE: As shown in section 2.1.2, directing or following gaze and the resulting joint attention is necessary to create a common ground. Thus, to be able to identify communication errors, like losing the engagement of the interaction partner or figuring out if the interlocutor shares a common ground, the *VFoA* plays an important role. However, it is not enough to only detect the current *VFoA*. The context plays a significant role. Knowledge about the current *FoD*, for example, is necessary to detect following gaze, or perceptual and conceptual "with-me-ness" (see subsection 2.1.2). Furthermore, current discourse information is considered in my model. The current *FoD* plays an important role in monitoring the common ground. Therefore, the model compares the current *VFoA* with the current *FoD* to detect successful directed gazes and the resulting joint attention.

DISCOURSE HISTORY: Additionally, the model distinguishes between a discourse change and an ongoing topic and thus considers the discourse history. Especially for a lack of mutual attention, it incorporates the dialogue history. Therefore, it distinguishes two different situations: (a) the current *FoD* changes and (b) the *FoD* is ongoing. In (a) the model classifies the inattention as an understanding problem because the interaction partner did not follow the attention shift. This means, the user does not look once at the new *FoD* within a time frame. For an ongoing *FoD* (b), the user is inattentive whenever the *VFoA* neither matches the current *FoD* nor the agent itself. The model classifies this state as an engagement problem because the agent misses joint attention throughout the current *FoD* and without a discourse change.

TASK PROGRESS: In addition, the progress of the current task is included in the model if the interaction contains a practical task for the human interaction partner. The interaction partner is inattentive when task progress is missing. This need further monitoring capabilities apart from the recognition of the *VFoA* and depends heavily on the task at hand. I discuss this in detail in section 10.3.

COMBINATIONS: It is possible to combine the listed features to recognize the human's attention or inattention.

Based on these results, the model decides if the dialogue manager should (re-)act. When the interlocutor is inattentive—the model assumes that the agent is either losing engagement or understanding—the agent should react with a dedicated intervention strategy. The other way around, if the model assumes that the human is attentive again, it stops the current hesitation strategy and the agent continues speaking.

4.2.2 *How to (re-)act: Hesitation Intervention Strategy*

When the agent identifies a state that requires an intervention action, the issue is to react appropriately. In the literature, only few strategies can be found to deal with missing attention. These strategies are mostly interrupting. For example, Eichner et al. uses waving a hand and explicit comments about the missing attention of the user to reacquire it [Eic+07]. Another strategy containing explicitly attention catching phrases, such as "Excuse me!" [YBH15]. A further possible strategy to deal with missing attention would be to speak louder. All these strategies are feasible actions, but assume that human attention has to be reacquired immediately. When we see attention more as a resource, we can give the human the time s/he needs. In the *HHI* literature, hesitations are mainly produced to *buy time* for the speaker. However, the idea of the *AHM* is to use hesitations as an intervention strategy to buy time for the listener. This has several possible advantages:

- It is not explicit: Hesitations are a non-intrusive way to deal with missing attention. The agent does not interrupt the current topic of the interaction and explicitly address the missing attention. Rather, it makes it implicitly by hesitating.
- No information of the dialogue is needed: For the production of hesitations, no further information of the current topic of the dialogue is needed. They can be produced at any time.
- No further information is presented: Hesitations do not have to present further information about the current focus of the interaction.

A variety of repair mechanisms are possible, including different kinds of hesitations (see section 2.2). So, which disfluent repair actions are appropriate? Based on the findings of Goodwin (section 2.2.3) speakers use hesitations in different intensities or characteristics when they lose the listener's attention, which I discuss in the following.

UNFILLED PAUSES: One phenomenon is self-interruption. The speaker often pauses the speech stream until the listener's attention returns. This is a relatively simple action. The only variable is the length of the pause.

HESITATION VOWELS: Hesitation vowels, so-called fillers, are mostly used with pauses. This increases the number of variables substantially: length of the pause before a hesitation vowel, after the vowel and the type of the vowel itself, e.g., "uh" or "uhm".

LENGTHENING: A lengthening of a phoneme can often be observed before a pause starts. Here again, the length of the prolongation is a variable.

REPETITIONS: Goodwin reported restarts in different executions (e.g., repetition of the last speech segment). A variable for a repetition can be the number of repeated words, but also the type of the restart can be a variable, e.g., a simple repetition or combinations with insertions, deletions, or substitutions. Repetitions in particular have already been successfully tested in various interaction scenarios (see e.g., [Dan+16; Pit+16]).

COMBINATIONS: It is possible to combine the listed instruments (see e.g. [BH14]).

This results in a large feature search space for designing hesitation strategies. The goal of this thesis is not to evaluate each feature, but rather to find an appropriate hesitation strategy that can be used to improve the interaction. Furthermore, different hesitation strategies may should be used to deal with missing attention based on engagement or understanding problems. In the following, I distinguish theses two strategies:

Working definition: Re-attention strategy

A Re-attention intervention strategy is a reaction to inattentive interaction partners by reason of engagement problems.

Working definition: Highlight attention strategy

A Highlight attention intervention strategies is a reaction to inattentive interaction partners by reason of understanding problems.

While a simple pause can be realized fairly easily without the need of an incremental synthesis, e.g., by pausing the speech stream, a pause should not be underestimated regarding timing. Other hesitations are more complex to implement. The production of hesitation vowels suffers from the fact that they are normally not part of the training corpus for speech synthesis. Therefore, filled pauses, but also lengthening and restarts, pose special demands to the speech synthesis. Some general requirements for the integration of such a system are discussed in the next section.

4.3 DIFFERENCES TO OTHER MODELS

The presented *AHM* has some differences to previous proposed models to incorporate either the human's inattention into the dialogue system or propose hesitations as intervention strategy. However, especially the in combination of these aspects are not considered in the previous work.

HUMAN ATTENTION AS A RESOURCE: The first difference is, that the *AHM* is based on the *capacity theory of attention*—meaning it acknowledge that the human attention is a valuable resource. In contrast to some other models, e.g., which uses hand waves and explicit statements on the inattention [Eic+07] or attention-grabbing phrases such as "Excuse me!" [YBH15], it gives the inattentive interlocutors the time they need. To this end, it uses hesitations to "buy time" for the listener.

HESITATIONS AS A COMMUNICATIVE ACT TO REACQUIRE ATTENTION: System hesitations are used more recently. However, Bohus and Horvitz use it as disengagement action [BH14]. Furthermore, hesitations are used as a communicative act in some systems to signal delays, more precisely to buy time for the system and retrain the floor(e.g., [Shi+08; Gal+17; SH13]). Only, little research evaluated the effect of hesitations to buy time for the listener (e.g., the effect of unfilled pauses [Pal+15] or interruptions and repetitions [Kou+14]). However, they don't use hesitations as a communicative act to deal with missing attention or to reacquire it. To this end, my model is the first one, which uses hesitations as a communicative act to react on missing attention and thereby buy time for the listener.

REACTION DURING A SPEECH ACT: Other system checking the attention, especially at the beginning or end of the turn and do not react during losing the attention during the speech act (e.g., [YBH15; BH14; Dan+16]). In contrast, the *AHM* monitors the agent throughout a speech act.

DIFFERENTIATE BETWEEN REASONS FOR INATTENTIVENESS: Lastly, the *AHM* differences between two possible reasons for inattentiveness, whereas other models only incorporating engagement (e.g., [BH09; SG09]) or understanding (e.g., [Lem+16]).

In addition, even though some models incorporate many of these aspects (e.g., [BH09; Lem+16; YBH15]), other drawbacks can be found. Several researchers provide comprehensive models to detect disengagement or understanding problems but do not provide how to deal with missing attention in the interaction (e.g., [BH09; Lem+16]). Furthermore, some researchers do not even evaluate the proposed model in an interaction study (e.g., [YBH15]), or evaluated not explicit the effect of the used hesitation, but rather other aspects such as the general turn-taking behavior (e.g., [SG09]).

5 SUMMARY OF PART I

In the first part of this thesis (*From HHI to HAI: Developing a Model*), I investigated RQ 1 based on literature from *HHI* and *HAI* research. To this end, I focussed on research from *HHI* in chapter 2. In this process, I revealed the following considerations for the incorporation of human attention in *HAI* based on the *capacity theory of attention*: (1) the human attention capacity is limited, (2) changes in the environment can influence the attention policy of the human and (3) the process of attention allocation can be top-down or bottom-up. Furthermore, I showed that the *VFoA* plays an important role, both in *HHI* (section 2.1.3) and *HAI* (see section 3.1) and that the human gaze is a reliable indicator for their attention and higher cognitive processes.

On the findings from the linguistic and psychological research on *HHI* (see section 2.2) and findings on research focussing on hesitations in *HAI* (see section 3.2), I proposed system hesitations as a reaction to listener's inattentiveness in chapter 4. The presented model—the Attention-Hesitation Model (AHM)—proposes how to incorporate the attention of the human interaction partner into the dialogue system. The maintaining phase of the interaction is thereby of special interest. I presented my model that uses the *VFoA* to draw conclusions about the inner mental state of the human interaction partner. More precisely, it interprets missing human visual attention to identify two disturbed interaction states: "missing engagement" when the human is not fully engaged in the interaction and "missing understanding" when a communication or comprehension problem has occurred, in terms of grounding errors. Both states require different intervention strategies as repair mechanisms. The proposed model differentiates between re-attention and highlight attention strategy and thereby uses hesitations as an additional, new means for these strategies.

Part II

FUNDAMENTALS FOR AUTONOMOUS HAI

6 MODELING DIALOGUE FOR HAI

In the last part of this thesis, the *Attention-Hesitation Model (AHM)* was developed as a tool to coordinate human attention and system speech in a human-agent dialogue. In this part, I investigate which requirements do such a model pose to the design of dialogue systems. To this end, a closer look is taken at the architecture of spoken dialogue systems (section 6.1), starting with the natural language processing pipeline (section 6.1.1) and their drawbacks, especially for modelling the *AHM* (section 6.1.2). Then, various options for implementing the dialogue management component are discussed, the choice of the toolkit used is justified, and the improvements required are shown in section 6.1.3. Furthermore, the relevance of incremental dialogue processing–for both investigating my *AHM* and dealing with the incremental nature of human dialogue—is explained (section 6.1.4). This chapter concludes with resulting hardware and software requirements are posed to the design of dialogue systems from a system engineering perspective (section 6.2).

6.1 ARCHITECTURE OF DIALOGUE SYSTEMS

By definition, *dialogue* is "a conversation between two or more persons" and also "a similar exchange between a person and something else (such as a computer)" [Mer19b]. This already demonstrates, that a dialogue—similar to interaction itself (see chapter 1)—can have various types of interaction partners. A dialogue between humans is just as conceivable as a dialogue between a human and an artificial agent. Furthermore, it can happen in various constellations between humans and agents. In this thesis, the focus is on verbal, dyadic dialogues between one human and one agent. Another definition is the general exchange of ideas and opinions [Mer19b], which can be of course also appear in written form. In this thesis, this kind of written dialogue is not taken into consideration. Speech is an important element of dialogue, but other modalities will also be considered. Figure 6.1 depicts a short interaction from an information request scenario between a human and an agent. To be able to model this short—but already challenging—interaction scenario, a complex dialogue system is needed. In the following, I first present the conceptual architecture of spoken dialogue systems. Even through the focus of this thesis is the coordination of humans attention with dialogue, the overall architecture plays an important role and influence the possibilities. After a short overview of architectures, I present the state of the art

Figure 6.1: Short interaction example of an information request scenario.

in dialogue control systems and discuss several concepts in dialogue modeling. Modelling dialogue systems is a challenging task. Similar to Eckert et al.'s schematic conversational interaction between a human and an agent (see fig. 6.2 on the left), the conceptual architecture of speech-based dialogue systems often consist of a pipeline of five components (see fig. 6.2 right), I name these *the classical natural language processing pipeline*. In the next section, a closer look at it is taken.

6.1.1 *The Classical Natural Language Processing Pipeline*

The classical natural language processing pipeline consists of five component. In the following, these components are introduced.

AUTOMATIC SPEECH RECOGNITION (ASR): The first component in dialogue systems is the *Automatic Speech Recognition (ASR)*, it **recognizes** the spoken language by transforming an audio signal into computer readable text [YD16]. A distinction is made between three types of approaches: Acoustic phonetic, pattern recognition, and artificial intelligent approach [KC16]. Usually, the *ASR* receives the sound of human speech and produces a speech hypothesis with a sequence of words that correspond to what the human has said. Sometimes this recognizer produces an N-best list and/or a confidence for this result. In fig. 6.1, the *ASR* will produce the list of words, in this case "how will the weather be tomorrow evening".

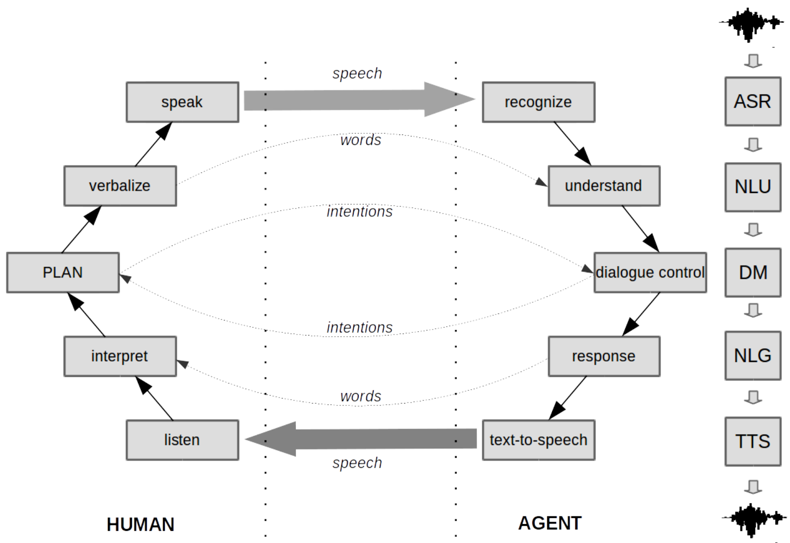


Figure 6.2: Left: Schematic conversational interaction between a user and an agent (after Eckert et al.). Right: Architecture of several spoken dialogue systems: the classical natural language processing pipeline.

NATURAL LANGUAGE UNDERSTANDING (NLU): The *Natural Language Understanding (NLU)* produces a semantic and sometimes syntactic representation of the speech hypothesis, often in the form of a semantic frame with slots to fill. The task of this component is to **understand** the intention of the words and prepare this information for the decision component. One information could be the type of the dialogue act, for example, whether the human *greets*, *farewells* or *asks* a question. In our example, the human *asks* information about the weather. A number of techniques exists to deal with errors made in the *ASR*. These can be using key word spotting, using grammars, or statistical approaches [KKS13]. Cambria and White provide an overview of current research in this field [CW14].

DIALOGUE MANAGEMENT (DM): The *Dialogue Management (DM)* interprets the semantic representation from the *NLU* to decide when and how to (re-)act. It **controls the dialogue**, e.g., by producing a response. In our example the *DM* decided that the agent should react to the human. In this case, with an *answer* to the information request and a self-initiated *statement* with a recommendation. The dialogue control is a main part of this thesis. Therefore, I discuss the state of the art in this research field in more detail in section 6.1.3.

NATURAL LANGUAGE GENERATION (NLG): The *Natural Language Generation (NLG)* cis the counterpart to the *NLU* and a relatively new research field [PSM13]. It converts a semantic representation of the next dialogue act into a natural language **response**. It is often a template-based language generation. In our example, it converts the dialogue acts *answer* and *statement* to the response text: "They say it will rain, don't forget your umbrella".

TEXT-TO-SPEECH SYNTHESIS (TTS): The *Text-to-Speech Synthesis (TTS)* is the last component in the classical dialogue pipeline. It synthesizes written text into an audio signal (**text-to-speech**). Common speech generation approaches use *Hidden Markov Models (HMMs)* and *Gaussian Mixture Modelss (GMMs)* [Bre92; MBM15], but also *Deep Neural Networks (DNN)* are used recently [Lin+15].

6.1.2 *Drawbacks of the Conceptual Architecture*

The *classical natural language processing pipeline* as a schematic representation of dialogue architectures (depicted in Figure 6.2) is a common description of dialogue systems. Examples can be found in various publications, sometimes with small modifications, such as a backend component [McT04; Kul04; BR09; LT00; ABS13; PH05; Pel+12]. However, this model has many simplifications and certain aspects are not considered, which are important for the integration of the *AHM* in an autonomous speech system. In the following, I discuss several aspects of dialogue systems, which are not (directly) part of this schematic representation and their influence on both the conceptual work and understanding of dialogue as well as the implementation of these systems.

OMITTED MULTI-MODALITY: The architecture in Figure 6.2 does not really deny the possibility of multi-modality, but also not explicitly point out the fact that (human-human) dialogue is highly multimodal. Modalities like facial expressions, eye-gazing, or gestures play an important role in communication, but are not represented in this schematic architecture. To acknowledge that dialogue is not solely speech-based, it is important to make these modalities and their contribution to the dialogue explicit in the architecture.

On the input side, the *ASR* usually produces speech hypotheses as recognized words and does not work multimodal. Nevertheless, there are some approaches for audio-visual automatic speech recognition (see [Pot+04] for an overview). Further, input sources are needed—in addition to the *ASR*—which provide information from other modalities. The *NLU* can then use this additional information to create its semantic representation of the speech hypothesis, or the *DM* can directly act on this additional information. Several scientific works investigating the different levels of integration, which is often referred to as the early vs. late fusion decision [And+13; EPK17; GP05].

On the output side, multi-modality can be added as well. The agent can not only react verbally, but also with facial expressions, pointing, or gazing behavior. Like on the input side, the question on which level the other modalities are integrated arises here too (multimodal fission). The *DM* can produce multimodal output, or the *NLU* component can convert a semantic representation into different modalities. In both cases, the different modalities need to be synchronized (e.g., via a suitable realizer such as [VYK14]).

Especially for the *AHM* other modalities beside the speech input are important, such as the eye-gaze to measure visual attention.

NO DEFINITION OF INTERFACES: The presented architecture is only schematic and not well-defined. There is no clear specification of the interfaces between these components. Each dialogue system defines its own interfaces between the modules, inspired by the current domain, the task, and implementation[1]. Thus, achieving compatibility and interchangeability of individual modules is difficult. Furthermore, in some dialogue systems single components fulfil several task of this pipeline at once. This is the case, e.g., when a dialogue manager directly produces a textual response, circumventing the *NLG* (e.g., [PW10b]). From the technical point of view, this make it difficult to compare different dialogue systems or reuse existing components to build a new system. As this research topic involves different disciplines and the components have distinct responsibilities, it would be favorable to be able to utilize previous research and processing components. However, to be able to integrate previous work, a modular system with clear responsibilities and standardized interfaces is mandatory. Furthermore, this modularity allows flexibility in practice. On the one hand, it supports simple exchange of processing modules. On the other hand, it allows the comparison of modules and whole systems. However, this leads to more effort, in terms of specification and implementation. For the integration of the *AHM* it is essential to have a clear structure of responsibilities, as it is a sub-part of the dialogue management and has to be well embedded in an (existing) system.

VAGUE SPECIFICATION OF TASKS: Due to simplification and not well-defined interfaces, several modules are not part of the architecture—like the already mentioned components for the recognition and generation for nonverbal modalities. For these, it is still questionable, where data fusion takes place in this pipeline, in the conceptual architecture as well in the implementation.

From the technical point of view, the lack of specification of interfaces makes it difficult to keep a clear distinction between these processing modules. Several *databases*[2] are not explicitly mentioned. Examples therefore are the dialogue history and the dialogue context. Most dialogue architectures have an extra back-end module for the

1 Or don't even specify these interfaces.
2 These databases could be parts from other modules, e.g., the *DM*

communication with external applications or hardware units. Furthermore, even speech processing modules like the *Voice Activity Detection (VAD)* [GZF18] or *Automatic Addressee Recognition (AAR)* [RK16] are not explicitly mentioned. *VAD* classifies the audio signal into speech and non-speech, and is typically a submodule of the *ASR*. The *AAR* estimates whether the agent is addressed or not, which is part of the dialogue control. In some dialogue systems these modules are separate (e.g., a separate *VAD* [BSG07] before the *ASR*): They provide their information to other modules and are not a subtask within the classical processing schema. The same applies for the *AHM*—it can be a module in the *DM* or a separate one. This unclear task specification in the classical pipeline not only results in different dialogue system structures, it raises whether we identified the correct (and all) tasks in our schematic architecture in the first place.

NON-INCREMENTALITY: The incremental nature of dialogue is disregarded in the default pipeline. Dialogue processing in *Human-Human Interaction (HHI)* is highly incremental and recent research focus on this topic (e.g., [SS11; BS12; SH10; Mic20]). Non-incremental systems have several problems and drawbacks. Because processing starts after the input is finished, these systems are not only slow by design, but cannot consider various feedback mechanism from *HHI*. It is not possible to react during the human speech, and the agent cannot react on human feedback signals while it is speaking. For the *AHM*, incremental capability are necessary. Incrementally can influence various aspects of the dialogue system, such as the interface design, the module behavior and even the information flow in the system. I discuss this topic in more detail in section 6.1.4.

SIMPLIFICATION OF THE TOPOLOGY: Based on the previous issues, the general architecture as a pipeline is an enormous simplification of the topology of natural language processing. From the technical point of view, regarding the *DM*, the question arises on which basis the decision making is performed. If we try to write down the function of the decision-making process, which variables are required? In the figure above, the only input of the *DM* is the output of the *NLU*. Certainly, other variables can influence the decision-making process, e.g., the system status, dialogue history or the current context. Simultaneously, the *DM* can influence the *NLU* with the current dialogue context (e.g., [PH05]). The pipeline above does not allow feeding information back to previous modules. Furthermore, the decision-making part of a dialogue system is not necessary a single component, but rather can be spit into a number of components or processes. These make discsions at different levels and can be performed top-down or bottom-up. It is questionable whether a single *DM* component is required. Whether a single component or a set, a dialogue system needs a subsystem for decision-making which is responsible for the actions is needed.

Besides the representation as a pipeline, few other representations exist with different levels of detail. It can be a single black box for the whole dialogue system [Kip15; ZE16] or a star topology, such as implementations of the information-state architecture [LK16] or voice dialogue systems [SBH95]. The use of a end-to-end learning for the whole dialogue system (e.g., [ZE16]) makes it impossible to reuse existing components. Furthermore, the underlying interaction models are not necessary transparent and comprehensible. Therefore, debugging and the adjusting the agent's behavior it's difficult or rather impossible. Schlangen and Skantze present a conceptual framework—the general, abstract model of incremental dialogue processing—which can deal with different network topologies [SS11]. I discuss this in more detail in section 6.1.4. Furthermore, the resulting requirements are discussed later in this thesis (see section 6.2.1). In the following, a closer look is taken on the dialogue coordination.

6.1.3 *Coordination of Dialogue*

The *DM* component is usually the decision-making part of a dialogue system. It controls the dialogue flow and usually receives its input from the *NLU*. Based on the semantic representation of spoken utterances, the *DM* produces some output to the *NLG*. Basically, the *DM* has two responsibilities:

1. Decide *when* to (re-)act
2. Decide *how* to (re-)act.

There are several types of *DM* and also approaches for grouping these. Schlangen propose three groups: models of *(1) what was said*, models of *(2) what to say next*, and models of *(3) when to say it* [Sch05]. This grouping is based on the function of the underlying approaches. They can be further divided into different approaches, which is explained here in more detail.

FINITE-STATE BASED: One of the simplest approaches to modeling dialogue is to represent the dialogue manager as a finite state machine. In this approach, the whole dialogue is structured as a graph. This graph specifies all legal dialogues. A set of states represents all possible system dialogue actions, whereas a set of moves between states represents human responses and the transition to a new state. The main advantage of this approach is the possibility of rapid prototyping. It is possible to sketch a *Human-Agent Interaction (HAI)* without the need of training data or complex models. For the developer, this kind of representation is transparent. The major advantage is simplicity and intuitiveness for the developer. Quite a number of approaches are based on finite state machines or state charts (e.g., [BWB09; SA12]), especially in the *Human-Robot Interaction (HRI)* this is a common method. However, to model less restricted interactions, which require more states, the dialogue graph has to be enriched leading to a fast growing difficult to handle population of states.

FRAME-BASED: Another common approach, particularly in the information retrieval community, is frame-based. This is especially appropriate for classical form-filling applications. The central data structure is a frame with slots to fill. A frame consists of a set of needed information, forms the context for utterance interpretation and dialogue progress. The goal of the *DM* is to monitor the current frame and to fill in slots. Thereby, multiple inputs and flexible order are possible, which makes it well suited for complex information access tasks. Unfortunately, it is ill-suited for complex problem solving tasks.

INFORMATION STATE UPDATE: Traum and Larsson propose the information state approach for dialogue management to account for updates of information through the ongoing interaction [TL03]. Their information state-based theory consists of various components. First, a description of informational components (e.g., beliefs, desires, user models...) is needed with a corresponding formal representation. Next, a set of dialogue moves must be defined. Additionally, a set of update rules in needed, which change the current information state based on the last dialogue move. The update strategy decides which update rule is selected. According to Traum and Larsson this approach combines the advantage of both previous concepts[3]. There are multiple implementations of this approach, e.g., TrindiKit [TL03]. However, Peltason could show in her thesis, that this approach does not provide a systematic solution for the asynchronous nature of the underlying processes in *HRI* [Pel14].

PATTERN-BASED: Peltason proposed to model dialogue for *HRI* based on interaction patterns [Pel14]. An interaction pattern describes recurring and configurable dialogue structures on a general level and can be formalized as a transducer augmented with internal state action. In combination with a generic protocol for task representation [Lüt+11], the concept of interaction patterns supports rapid prototyping of *HAI* dialogues. The concept of task allows the dialogue to react on events during a system action. Figure 6.3 visualize a life cycles and the corresponding task events. This approach has this advantage of the finit-state based approach, however with a better scalability. Furthermore, by interleaving interaction patterns, flexible interactions become possible [PW10a]. Peltason and Wrede implemented this approach in the *Pattern Based Mixed Initiative Interaction*

3 for more information see e.g., [Sch05]

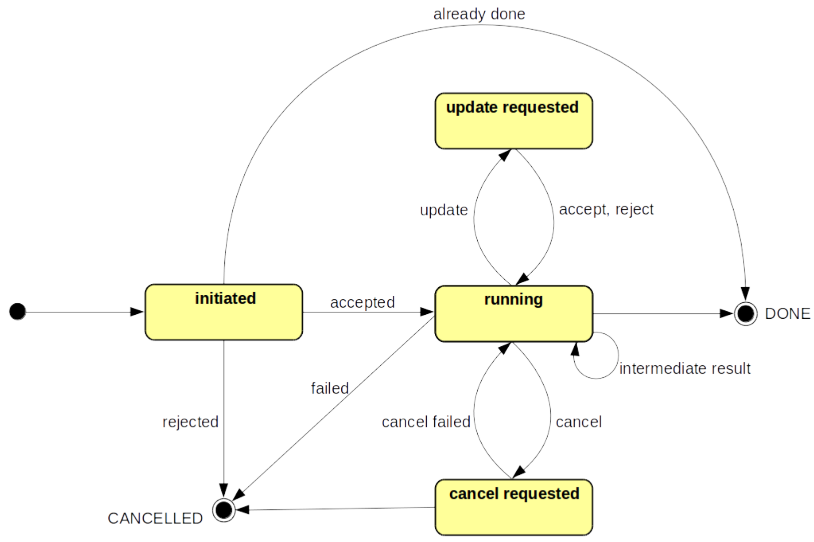


Figure 6.3: Life cycles and corresponding task events of a system task after Lütkebohle et al. [Lüt+11]

Toolkit (Pamini) [PW10b]. The authors argue that their toolkit meets the requirements of advanced dialogue modeling for *HRI* and at the same time exhibits a better scalability than existing concepts. In doing so, the toolkit supports rapid prototyping and counteracts the lack of generalizability of previous *HRI* dialogue systems. In contrast, this approach does not address the incremental nature of human dialogue.

STATISTICAL: Agents should interact in dynamic and uncertain environments. More sophisticated approaches to deal with this challenge are based on *Bayesian Network (BN)* or *Partially Observable Markov Decision Processes (POMDP)* [DR07; HSC05; Lis13]. These models focus on modeling the dynamic and uncertain environment that robots have to deal, by explicitly representing these uncertainties. Another advantage is the reduction of development costs, by learning dialogue policies. However, the probability parameters need to be learned or crafted manually, which tends to be expensive for more complex interactions. A major disadvantage of using these methods is that lots of training data is required, which is often not existing, particularly for rapid prototyping of new interactions. Furthermore, the resulting model may perform well, but does not necessary help us understand the underlying mechanisms[4]. This applies to both the developer creating such interactions and the human in the interaction. In addition, such models are not always simply extendable without retraining.

4 For more information on this problem, see the 'explainable ai' research area, e.g., in [Goe+18].

Of course, other different approaches to the classification of dialogue management systems exists. Systems can be divided into whether the human or the system takes the initiative. Speech assistants, such as *Smart Personal Assistant (SPA)* Alexa or Google, are typically human-initiative systems. On the other side, virtual assistants (e.g., travel assistants) are usually system-initiative which fits well to the information retrieval problem. Mixed-initiative approaches, such as *Pamini* [PW10b] are somewhere in between. These are well suited for the mixed initiative nature of *HAI*.

There are multiple different approaches to model the interaction between a human and a (virtual) agent, with different benefits and drawbacks. To investigate whether an agent can use hesitations—based on the attention of the human interaction partner—to improve the *HAI* in a smart home, an appropriate dialogue system need to be built.

The recent trends to learn a dialogue policy is promising. However, it requires training data, which is not always available. Especially for rapid prototyping of interaction scenarios and behaviors, this is not suitable. Furthermore, the resulting system has limited transparency for the developer, and is not easily extendable without retraining. The information state approach does not provide a systematic solution for the asynchronous nature of the underlying processes of *HRI*. Simple state charts do not scale with the amount of interaction options, but are well-suited for local interaction patterns. With a generic interface for system tasks and the possibility to combine local interaction patterns, *Pamini* provides an excellent basis for dialogue modeling. Only the possibility to integrate more fine-grained control mechanisms (e.g., to react incrementally on results) is missing. The need of the concepts of incremental dialogue processing is discussed in the following section.

6.1.4 *Incremental Dialogue Processing*

As noted above, dialogue in *HHI* is highly incremental and recent research focuses on this topic in *HAI* as well (e.g., [SS11; BS12; SH10; Mic20]). Previous dialogue systems neglect the incremental nature of dialogue, which leads to The following problems in dialogue processing in general and for the integration of the *AHM* specifically.

Problems of Non-incremental Processing

Non-incremental processing is slow by design. In current systems, the processing of spoken utterances starts after the automatic speech recognition produces a result. This is usually after a turn-end (typically a short silence) is detected. The agent takes some time to respond, which consequently leads to disfluent *HAIs*. Incremental processing can provide several benefits for *HAI*. Smaller interaction chunks lead to faster responses and quicker feedback. The human input is processed earlier, which makes it possible to react earlier.

Another problem of non-incremental processing is, that it cannot deal with speech disfluencies or distinguish between hesitations and turn-endings. Silence is not always a marker of a turn-end. Other modalities, such as gazing behavior, should be considered, to distinguish a human speaking pause meant as a turn-end from a hesitation.

A third problem is, that it is not possible to respond during the human speech. With incremental processing, it is feasible to react while the user is still speaking, e.g., to show listener signals like smiling, nodding, shaking the head, or the visual attention. Interrupting the user—if needed—is conceivable, too.

Furthermore, with non-incremental processing, it is not possible to react to human feedback signals during the agent's speech. If the agent loses the engagement or attention of the human interaction partner, while it is speaking, in non-incremental systems it is not possible to adapt the current speech plan. The human interaction partner may not even be observed during system speech. However, successful interaction requires the coordination with the other interaction partner. Consequently, it is necessary for the agent to observe the human and to adapt its actions. With incremental processing, it is possible to react and adapt the current speech plan. One option would be a strategy to regain the lost attention of the human interaction partner. Therefore, it can be useful or even necessary to interrupt the robot's current speech act or reschedule it.

To sum up: incremental processing is a requirement to allow closed and faster feedback loops in *HAI*. Without such possibilities, an adaption of the ongoing behavior is not possible and the *AHM* cannot be integrated into an autonomous dialogue system. Incremental dialogue processing is a relatively new research field. Therefore, there are few works on how incremental dialogue processing influences the underlying dialogue system architecture. Schlangen and Skantze propose a conceptual architecture, which I will explain in the next paragraph.

A General, Abstract Model of Incremental Dialogue Processing

Schlangen and Skantze propose a general and abstract model for incremental dialogue processing [SS11]. The *IU-model* is based on the concept of *Incremental Unit (IU)*. *IUs* are the smallest 'chunks' of information that can be passed between connected *processing modules*. The length of such chunks thus determines the level of incrementally of the system. The underlying idea is a network of processing modules (*processor*), which have a *left buffer* and a *right buffer*. These modules take input from their *left buffer*, perform some kind of processing and provide output to their *right buffer*, which can be the input for the next processing module. Figure 6.4 visualizes an example of an *IU*-processing module.

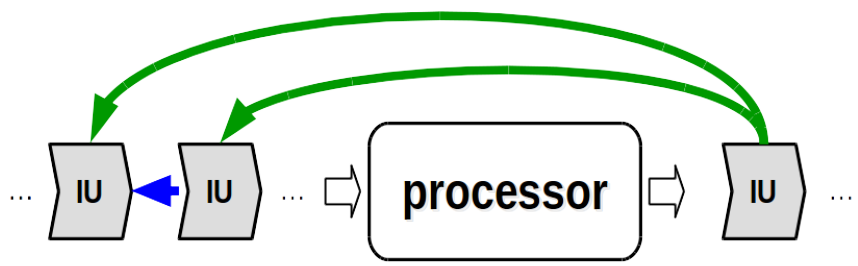


Figure 6.4: Example of an *IU* processing modules (*processor*): Green arrows represent *'grounded-in' links*, blue arrows are *'same level' links*.

Incremental systems consist of a network of processing modules (see fig. 6.5) that work on these *IUs*. They serve as the basic units,

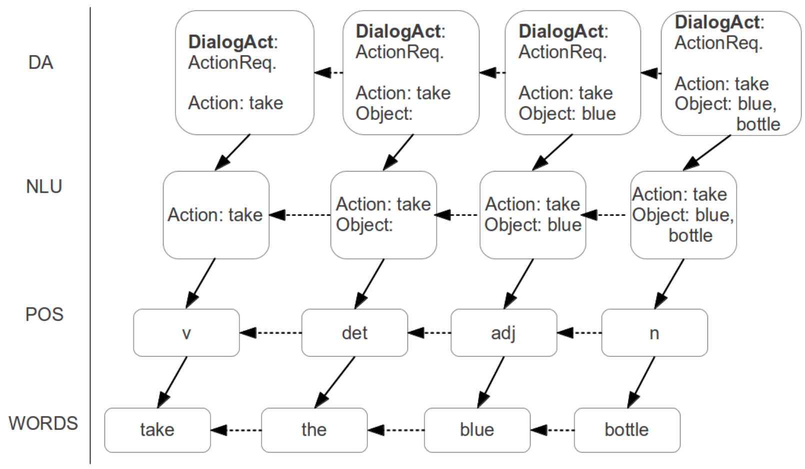


Figure 6.5: Example of an *IU* network. It describes the processing pipeline of *IUs* from the ASR results *wordIU* through to *DialogActIU*. Dashed arrows represent *same level links*, solid arrows are *grounded-in links*.

which can be subject to (post-hoc) changes during processing, e.g., affecting incrementally produced *ASR* results or subsequent syntactic or semantic parsing results. In the example in fig. 6.5 the *NLU* recognizes dialogue acts stated by the user. In doing so, it tries to build a representation (*DialogActIU*) based on the results of the *ASR* (*wordIU*). In our example "take the blue bottle" the *ASR* recognizes the single words "take", "the", "blue", and "bottle" successively. The *IUs* can be in one of different states:

ADD: indicates that a new *IU* has entered the processing module

REVOKE: indicates that a previously added *IU* has been updated or revoked

COMMIT: indicates that an *IU* has been finally committed and will not be changed anymore.

After an *IU* is added to a buffer, it is still possible that a previous module changes its hypothesis, e.g., by first generating a wrong hypothesis for the "bottle" such as "mottle". In such a case, the *IU* is revoked, and a new *IU* may be generated. If an incremental unit is marked as committed, it is considered unalterable and cannot be revoked. Importantly, *IUs* can be part of a larger unit, e.g., words that can be combined to a phrase. The *IUs* have two different kinds of connections. *Same level links* connect *IUs* which are produced by the same module and reflect their temporal order. *Grounded-in links* represent on which *IUs* they depend, e.g., a phrase depends on individual words. This results in the possibility to represent the hierarchical structure of the processing. This model of incremental processing is partly implemented by Baumann and Schlangen in the *Incremental Processing Toolkit (inprotk)* [BS12].

Benefits and Challenges of Incremental Processing

As already noted, incremental processing has considerable advantages over traditional dialogue systems:

- It should be faster by design.
- It allows the distinction between hesitation and turn-end.
- It allows incremental feedback during a humans' speech.
- It allows reacting to feedback during the agent's speech.

However, these new interaction possibilities results in new challenges for dialogue system modeling. First, the processing modules have to deal with adaption or retraction of their inputs. Especially, for the *DM* this is an enormous challenge. It is still possible, that the *DM* only reacts to unalterable results, but for fast responses it is necessary to react to incomplete incremental results in some reasonable way. When such an incremental result needs to be revoked, the *DM* should be able to revert unnecessary actions. Of course, this is only possible if the following modules allow such behavior. To act on incomplete input, always bears the risk of adaption and retraction. A balance must be found between the benefit of fast reaction time and the cost of repairing overhasty actions. It should be carefully considered, which actions can be reverted easily.

However, for my research question, incremental processing lays the foundation for the adaptive behavior of the agent. It is not possible to determine the moment in which the agent loses the attention of the human interaction partner in advance. Consequently, it is necessary to adapt the ongoing behavior of the agent "just in time". Therefore, incrementally—at least on the output side—is mandatory.

6.2 RESULTING REQUIREMENTS: SYSTEM ENGINEERING PERSPECTIVE

To manage the dialogue between a human and the agent, two main questions need to be answered: (1) *when* to (re-)act and (2) *how* to (re-)act. While implementing a hesitating agent—based on the attention of the interaction partner—several requirements from the system engineering perspective have to be met. I discuss them in the next section, divided into software and hardware requirements to implement the *AHM* in an autonomous *HAI*. Dialogue systems for smart homes need to full fit some general software requirements to be able to use and benefit from the various sensors and actuators within it. In section 6.1.1, I discussed the conceptual architecture of dialogue systems. Based on its drawbacks, I extract several requirements for the general dialogue architecture in section 6.2.1. Furthermore, to integrate the *AHM* into dialogue coordination, the *DM* need to full fit requirements as well (see section 6.2.1). This section closes with resulting hardware requirements in section 6.2.2.

6.2.1 *Software Requirements*

The resulting system has to meet a number of essential software requirements, which I list briefly.

Requirements for General Dialogue Architecture:

It is difficult to satisfy the various requirements of dialogue systems and simultaneously have a comprehensible representation of the system. From the previously discussed aspects, I extract the following requirements for the general architecture of dialogue systems:

MULTI-MODALITY: Explicit specification of *multi-modality on the input* as well on the *output* side of the system. The agent needs to be able to perceive the human interaction partner. To answer the first question—*when to react?*—it is necessary to observe the interlocutor. Different input signals have to be integrated into the dialogue system. Not only speech input, but also the eye-gaze plays an important role in the interaction. Furthermore, the agent should be able to express its own visual attention via its gazing behavior.

INCREMENTAL PROCESSING CAPABILITIES: Explicit *consideration of the incremental nature of human dialogue*. To realize the proposed hesitation intervention strategy—in this case the answer of the question of *how to react?*—a possibility to change the ongoing speech plan is needed. Therefore, the capability of incremental processing (see section 6.1.4), especially on the synthesis level, is mandatory.

TOPOLOGY: The system should be *modular*, i.e., it should easily be possible to exchange different components, such as *ASR* or *TTS*. Therefore, a clearly defined interface between these components is needed, and an appropriate middleware has to be chosen for the communication between these components. Furthermore, a clear definition of the interfaces between the modules—which are based on the tasks of the modules—facilitate the integration of the *AHM* as a separate module in the dialogue system. Based on the previous requirements, the topology of the dialogue systems is a *network*, rather than a single pipeline. To this end, hierarchical decision-making becomes possible, as well as feeding information back to previous modules.

GENERALIZABILITY: Furthermore, the resulting system should be *platform-independent* to be able to run on different agents within a smart home, regardless of whether robotic or virtual agent. Additionally, the resulting dialogue system should be *scenario independent* to allow reusability and the generalizability of the system.

These requirements not only influence the software engineering part of modeling dialogue. They also have an impact on the understanding of dialogue as a cognitive process and allow insights into human interaction. Especially dialogue coordination is influenced by these requirements.

Requirements for Dialogue Coordination:

In section 6.1.3, different approaches to model the *DM* between a human and a (virtual) agent, with different benefits and drawbacks. In the last section, several requirements regarding the general dialogue system are defined. However, to investigate whether an agent can use hesitations—based on the attention of the human interaction partner—to improve the *HAI* in a smart home, an appropriate dialogue coordination system need to be built. First, the *DM* should consider the previous mentioned requirements. For multi-modality, the dialogue coordination can either take different modalities directly into account or indirectly by multi-modal fusion or fission in the *NLU* and *NLU* components. Incremental processing capabilities are needed for the *DM* component as well. It should be able to react incrementally to interrupt the ongoing dialogue act and start an intervention strategy. The topology also plays an important role to be able to integrate the *AHM*. Lastly, the *DM* should support the generalizability of the system. Besides this, the dialogue management component needs to fulfil the following additional requirements.

System-initiative as well as agent-initiative should be supported. This allows various interaction possibilities in the smart home, starting from simple question-answer interactions up to more complex interaction such as supporting during cooking. Furthermore, the underlying model needs to be transparent and comprehensible to facilitate debugging. For the integration of the *AHM*, two aspects are important. At first, the *DM* needs a *representation of human mental states*. Based on the perception, a decision must be made whether the interaction partner is attentive or not. Therefore, the agent needs a concept of attention and opportunities to detect it. As a feature, the eye-gaze of the interaction partner should be included. Second, *DM* needs corresponding *intervention strategies* to deal with this disturbed interactions. To this end, the current dialogue act must be interruptible, and a hesitation strategy be started.

As already state in section 6.1.3, the toolkit *Pamini* fulfils several of these requirements. The use of interaction patterns with system task descriptions counteracts the lack of generalizability of previous *HRI* dialogue systems. Furthermore, system-initiative as well as agent-initiative is supported and the concept of interaction patterns facilitate the comprehension of the underlying structure of the dialogue. However, this approach does not address the incremental nature of human dialogue. In addition, it has no representation of interaction partner's mental states or the opportunity to initiate intervention strategies during a dialogue act.

6.2.2 *Hardware Requirements*

The hardware has to meet several requirements as well.

POSSIBILITIES OF EXPRESSION: To establish joint or shared attention an anthropomorphic appearance is preferable. The agent should have something like eyes, to look at specific objects or the interaction partner. Additionally, possibilities for the expressions of (facial) emotions can be interesting. For socially intelligent agents, humanoid appearance facilitates a natural interaction. Although, it should be considered, that the agent does not fall into the uncanny valley [Mor70]. Furthermore, the agent should be able to express the current *Focus of Discourse (FoD)*. Besides the verbal modality, other actors are desirable.

PERCEPTION OF THE INTERACTION PARTNER: Besides its possibilities of expression, the agent needs sensors to be able to percept the human interaction partner. At least a microphone and camera are necessary for a first realization of my *AHM*. Of course, other sensors, e.g., laser-scanner, can improve the perception of the user.

NATURALNESS OF THE ENVIRONMENT: The environment plays an important role for the interaction. Participants should feel comfortable, and the environment should not look like a laboratory. Of course, it would be the best to test the *HAI* "in the wild". However, this introduces further variables into the evaluation. A balance must be found, between the possibility of a natural and free interaction and a controlled environment.

7 REALIZATION OF THE DIALOGUE SYSTEM

In the last chapter, concepts of modeling *HAI*, especially regarding the speech-based dialogue and the requirements posed by the technical realization of my *AHM* are discussed. In the following, the technical realization of a dialogue system which is the fundamental work that allows autonomous *HAI* and the investigation of the effects of my *AHM* in interaction is presented. To this end, the choice of the research platform is justified in section 7.1. In addition, the general architecture (section 7.2) of the smart home system and the concrete modules for the dialogue system (section 7.3) are presented. In section 7.4, it is discussed shortly how the presented system implementation meets the requirements. The chapter concludes with first scenarios and research studies—using this dialogue system—including a first integration of human gaze (section 7.5) and a summary of the second part of this thesis (chapter 8).

7.1 RESEARCH PLATFORM

In this section, the used research platform is described, in particular the Cognitive Service Robotics Apartment (CSRA) and its conversational agent *Flobi* (section 7.1.2). The *Cognitive Service Robotics Apartment (CSRA)* is designed for a living lab approach and serves for (1) demonstrating the state of the art in smart homes, (2) performing interaction experiments within it and (3) being a working and meeting space for researchers.

7.1.1 *Cognitive Service Robotics Apartment*

The research platform Cognitive Service Robotics Apartment (CSRA) offers a good balance between the possibility of natural and free interaction and a controlled environment [Uni19]. Figure 7.1 shows a floor plan of the intelligent apartment. In total, the apartment consists of three rooms and covers an area of approx. $60\,\mathrm{m}^2$. There is a large living-dining area with an open kitchen and a separate bathroom. Additionally, there is an extra multi-functional room, currently the robot's room. We equipped this research environment with various sensors (e.g., motion sensors, a tactile floor, or sensors for the opening state of the cupboards) and actors (e.g., speakers, lights, and visual displays). In total, more than 300 sensors and actors allow the realization of multi-modal interactions with and within this intelligent environment [Hol+16]. Via network-enabled Basler cameras and Rode

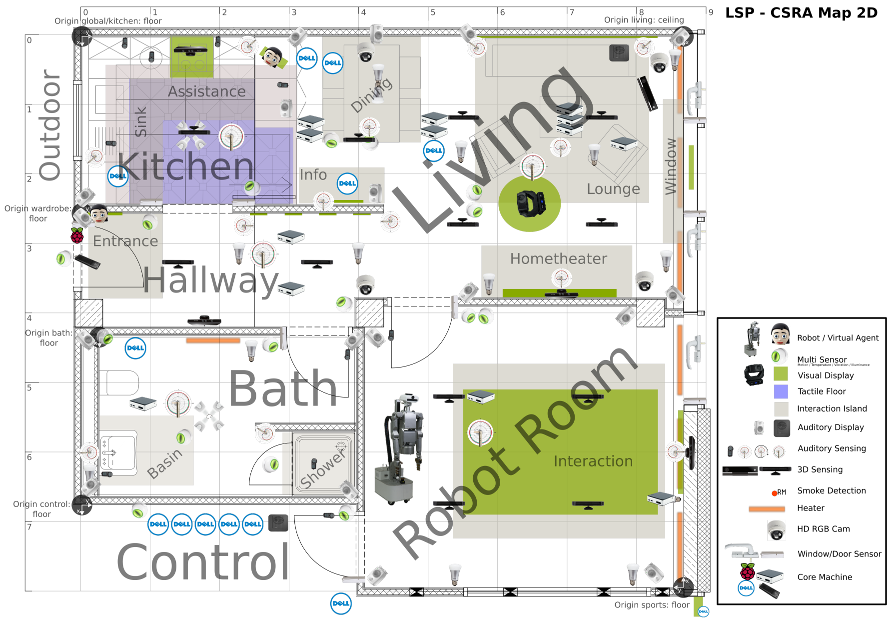


Figure 7.1: The 2D-map of the CSRA including the position of the sensors and actors.

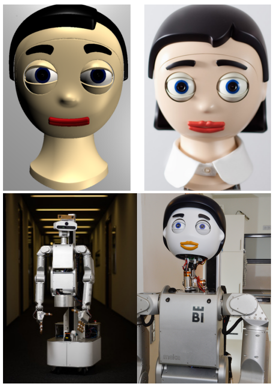
NT55 cardio- and omni-directional microphones mounted at the ceiling of the apartment, it is additionally possible to observe the whole interaction area.

We created two interaction islands with virtual agents in the apartment. As a virtual interaction partner, we use a simulation of the anthropomorphic robot head *Flobi* [Lue+10]. Additionally, the robot *Floka* is part of the apartment and can be used in two different configurations—with a sensor head or a social head—for different tasks within in the apartment. Figure 7.2b shows a picture of all conversational partners.

7.1.2 *Anthropomorphic Robot Head Flobi*

The anthropomorphic robot head *Flobi* is developed at the Bielefeld University. According to Hegel et al., several key concepts influenced the design of the Flobi head [HEW10]. To avoid the uncanny valley effect, Flobi is designed as a cartoon-like character, which utilizes basic concepts of the baby face schema. It has a small chin, big eyes and a very fine nose. These research prototype is specifically designed for social interaction with humans and therefore equipped with social skills. Flobi has 18 *Degrees of Freedom (DoF)*: three in the neck (pan, tilt, roll), six for the lips and the remaining for the eye movement. Each eyebrow can be rotated individually, and each of the four eyelids can be opened or closed. The eyes itself can look up or down together (tilt) but separately to the left and right (pan) to focus objects or persons. It can also show emotional facial expressions. Figure 7.3 visualizes the

(a) *CSRA*.



(b) Conversational partners in the *CSRA*.

Figure 7.2: (a) The *CSRA* research platform and (b) conversational agents within the apartment: (upper left) the simulation of *Flobi*, (upper right) the original head and (below) the different forms of the robot *Floka*.

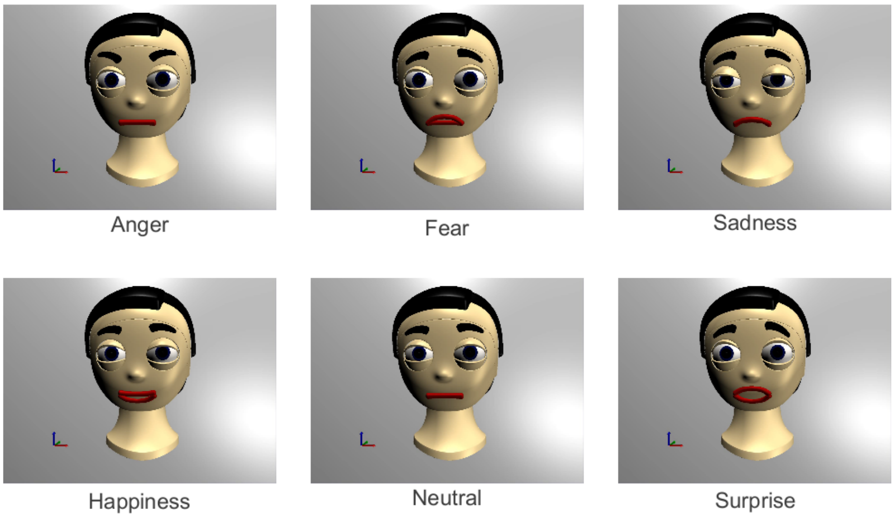


Figure 7.3: Basic emotions of the Flobi simulation.

neutral expression and the five basic emotions: anger, fear, sadness, happiness and surprise. Additionally, it can look at specific targets. Thereby, it not only moves it eyes but can also use its head [Sch20]. To demonstrate responsiveness, Flobi looks at faces, blinks, and simulates breathing.

Both Flobi heads are animated in a simulation based on the *Modular Open Robots Simulation Engine (MORSE)* [Ech+11]. To control the simulated actuators the same API is used as for the movements of the "real" Flobi head. This allows us to test facial expressions, head movements or whole scenarios with the simulation and use them with the robot without changing the implementation. Additionally, the simulation itself can also be used for human-agent interaction scenarios.

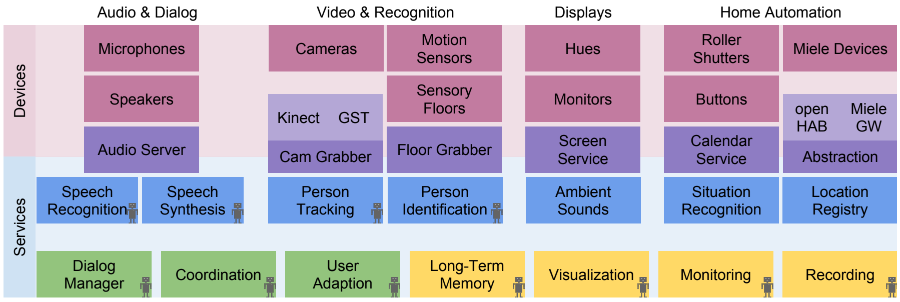


Figure 7.4: Functional system overview of the CSRA [Wre+17]. Abstraction layers are depicted vertically, starting with exemplary hardware devices at the top, through the corresponding grabbing services to behavioral components at the bottom.

The social robot head of the *Floka* is a subsequent version of the *Flobi* head. These agents have the required expressive and sensory capabilities for the *HAI* I investigate. Furthermore, additional information can be obtained from the apartment and its sensors.

7.2 GENERAL SOFTWARE ARCHITECTURE

In the following, the general architecture of the *CSRA* is presented. Of special interest are the abstraction layers of the apartment, the used middleware, and the continues integration approach.

7.2.1 *Abstraction Layers of the CSRA*

We designed the *CSRA* architecture with the abstraction layers depicted in fig. 7.4, in which we distinguish between *devices* and *services*. In the following, I explain this architecture further with dialogue relevant examples[1].

SENSORS AND ACTUATORS are depicted at the top of Figure 7.4 (in pink). Exemplary *hardware devices* are microphones and speakers.

HARDWARE ABSTRACTION LAYER are depicted in purple and consists of components which handle raw data, e.g., the *Cam Grabber* or the *Audio Server*. They form the bridge between hardware devices and software services.

SUBSEQUENT PROCESSING COMPONENTS are depicted in blue. They do not necessarily need to take place on the computer which is connected to the hardware device. An example is the *ASR*, which reach its input from the *Audio Server*. Including with the hardware abstraction layer, they form the *base services*, which do not generate apartment behavior.

1 For more information about our decisions regarding the software architecture of the CSRA and the need for such distinctions within the apartment, see [Wre+17].

BEHAVIORAL SERVICES are depicted at the bottom of Figure 7.4 (in green). They can produce behavioral actions within the apartment. An example is the *DM*.

CROSS-CUTTING SERVICES are depicted in yellow and are connected to all layers, e.g., *Recording* video and system events.

The configuration and organization of smart devices and their functionalities is performed using the smart environment framework *Base Cube One (BCO)*, a location-addressable and service-oriented architecture [PLH19]. I implemented several general apartment *base services*, such as the *Speech Recognition* and *Speech Synthesis*. Furthermore, I integrated behavioral services, e.g., *Pamini* as *Dialogue Manager*. These components are discussed in more detail in section 7.3.

7.2.2 *Middleware and Interfaces*

A message-orientated and event-driven middleware is chosen to be able to build this modular architecture and support the communication between these services. The *Robotics Service Bus (RSB)* [WW11] allows the implementation of services in the programming languages, Java, Python, Common Lisp, and C++ and supports the communication between different compute machines. To be able to interact with the other components, each *service* of the dialogue system needs an integration into the apartment middleware. Furthermore, each module can be exchanged easily because of well-defined interfaces. Each interface is defined declaratively, as a data type, specified using an *Interface Definition Language (IDL)*. To increase the usability, I added all data types concerning the dialogue system to the *Robotics Systems Types Repository (RST)* [Bie19a].

7.2.3 *Continuous Integration*

We developed this reusable service architecture in a continuous development and integration process to cope with the complexity of the *CSRA* project and improve its reproducibility. To this end, we used the *Cognitive Interaction Toolkit (CITK)* [LWW14] and I integrated all dialogue components into the apartment distribution.

Using the architecture we developed in the *CSRA*, I could fulfil several requirements concerning dialogue architectures (cf. section 6.2.1 and section 6.2.1). The concept of services and application of well-defined of interfaces support the *modularization* of the dialogue system. This allows *platform* and *scenario independence*, and makes it possible to easily exchange single components without (re-)implementing the system from scratch.

7.3 DIALOGUE MODULES

In the following, I shortly discuss the different parts of the dialogue system with its services and interfaces. All services regarding the dialogue system are available at the interaction islands, as well, as on the robot and can be configured for a specific scenario within the smart environment. Furthermore, I identified two key concepts: (1) the use of interaction patterns with system task descriptions for rapid prototyping of interaction scenarios and (2) the concept of the *IU* model to deal with the incremental nature of human dialogue. Consequently, I combine the two toolkits *Pamini* and *inprotk*, which implement these concepts. In addition, the integration features for the attention concept and the hesitation intervention strategy are presented.

Figure 7.5 visualizes an example system overview of my dialogue system, which is divided into four layers: (1) sensors and actuators, (2) hardware abstraction layer, (3) further processing modules (4) decision management. Grabbing and pre-processing units pass their information to the attention-hesitation module, the *DM*, and the scenario coordination. These in turn communicate with the *TTS* and highlight service, as well as hardware abstraction for the home automation (*BCO*) ant the robot control (*High Level Robot Control (HLRC)*). It should be noted, that this communication is bidirectional and (as mentioned in chapter 6) the attention management can alternatively be modeled as a submodule of the dialogue management component, due to the close interplay of these components. The different parts of the dialogue system are presented in the following.

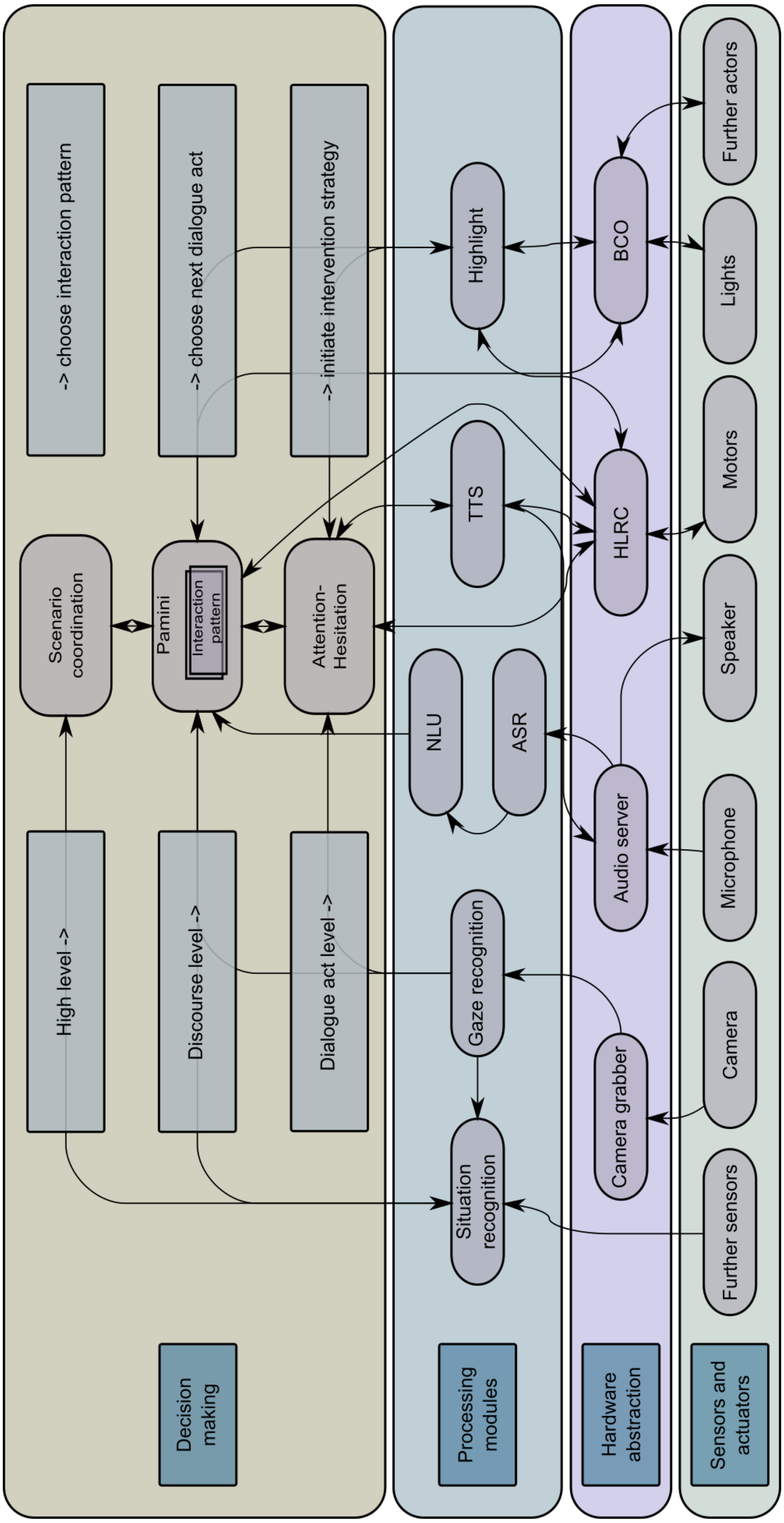


Figure 7.5: Example speech system overview. At the button, possible sensors are depicted, followed by the hardware abstraction layer, and further processing modules. The decision management is divided into three levels.

7.3.1 *Speech Recognition and Understanding*

SPEECH RECOGNITION SERVICE: For *ASR*, the incremental speech recognition module of *inprotk* is used, extended with *RSB* interfaces. To be able to grab the data from different microphones and be independent from the actual hardware configuration, between the following configurations can be chosen. It is a possibility to grab directly from a microphone, which allows fast processing. The second option is an *RSB* input using the *RST* SoundChunk interface (see listing A.1 in appendix A). This allows the speech recognition to be run independently of the microphone grabbing. The *ASR* component in *inprotk* produces wordIUs (cf. section 6.1.4). It supports different speech recognition tools, such as Sphinx [Lam+03], Google ASR [Clo19], or Kaldi [Pov+11]. The service produces either a single speech hypothesis depicted in listing A.3 in appendix A, or multiple speech hypotheses (see listing A.2) if the chosen *ASR* provides n-best results. Because of privacy concerns with cloud services, we mainly use Sphinx or Kaldi, which both working offline and are suitable for incremental systems [Bau+17]. For simple smart home commands, such as "switch the light on" or "Which new technologies are in the apartment?", Sphinx is used with a JSGF grammar [Sun10].

GAZE RECOGNITION SERVICE: To create a first model of attention, I integrated a gaze detector into the system. To this end, the gaze detector by Schillingmann and Nagai is used [SN15]. Therefore, a *RSB* in- and output was implemented. This gaze detector is based on dlib [Kin09], a C++ toolkit for machine learning and can estimate mutual gaze with 96% accuracy at 8° tolerance and one-meter distance to the agent based on head pose features, eye region HOG features and noise robust pupil detection. Furthermore, it can estimate gaze and detect mutual gaze even in low-resolution images of 640×480 pixels, with a frame rate of 30 fps and a latency of about 80ms, which makes the system suitable for real-time processing. [SN15]. Figure 7.6 depicts the visualization of the gaze detector estimating mutual gaze. In contrast to a head detector—often used to estimate the current *Visual Focus of Attention (VFoA)*—it uses additionally eye region HOG features and pupil detection. This allows to provide information about the detected face as well as the horizontal and vertical angle of its gaze using a camera, next to the head of each agent (see listing A.8). In doing so, no special hardware need to be wear which could influence the interaction. In addition, the detector recognize no mutual gaze even if the head is pointed to the agent while looking to the side, as depicted in fig. 7.7 (upper). Furthermore, the horizontal and vertical angle of its gaze is roughly classified using configurable thresholds for the angles (MUTUAL_GAZE, LEFT, RIGHT, UP, DOWN, UNKNOWN). The fact that humans not always

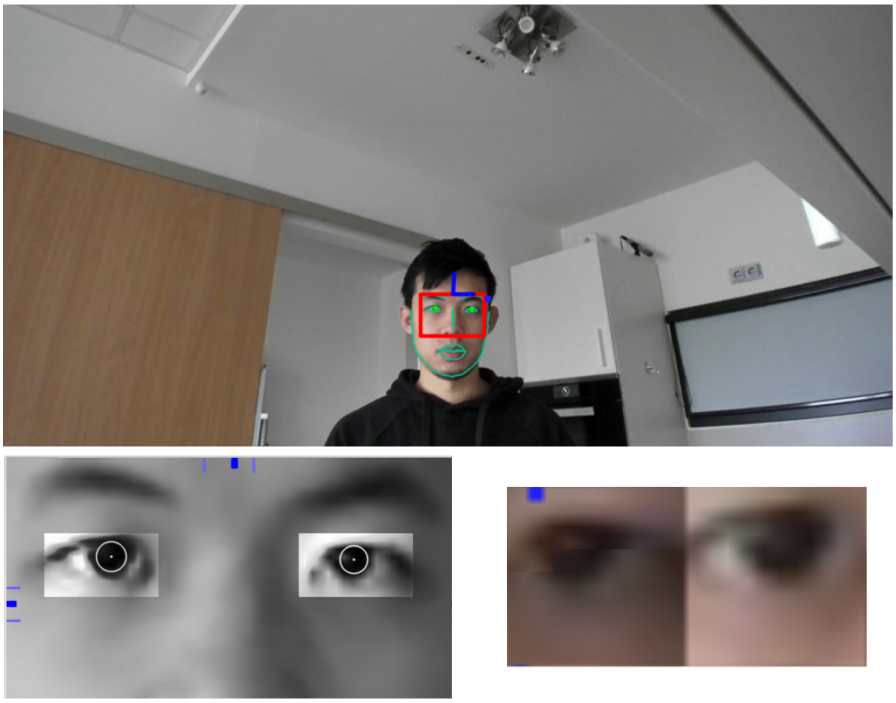


Figure 7.6: Visualization of the gaze detector, estimating mutual gaze.

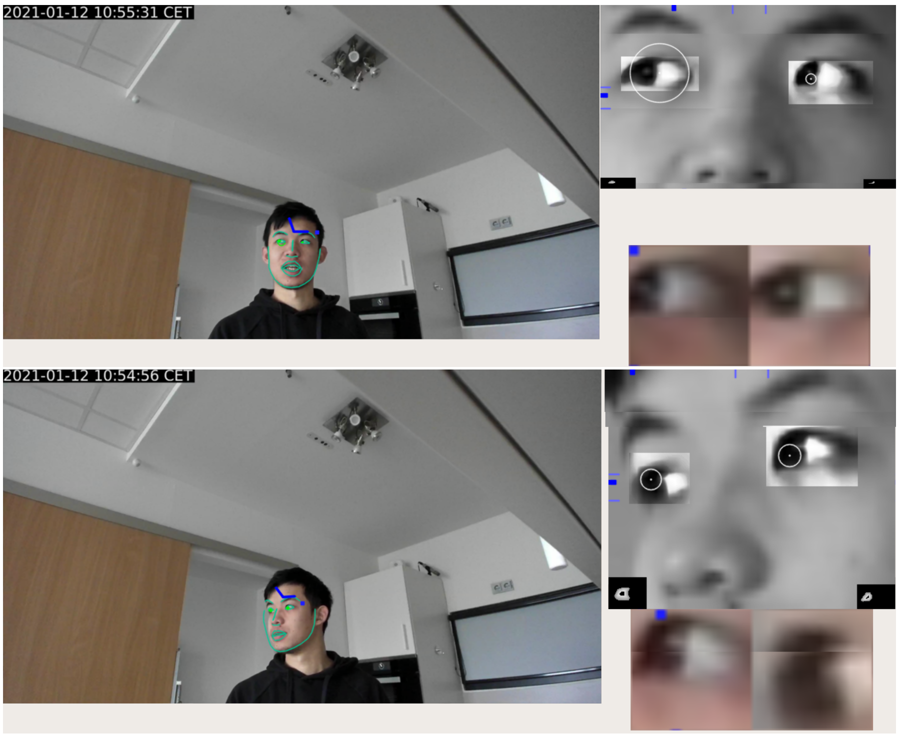


Figure 7.7: Visualization of the gaze detector estimating no mutual gaze: (Upper) the human's head is pointed to the camera while looking to the side.(Lower) the human looks to the side.

fixate the other's eyes, but rather different points in the face of their interaction partner is addressed by classifying gazes at the whole display at MUTUAL˙GAZE. Furthermore, short gaze aversions (less than 1s) are ignored.

OTHER RECOGNITION SERVICES: The apartment provides further recognition services, which outputs can be used in the dialogue systems. A person tracking system provides person hypotheses about the current location of users in the apartment. The *situation recognition* delivers information about current events, e.g., is someone cooking or is there a demonstration. In addition, another service recognizes pointing gestures.

SPEECH UNDERSTANDING SERVICE: The *DM Pamini* needs a representation of a dialogue act. Based on the underlying *ASR* framework, I implemented a simple *NLU* module, which produces either a simple speech hypothesis or a high-level dialogue act (see listing A.4 in appendix A). I applied two techniques. The first one, is grammar-based and parses single words with the corresponding grammar. This is necessary because even though Sphinx can be used with a grammar, the result contains no information about the matched grammar tree. Based on the grammar, the *NLU* produces a dialogue act and the corresponding grammar tree. The second approach uses an open-source machine learning framework for intent classification and entity extraction [NLU19]. Listing A.2 visualizes these interfaces, both relying on the *ASR* results. The speech recognition module in *inprotk* produces incremental results if possible. To preserve this feature, the concept of *IU* is realized in the later modules as well as in the design of interfaces. The differentiation between speech hypothesis and dialogue act is mandatory, to integrate other modalities. This allows, for example, to easily integrate a component which incorporate the results of a head nodding recognition, which can also be interpreted as the dialogue act H.confirm or a hand waving as H.greeting. Furthermore, other modalities can be used to add further information to the dialogue act. In the apartment we use, for example, pointing gestures to disambiguate simple commands such as "switch this light on".

7.3.2 *Decision Management—the Dialogue Manager*

The following dialogue depicts an exemplary interaction within the apartment.

Example 7.3.1: A dialogue between the agent (A) and a human (H) in the CSRA: The agent welcomes the visitors and becomes acquainted with them.

Interaction pattern		Action
		A person enters the apartment for the first time and looks at the agent Flobi.
System-initiative task to learn the face:	A:	I have never seen you before. I would like to learn your face, please enter your name and look into my camera, while I'm learning your face.
		The person types "John" on a virtual keyboard on the screen, and Flobi learns the person's face.
	A:	Ok, thank you, John, now I know you.
Human-initiative information request:	H:	Can you tell me something about you?
	A:	Yes, sure. I'm Flobi, the virtual contact person here in the apartment. I'm able to see you, hear you and with my help, you can adjust several settings within the apartment. In the living room is Floka, it can bring you something to drink.
		Flobi looks at the Floka robot.

I distinguish between the following levels of decision-making in the coordination of the *HAI*.

HIGH LEVEL: the overall goal/plan of the agent or respectively the *CSRA*

DISCOURSE LEVEL: interaction managed based on conversational acts

DIALOGUE ACT LEVEL: adaptions or repair strategies during a conversational act

The high-level decision-making depends on the overall agenda or plan of the agent. Example 7.3.1 depicts a short interaction example within the apartment. Here, the goal is to welcome new visitors and familiarize with them. Within the apartment structure, the *Coordination* component [Uni19] activates these goals by triggering the corresponding *Interaction Pattern (IP)* or a sequence of *IPs* from the *DM* (see section 7.3.2). On the discourse level, each *IP* is responsible for the local decision management and selection of the next dialogue act. For the dialogue act level, an additional module is needed (see section 7.3.2).

Integration of Pamini

As *DM* component *Pamini* (see section 6.1.3) is used and extended with *RSB* support and interfaces. In a first step towards integration of incremental dialogue processing, I addressed the following issues. First, for the communication with *inprotk*, *Pamini* has to distinguish between the different states of the incoming dialogue acts (i.e. added, revoked, updated, committed). Second, to allow the user to interrupt the agent (or to allow self-interruptions), system components (pertaining to both, verbal and non-verbal actions) need to be interruptible during execution. I addressed the first issue by extending *Pamini* to react on different states of dialogue acts. *Pamini* is based on two main concepts, which allow the fast implementation of asynchronous interaction scenarios—the concept of generic *IP* and a back-end communication via the *Task State Protocol (TSP)*. Through the *IPs*—a generalized description of the structure of *HAI*—concrete scenarios can be configured easily. They have, furthermore, been targeted to deal with asynchronous events of the agent's internal processes which can invoke specific dialogue actions. *Pamini* handles these asynchronous internal system processing events and thus provides an interface to the back-end processing modules through the *TSP*.

INTERACTION PATTERN: Let's take a closer look at the concept of interaction patterns.

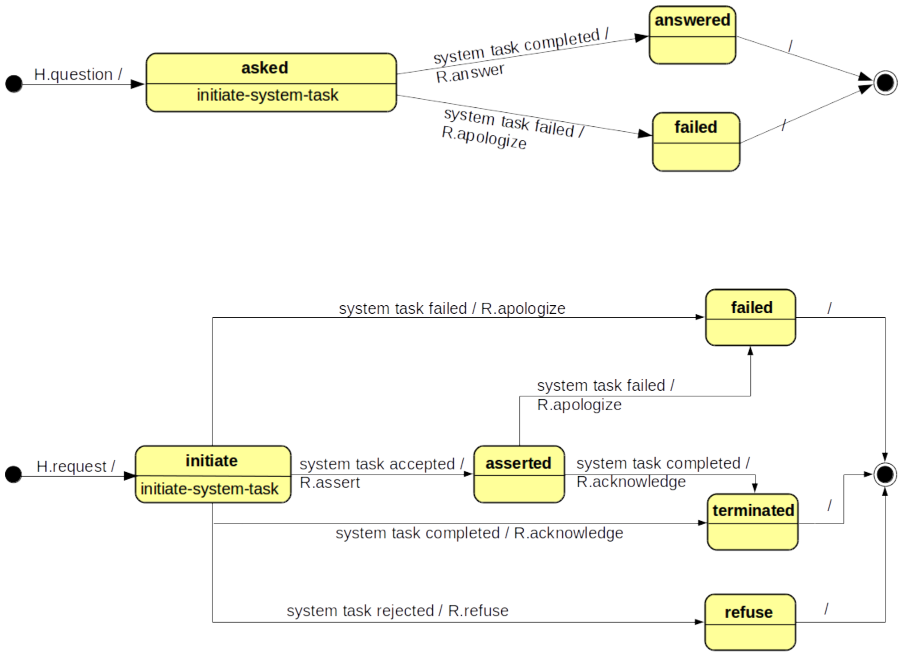


Figure 7.8: Two examples of interaction patterns: (Upper) *Human Simple Question*, (Lower)*Human Simple Action Request*.

Example 7.3.2: A dialogue between the agent (A) and a human (H) in the CSRA.

		The person enters the living room before it want to leave.
Human-initiative greeting:	A:	Hello Floka.
		Floka greets back by waving it's hand.
Human-initiative information request:	H:	How is the weather outside?
	A:	It's raining.
Human-initiative system task:	H:	Give me my umbrella, please.
	A:	Sure.
		Floka drives to the wardrobe and fetches it.
	A:	Here is your umbrella.

Figure 7.8 visualizes two different interaction patterns with corresponding example interactions on the right. The upper pattern depicts a *Human Simple Question*, whereas the lower represent a *Human Simple Action Request*. *Pamini* distinguishes two types of system actions. Example 7.3.2 depicts a dialogue for each interaction pattern. The first

type is a communicative dialogue act, for instance the *R.answare* in the *Human Simple Question* interaction pattern. Here. the human asks a simple question like "How is the weather outside?", which results in the answer "It's raining.". Other system actions need a longer period of time, especially in robotic systems. The second pattern in fig. 7.8 visualizes such an action: after the human requests "Give me my umbrella, please." the robot started with the system action *grasping* and gives incremental feedback during this action, using the concept of tasks. Those system tasks can receive an update or a cancel request, which is in line with the incremental processing requirement. It is possible to start such a task (e.g., while the human is still speaking) and update or cancel it if needed (e.g., if the corresponding dialogue act is reverted). The communicative dialogue acts are not represented in this way in *Pamini*. Here, a simple "fire-and-forget" approach is used, although, e.g., in the interaction example 7.3.1, the agent's answer could be considerably long. For more information about the concept of interaction pattern, see [PW10a].

ALLOWING INCREMENTALLY: In the original implementation, *Pamini* only reacts to input which is in the 'committed' state. To this end, I created a new input source in *Pamini* that uses dialogue acts of *inprotk*. When an incoming dialogue act is committed, the input can be processed by *Pamini* as usual. Otherwise, a more careful processing strategy is needed. Depending on the input, some output may be produced. However, we have to consider that the dialogue act may be revoked when the dialogue act generation module changes its hypothesis based on new input. Therefore, I allow this currently only for easily revertible actions, such as switching the status of the lights within the apartment. Achieving interruptability on the output side is even more difficult.

The interaction patterns are in a dialogue state before or after a communicative dialogue act, the time period in which the act is performed is neglected. Adding a state for each robotic dialogue act, to solve this problem, would blow up each interaction pattern. I decided to implement a new *Attention Module* that acts as an additional dialogue policy, which observes the current dialogue state: the attention module.

Attention-Hesitation Module

It is possible to integrate the concept of attention and hesitations into this dialogue management system using different approaches. I list them here briefly, starting with the concept of attention. The implementations of the individual hesitation strategies, as well as the concrete concept of attention are presented in the evaluation of the corresponding evaluation of the *AHM* in chapter 10. The *Attention-Hesitation* Module receives the following information from the dialogue management:

- Currently active *interaction patterns*
- Current state of active *interaction patterns*
- Speaking state of the human
- Speaking state of the system
- *FoD* of the current speech phrase
- Discourse history.

Furthermore, it observes the current output, the current verbalization and other system actions and make it possible to pause or stop the current action. If a hesitation is needed, a corresponding command is sending to *inprotk*. This could result—corresponding to the command—in a hesitation action or a stopping of the current speech act.

ATTENTION FEATURE INTEGRATION: As a first feature to integrate the interlocutor's attention, the gaze hypothesis by the gaze recognition service is used to estimate *mutual gaze*. This information can be utilized throughout the phases of the interaction in the dialogue system. In the establishing phase, to decide whether the person wants to interact with the agent and in the maintaining phase in the *Attention-Hesitation* module, to figure out if a re-attention or highlight strategy is needed. However, to distinguish further, between engagement and understanding problems, an additional task related information is needed—the *FoD*. To recognize *directed gaze*, the *FoD* of the current speech phrase of the dialogue act is compared with the results of gaze estimation. Therefore, the *FoD* need to be configured in the pattern configuration(more information can be found in section 10.2.2). Through the *location registry* of the *CSRA*, the position of objects (the current *FoD*) can be compared with the *VFoA*. Based on the *history of the discourse*, the module differentiates between an ongoing *FoD* and a *discourse change*. These features can be integrated without further information—beside the *FoD*—of the current scenario and the task at hand. To integrate the *task progress* itself as a feature, additional information is required. Additional identification services are required for this, which of course depend on the specific task description. This will be discussed in section 10.3, in which the interaction includes a practical task in the interaction.

HESITATION INTEGRATION: The integration of hesitations affects different components. At first, the synthesis module needs to be cable of producing hesitations. The synthesis module of *inprotk* can already pause after an ongoing word or phoneme. This allows the integration of the *unfilled pause* as a feature for the hesitation strategy. However, this behavior needs to be initiated and terminated by the coordinating component. More complex hesitations, such as producing a hesitation vowel or a repetition, need an adaptation of the ongoing speech plan and are currently not part of the inprotk-system. To make this possible, I implemented an additional *IU* processing module, which serves as a bridge between the attention module and the synthesis module. It receives its input from the dialogue management and passes the text to speak to the synthesis module. As soon as a hesitation strategy starts it change the left buffer of the synthesis module for, e.g., initiating the lengthening. Further implementation details are presented in the evaluation of the corresponding strategy (see chapter 10). In addition, other modalities are integrated into the strategy, e.g., a repetition of the highlighting of a current *FoD*. Therefore, other modules, such as gaze or pointing module of the agent, may need to be informed as well.

7.3.3 *Speech Output and Other Actors*

TEXT-TO-SPEECH SERVICE: The incremental synthesis module of *inprotk* is used as the *TTS* module. The underlying synthesis tool is Marry TTS [ST03]. In analogy to the *ASR*, it can produce different outputs, like streaming the data directly to a loudspeaker or publishing it via *RSB* using the *RST* Sound Chunk interface. For the input of this component, I use the concept of a system task with the additional play back options, pause and resume (see listing A.7 in appendix D). The pause can be further configured, in terms of what kind of hesitation strategy should be performed. Additionally, I implemented a feedback mechanism to inform about the current state of the task and the produced phonemes to coordinate lip movements of the agent.

HIGHLIGHT SERVICE: We designed a special highlight services which should be able to guide the attention of the user to a specific location. The attention module provides the following possibilities to highlight the current *FoD*:

GAZE: Use gaze to refer to the target that should be highlighted. Flobi is able to look at specific directions. To this end, it not only moves its eyes, but also turns it head in the desired direction. Therefore, the *humotion framework* is used [Sch+16].

GESTURE: Use gestures to refer to the target, e.g., pointing. Even though Flobi has no arms to point at specific objects, it can use its head to refer to one direction. The Meka as also able to paint with its arms and hands.

LIGHT: Use (ambient) light at the target, e.g., LEDs or surrounding lights. The fact that Flobi is connected to the smart environment allows developing and investigating interaction concepts beside human-like interfaces. One of these is the usage of lights, whether as ambient background light or by directly highlighting specific areas of the smart apartment. Another option is to use a spot(-light) that points to the target.

SOUND: Use (ambient) sound at the target, e.g., beeps.

Furthermore, the highlight module is integrated into the output possibilities of *Pamini*.

Listing 7.3.1: Example configuratuion for the answere of the question "Where can I find the spoon?".

```
robotDialogAct state="asked" type="R.answer" fod="cutleryDrawer"

  verbalization text="Da in der Schublade."
  attention targetId="cutleryDrawer" modality="GAZE" duration="3000000"
  attention targetId="cutleryDrawer" modality="AMBIENT_LIGHT" duration="5000000"
```

The listing 7.3.1 depicts an example configuration of the agent's answer of the question "Where can I find the spoon?". In addition to the verbalization "There in the drawer.", the agent also gazes at the specific cutlery drawer in the kitchen for 3 seconds. Furthermore, the corresponding the handle of the cabinet door lights up for 5 seconds. Note, that the targetId is the current *FoD* of the dialogue act, but can be overridden (or specified more precisely). The location of the object with the id cutleryDrawer can be found in the location registry of the apartment.

OTHER ACTOR SERVICES: Besides speech and highlight output, the apartment allows other possibilities of expression. The agents have several actor services, such as showing various facial expressions or looking at special points of interest. Furthermore, the apartment has several actuators, which can be manipulated by other components, such as lights, displays, or even smart kitchen devices. These actions can be triggered via a system task (see listing A.6 in appendix A).

7.3.4 *Introspection Capabilities*

Introspection capabilities, especially at run time, are relevant for the implementation and testing of interaction scenarios. The *RSB* middleware already provides comprehensive introspection capabilities, command line tools to query the current configuration of the middleware, communication participants and the published events [Bie19b]. Besides this, I create an extra visualization to inspect dialogue relevant *RSB* communication. This visualization can be started at any time during the interaction and displays the main results of the processing modules.

7.4 MEETING THE REQUIREMENTS

In chapter 6, general concepts of the design of human-agent speech-based dialogue are presented and requirements for the design of dialogue systems as fundamentals for an autonomous *HAI* are extracted. In this chapter, it is shortly discussed how the technical realization of a dialogue system—which allows further investigation of my research question—meets these requirements for both the hardware (section 7.4.1) and the software (section 7.4.2).

7.4.1 *Hardware Requirements*

Section 6.2.2 presents three hardware requirements for the implementation and evaluation of the *AHM*: (1) the possibilities of expression (2) possibilities of interaction partners perception and (3) the naturalness of the environment.

The possibilities of expression are mandatory to establish joint or shared attention. This is achievable through the agent's gazing behavior. The social agents Flobi and Floka (see section 7.1.2) have both comprehensive expression capabilities, via eye gaze and head behavior and the possibilities of showing facial expressions. In addition, the *CSRA* itself (see section 7.1.1) is equipped with additional actors, such as lights, displays, and additional boxes. Furthermore, several possibilities of interaction partners perception are given. In contrast to commercial *SPAs*, the agents in the *CSRA* can use the various sensors in the smart home, ranging from pointing gestures to disambiguate speech to gaze hypothesis for monitoring the visual attention of the user. Through the use of a gaze detector based on VGA images, it is not necessary to wear eye tracking glasses. The benefit of more fine-grained results does not outweigh the disadvantage of wearing an extra tracking system. In addition, the *CSRA* offers the possibility of a more natural environment because it is furnished like a real apartment. However, through the observation possibilities, the *CSRA* provides the right environment for the evaluation of my model in human-agent interaction studies in a smart environment.

To sum up, the *CSRA* with its conversational agents fulfill the hardware requirements for investigating the effect of my *AHM* in a *HAI* in a smart environment.

7.4.2 *Software Requirements*

Apart from the hardware requirements, several software requirements my model pose to the design of dialogue systems and the fundamentals for an autonomous *HAI* are collected in section 6.2.1: (1) multi-modality, (2) incremental processing, (3) topology, and the (4) generalizability of the system.

Multi-modality is achieved by using various services of the apartment. On the input side, services such as the *Speech Recognition*, *Face Recognition*, or *Pointing Recognition* provide information which can be further processed, e.g., in the attention module. On the output side, the generation of multimodal dialogue acts can be configured, including, e.g., verbal output with facial expressions or head animations. In addition, further multimodal system actions can be triggered using the task state interface.

Explicit consideration of the incremental nature of dialogue processing is achieved in two ways: (1) incremental processing capabilities are given by the speech recognition and speech synthesis modules by *inprotk* through the concept of *IUs* and (2) the concept of tasks allows the interruption of system actions. This is achieved through the combination of the two toolkit for dialogue modelation *inprotk* and *Pamini*.

Furthermore, by using the concept of *services* within the *CSRA*, the dialogue system is modular. In addition, I defined interfaces to allow an exchange of single components. Based on the previous requirements, the topology of the dialogue systems is organized in layers. Through the used middleware, each component can observe or request information from other components. In addition, through the concept of *IUs*, the dialogue system can depict various topologies and is a network rather a single pipeline (see section 6.1.4).

Generalizability is achieved through the concept of generic interaction patterns of *Pamini*. In combination with a generic protocol for task representation, rapid prototyping and reusability is supported. Furthermore, the integration of the attention concept as well as the possibility to use hesitation as intervention strategy are enabled in the *DM* by integrating a separate attention module. In the following, different interaction scenarios are outlined, using the presented dialogue system.

7.5 DIALOGUE INTERACTION SCENARIOS

The presented dialogue system is used in several interactions scenarios and research studies. In the following, a short overview is given.

7.5.1 *Interaction Zones within the CSRA*

The dialogue system is deployed in the *CSRA*. The apartment has two permanent human-agent interaction zones: one in the entrance area and another one in the kitchen. In both interaction zones, the simulation of the Flobi head is the interaction partner. In the entrance area (see fig. 7.9), Flobi welcomes visitors and gives general information, e.g., about itself, the apartment, or its current development. In the kitchen (see fig. 7.10), Flobi assists visitors, e.g., during cooking

Figure 7.9: Interaction zone in the entrance area of the *CSRA*: Flobi welcomes visitors, learn their faces and give information about itself and the intelligent environment.

or helps to find the right place for kitchen utensils. Additionally, it

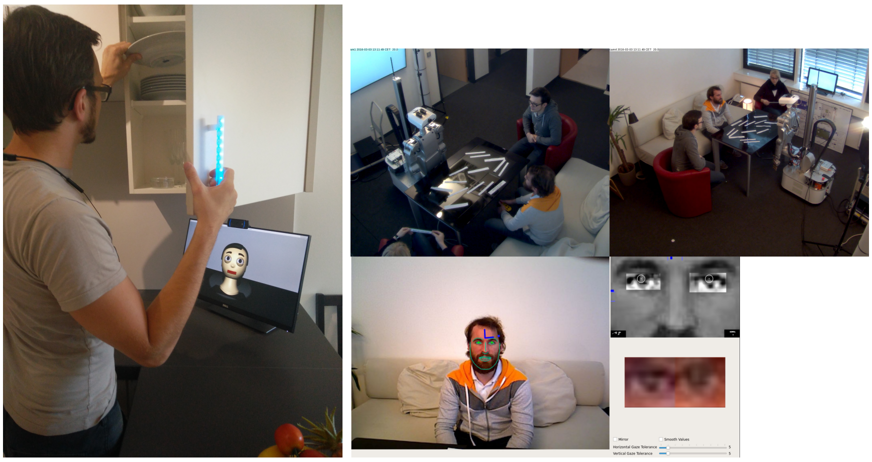


Figure 7.10: (left) Interaction zone in the kitchen of the *CSRA*: Flobi assists visitors, e.g., during cooking or helps to find the right place for kitchen utensils. (right) Interaction zone of the Floka robot: the robot answers simple questions and changes the state of the light within the apartment.

can give general information about itself and the apartment in this interaction zone. Furthermore, each agent can answer questions (e.g., *what time is it?*) or to solve simple tasks within the smart home, e.g., turning on the light in the apartment. Both instances utilize a webcam on top of the monitor. Through these cameras, the simulated Flobis can detect faces in front of them and focus them, thus demonstrating responsiveness and establishing shared attention. Furthermore, a microphone near each interaction zone and speakers are utilized as well. The third interaction 'zone' is the mobile Floka-robot itself. It has two microphones installed in his upper body. Depending on the configuration of the head, different cameras are utilized: two USB3 Ximea cameras (social head) or a Primesense short-range RGBD-sensor (sensor head). Further microphones and speakers throughout the apartment can be used for other interaction zones.

In order to explore intuitive verbal and non-verbal interfaces in smart environments, we recorded 63 user interactions in the *CSRA*. The resulting multi-modal corpus contains goal-directed actions of naive users in attempts to solve a number of predefined tasks with the apartment [Hol+16]. On this corpus, we explored which interfaces participants would intuitively and most frequently address. We found out that participants preferred physical interfaces whenever the task allowed to, and most participants used speech to control the smart home environment [Ber+16].

7.5.2 *Simple Service Robot Interaction Scenario*

In a scenario with the Floka robot, we addressed if it is possible to guide the attention of the robot towards a specific interaction partner in a multi-party interaction [Ric+16]. The implemented dialogue consists of a set of simple questions and tasks for the service robot Meka. The human can one of the following information or action requests:

- turn on/off the light in the apartment
- ask for the current time
- ask for a missed call
- ask for delivery
- request about possible ongoing experiments
- request which data is getting recorded
- ask for more information about the Zen-garden in the apartment.

Additionally, a greeting of the robot is possible. In addition to the dialogue system, a gaze management integrated [Fac19]. This guides the robot's gaze to the most interesting point, based on hierarchical prioritization of multi-modal sensor input streams (for more information see [Ric+16]). In addition to this button-up attention management for the robotics gaze, a top-down override, e.g., from the dialogue management component is possible.

We conducted an interaction study with the autonomous system to evaluate the different aspects, (i) the integration of human eye-gaze, i.e., the evaluation of mutual gaze as cue for addressee recognition and (ii) the possibility to guide the attention of the robot towards a specific interaction partner in a multi-party interaction. The study has been carried out with German native speakers. In total, we recorded approximately 53 minutes of interaction in 5 trials with 2 female and 13 male participants. A typical trial takes approximately 10 minutes. Altogether the dialogue system detected 874 human dialogue acts, 152 of these would have triggered a verbal response or a corresponding system action (light on/off). To evaluate the means of different approaches to addressee recognition, a ground truth annotation was carried out for each dialogue act. We found that it is possible to achieve mutual gaze (even the robot is inattentive in the first place) using a bottom-up management system for the robotic gaze. Furthermore, the integration of gaze information about the human interaction partner improve the dialogue in terms of reducing false-positive reaction, but more sophisticated addressee recognition would be useful for further improvements [Ric+16]. The approach of integrating attention into the dialogue system in the initial phase of the interaction and the consequences for conversational role detection is further investigated by Richter [Ric20].

7.5.3 *Further Interaction Scenarios without Agents*

Besides this human-agent interactions, the dialogue system is used in some scenarios outside the *CSRA*. In the innovation cluster *KogniHome*, several demonstrators are developed which address the topic of technology-assisted living for people. The presented dialogue system was used in the various demonstrators, such as the *KogniMirror* or the *KogniChef*. Both demonstrators assist in daily tasks without an explicit embodiment of an agent. The *KogniChef*, depict in fig. 7.11, is a

Figure 7.11: The cognitive cooking assistive system *KogniChef*.

cognitive cooking assistive system that provides users with interactive, multi-modal and intuitive assistance while preparing a meal [Neu+17]. The *KogniMirror* is an intelligent mirror that can provide support for completing daily activities, e.g., it provides the user step-by-step instructions to help them tie a necktie correctly.

This overview of various interactions—using the whole or parts of my dialogue system presented in chapter 7—shows that the dialogue system can deal with different scenarios. Furthermore, it demonstrates that the dialogue system is platform-independent.

8 SUMMARY OF PART II

In the second part of this thesis (*Fundamentals for Autonomous HAI*), I investigated RQ 2. After a brief introduction into dialogue modeling—particularly regarding the dialogue management component—in chapter 6, I illustrated the requirements posed by the technical realization of my *AHM* (see section 6.2). Furthermore, I constituted the choice of the research platform (section 7.1) and the technical realization of a dialogue system (chapter 7), which allows further investigation of my research question. I identified two main concepts for dialogue modeling (1) the use of interaction patterns with system task descriptions to allow rapid prototyping of interaction scenarios and generalizability, and (2) the concept of the *IU* model to deal with the incremental nature of human dialogue. Through the combination of the toolkits *Pamini* and *inprotk* both concepts are considered in my dialogue system. In addition, the attention-hesitation module is integrated into the dialogue management to coordinate the interaction on the dialogue act level (see section 7.3.2). To show the applicability of my dialogue system, I presented scenarios and research studies, which used this dialogue system, including a first integration of human gaze.

The presented dialogue system meets the requirements (see section 7.4.1) to the general architecture in speech-based systems and is the fundamental work for autonomous *HAI*, and the investigation of the effect of my *AHM* in interaction.

Part III

LEARNING FROM EXPERIMENTS

9 EVALUATION METHOD AND HYPOTHESIS

In the first part of this thesis, the *AHM* based on the insights gained from *HHI* research studies (chapter 2) and literature on the topics of attention and hesitations in *HAI* (chapter 3) is developed. Additionally, in the last part, the requirements posed by the technical realisation of my model are posed. Furthermore, design decisions concerning the used platform and the general software architecture, and the implementation of specific software components for modeling *HAI* were depicted. In addition, design decisions for the integration of attention and hesitations were presented.

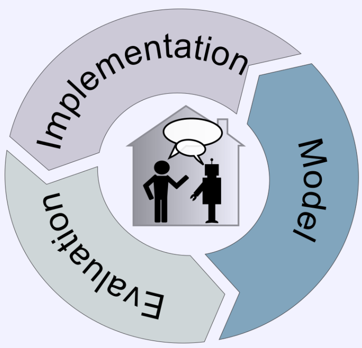


Figure 9.1: Three aspects of the research procedure to consider for answering the research question.

1. **Model**: A theoretical model based on results from *HHI* research.
2. **Implementation**: The implementation of a scenario which performs as autonomous as possible.
3. **Evaluation**: The evaluation in a real *HAI* study.

This part focuses on the last aspect of fig. 9.1—the evaluation of my model and the research hypothesis. Therefore, possible ways to evaluate *HAI* and their benefits and disadvantages are discussed in section 9.1. Afterwards, the chosen evaluation method of *Evaluation Cycles (ECs)* is presented in section 9.2, including an overview of the cycles to evaluate the effect of my *AHM* on the interaction.

9.1 EVALUATION OF DIALOGUE SYSTEMS

A persistent topic in the *Human-Agent Interaction (HAI)* community is the evaluation of the interaction between an agent and the human. There are several possible ways to evaluate interaction, which optimize towards different targets.

One way is to evaluate the dialogue system itself. Hung et al. provide an overview of different metrics and make a distinction between objective and subjective metrics [Hun+09]. One commonly used objective metric is time, e.g., the dialogue time or task completion time or the mean user/system respond time. It measures the efficiency of the dialogue system. Although, it might seem understandable to minimize the overall dialogue time—to make interaction more efficient—this does not tell anything about the quality of the interaction. A dialogue manager which is only optimized for efficiency regarding time will try to interact as little as possible. However, a longer interaction may indicate a more engaged interaction partner, which is often desirable. In contrast to the overall time, the system response time is a more reasonable metric for a dialogue system.

A similar approach, but independent of the exact execution, e.g., speaking rate, is the evaluation of the total number of user and system turns. Like the overall response time, it is only suitable as a criterion for optimization under some restrictions. While these metrics evaluate the efficiency of dialogue systems, other objective metrics directly asses the quality of the dialogue system. Typical examples are the number of user barge-ins or re-prompts. These metrics reflect—among other things—errors in the turn-taking behavior of the system. The general understanding of the system is measured as concept accuracy or inappropriate system responses.

Beside these objective measurements, Hung et al. present several subjective measures, e.g., the percentage of correct answers or contextually appropriate system utterances [Hun+09]. In general, the user satisfaction is an important consideration. It can be reported in questionnaires on one of the following topics: ease of usage, clarity, naturalness, friendliness, robustness regarding misunderstandings, or willingness to use the system again.

In the *Human-Robot Interaction (HRI)* community, several standardized questionnaires to assess subjective metrics exist. Regularly, the *Negative Attitudes Toward Robots Scale (NARS)*, developed by Nomura et al. [Nom+08] is used. This questionnaire is based on psychological scales and measures people's anxiety towards robots and the change in participants' attitude towards robots in long-term interactions. Another questionnaire—developed by Bartneck et al.—the *Godspeed Questionnaire Series (GQS)* measures five key concepts in *HRI*: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. The authors report reliability and validity indicators based on several empirical studies for each concept [Bar+09]. The idea behind the *GQS* is to have a standardized measurement tool for *HRI* and to be comparable between different robots and user studies. Weiss and Bartneck present a meta analysis of the godspeed questionnaire and conclude that the measurement of the five key concepts is relevant for the evaluation of social human-robot interaction [WB15].

Walker et al. postulate the PARADISE (PARAdigm for DIalogue System Evaluation) framework for the evaluation of spoken dialogue

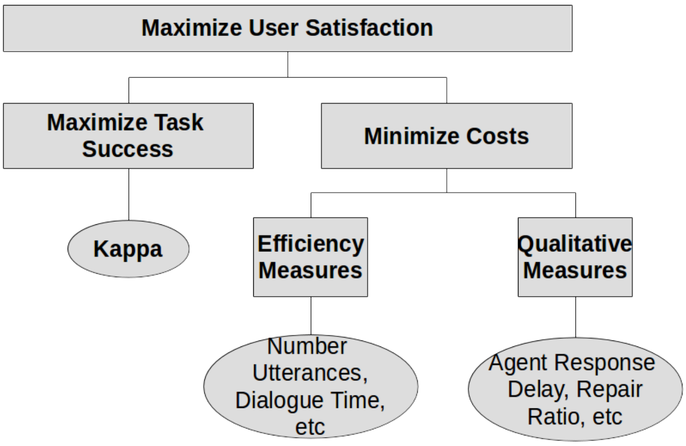


Figure 9.2: PARADISE's structure of objectives for spoken dialogue performance (after [Wal+97]).

agents [Wal+97]. The main goal is to maximize the users' satisfaction by maximizing the task success, while minimizing the dialogue costs. The main structure of objectives is illustrated in fig. 9.2. The authors define the performance p of a (sub-) dialogue with the section 9.1.

$$p = (\alpha \cdot N(\kappa)) - \sum_{i=1}^{n} w_i \cdot N(c_i) \tag{9.1}$$

In this case, κ describes the task success and c_i the different dialogue costs. The cost functions are normalized, weighted by their coefficient w_i. The task success is also normalized using a separate weight α. As an overall performance measure, the difference between the task performance and dialogue costs is calculated. The tasks are described using *Attribute Value Matrices (AVM)*, which consist of the information that must be exchanged between the agent and the user. The PARADISE framework was primary developed to compare different dialogue strategies in classical information retrieval domains, such as traveling systems. Therefore, the task success is defined as a slot-filling success rate and can be simply described as a *AVM*.

In *HAI* scenarios, such a slot-filling approach is not always possible. However, the objective measurement of task performance in general is a frequent metric, depending on the task at a hand. Especially in *HRI* the task success often plays an important role, even though it is not always closely related to the dialogue system itself. However, as each interaction is designed to full fit a special task, this is a reasonable approach.

As the dependent variables change, the methodology asses these metrics does too. These methods differ regarding their possibility to have the human "in the loop". In the research field of *Spoken Dialogue Systems (SDS)* special attention is paid to the use of corpus evaluation and user simulation techniques (e.g., [SGY05; GHL06], or see

[Liu+16] for an overview of unsupervised evaluation metrics). Eckert et al. proposed the use of simulated users to conduct dialogues with speech systems [ELP97b] already in the 90ies. They presented the advantages of simulated users, especially for the automatic evaluation and the generation of large data sets for statistical dialogue modeling and learning dialogue strategies. Besides the use of simulated users, corpus evaluation for supervised learning of dialogue policies is widely spread [Bug+04; Li+16]. To this end, many—usually domain specific—data needs to be collected. Based on a corpus, a dialogue policy can be trained, with classically machine learning techniques for training, validation and evaluation. Thus, using data set evaluation, is a quite obvious and a good opportunity for the evaluation of single parts or the whole dialogue system, also for robotic dialogue systems [Bug+04]. Eckert et al. also pointed out, that "Cumbersome manual work is greatly reduced by applying [...] automatic evaluation procedure[s]" [ELP97b]. However, they also believe "that tests with human users are still vital for verifying the simulation models." [ELP97b]. Beside the obvious advantages of automatic evaluation, a number of important difficulties are entailed. One main challenge is the collection of a data corpus, either for the evaluation itself or implicitly for the training of a simulated user. Often these corpora are text-based, which limits the possibility to integrate other modalities or user context, such as pointing gestures or the current *Visual Focus of Attention (VFoA)*, into the system evaluation. Even though, classical machine learning methods provide standardized evaluation scores which give a good initial impression of the learned policy, I doubt the informative value of such scores. While a policy with slightly improved values better covers the data set, it is not necessarily the better policy in general. In addition, for me, it is highly questionable which conclusions can be drawn from a poor performance and how the policy can be improved. Lastly but most importantly, it is not possible to evaluate feedback loops—the effect of the systems' behavior on the human interaction partner, which in turn affects the system behavior.

Another trend is the use of crowdsourcing platforms, such as Amazon Mechanical Turk (e.g., [Jur+11; Yan+10] or see [PE11] for an overview). Joosse et al. present several lessons learned from using crowdsourcing platforms as a methodology for gathering data as an *HRI* researcher [JLE15]. The idea is to run online studies with participants from all over the world. This methodology has several advantages. The greatest benefit of crowdsourcing is the possibility to collect a good amount of data quickly. Furthermore, it is possible to collect more diverse samples, such as participants from different countries, age, or cultural background. These online studies are often video studies in which participants rate different robotic behaviors. Another approach is to collect training data, by asking participants what they would do in specific situations, or what a robot should do. However, one should be aware that for validity checking and quality control an amount of manual labor is required. The biggest drawback is the limited opportunity to have the human "in the loop". As in the case of other online studies, the participant is not directly situated in the interaction. It is questionable if the human assessing the online interaction exactly reacts in the same way as in a situated interaction. Additionally, it is not possible to observe the human during the interaction—so important information is lost and the agent's possibilities to react to the human's behavior is narrowed.

Besides corpus evaluation, video studies, and questionnaires, another methodology in the evaluation of dialogue systems and *HAI* in general is the interaction study. This methodology is widely used in the *HRI* community [Bax+16; BM10] with various levels of autonomy—from fully autonomous to completely controlled by a wizard [Rie12]. However, Baxter et al. found, that "a majority of the research presented at the HRI conference does not involve interactive autonomous systems" [Bax+16]. So, even at one of the most influential conferences in the research field of *HRI*, the majority of interaction studies are controlled by a wizard. This is often a consequence of the motivation of the study. For example, the investigation of a special phenomenon not necessarily requires a fully autonomous system, with a *Wizard-of-Oz (WoZ)* interaction study, the robot can serve as a proxy for a human [Wei10; Rie12]. Riek discusses several concerns against this methodology, e.g., that "it is not really human-robot interaction so much as human-human interaction via a robot" and various ethical questions [Rie12]. Based on a comprehensive semantic review, she presents new reporting guidelines to help to circumvent these methodology concerns, including information regarding the robot, the users, the wizard, and the general experiment [Rie12].

Nevertheless, a more autonomous agent may enhance this type of study by reducing the influence of the wizards' human biases [Bax+16]. For other objectives, such as improving the agent or the interaction itself, a high level of agent's autonomy is essential. However, this methodology has several challenges [Bax+16; BM10]. While Bethel and Murphy discuss several insights from the psychology and social sciences for *HRI* studies, Baxter et al. examine publications in the HRI conference over three years and present some challenges and recommendations for the topics of *level of autonomy*, *participants*, *environment*, *study length*, *statistics*, and *replicability* [Bax+16]. The level of autonomy itself is one of the biggest challenges. The agent needs to be ready for the interaction with humans—to be sufficiently advanced to interact autonomously. Furthermore, the participant population—in *HAI* studies often drawn from university students and staff—plays an important role and needs to be considered when drawing general conclusions from a study. In addition, the participants should be balanced between the experimental conditions in terms of age, gender, and further demographics. The experimental environment and study length play further important roles and should influence the study design decisions. Another issue—which recently got more attention in the *HRI* community—is related to statistics and reproducibility. Baxter et al. present three main concerns towards the statistical analysis, the "arbitrary threshold for significance, replication sensitivity, and lack of effect size information" [Bax+16]. Especially, the variety of p-values over experiment replications highly questions its meaningfulness for inference and replication [Cum08]. Another aspect—especially in the *HAI* community—is the difficulty to replicate experiments, because of the nature of robotics hardware and agents software.

Besides these challenges and pitfalls, interaction studies provide the unique opportunity to have the human "in the loop". No other methodology provides such a profound insight into the *HAI* in general, the effects of system behavior on the participants, and vice versa—regardless of whether a quantitative or qualitative evaluation is performed. The human should be at the center of our research field—as we develop intelligence systems for humans. Consequently, the human should be at the center of the evaluation of such systems.

Besides the recent trend towards video and online studies, I perform user studies to answer my research question, as this method provides the only opportunity to have the human "in the loop". Even though this methodology is more difficult to operationalize and can be influenced by various external factors, I expect better insights into the *HAI* itself and a better understanding of modeling the coordination requirements of such interactions. I discuss how I counteract some previously mentioned pitfalls in chapter 9. Furthermore, I apply multiple methods of evaluation. To investigate my hypothesis, I assess the task performance. In addition, side effects of the systems are assessed through behavioral measures and self-assessments of the participants.

9.2 METHOD OF EVALUATION CYCLES

The background of the evaluation of *Human-Agent Interaction (HAI)* in section 9.1 shows that the method of interaction studies are providing the only opportunity to have the human "in the loop". I expect better insights into the *HAI* itself and a better understanding of modeling the coordination requirements of such interactions. Therefore, I evaluate my hypothesis that:

> **Hypothesis:** *The Attention-Hesitation Model (AHM) will increase the task performance in human-agent interaction.*

in five *Evaluation Cycle (EC)*. In each cycle (see fig. 9.3), the RQ 3 is addressed by an *HAI* interaction study. In addition, again the research questions RQ 1 and RQ 2 are addressed. Therefore, the model is iteratively improved, and different hesitation features are explored to find an intervention strategy that can improve the task performance without having negative side effects on the interaction. Besides the model itself, the implications in the implementation are discussed further. In addition, the evaluation approach in each cycle is improved. By this means, the results of each *EC* influence the next one. According to the guidelines by Riek [Rie12] and the recommendations

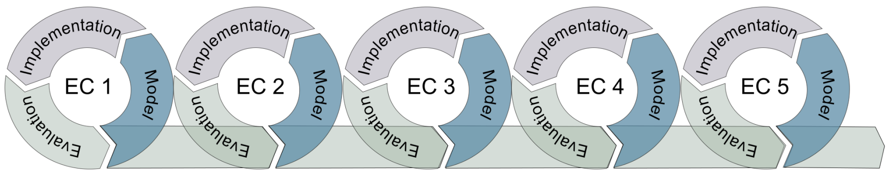


Figure 9.3: The method of evaluation cycles used in this thesis. The results of the previous cycle influence the model, its implementation as well as the evaluation study.

by Baxter et al. [Bax+16], key aspect for the design of interaction studies performed in these *EC* are presented.

SCENARIO The *HAI* take place in the *Cognitive Service Robotics Apartment (CSRA)*. As interaction partner, the virtual agent Flobi is used in all *ECs*. The scenario consists of a simple information providing situation, in which the agent provides information about itself or the intelligent environment to the human interaction partner.

ENVIRONMENT As environment, the *CSRA* is used. As discussed in section 7.4.1, it provides a good balance between the possibility of a natural interaction and a controlled environment.

LEVEL OF AUTONOMY As already discussed, the use of *Wizard-of-Oz (WoZ)* interaction studies has benefits and drawbacks. I use the benefit of controlled behavior to by performing *WoZ* interactions. In addition, I also perform fully autonomous interactions to counteract the drawback, that such controlled interactions are not replicable with autonomous agents.

PARTICIPANTS In my interaction studies, I mainly acquire test subjects from the university campus. It cannot be generalized to other groups. However, I balance between the experimental conditions in terms of age and gender.

REPLICABILITY The topic of replicability is addressed through the use of the *Cognitive Interaction Toolkit (CITK)* on the one side and use of multiple interaction studies on the other side.

9.2.1 *Experiment Procedure*

All *HAI* interaction studies are conducted in the interaction zones of the *CSRA* smart home environment (see section 7.1.1) and follow the same experiment procedure visualized in fig. 9.4:

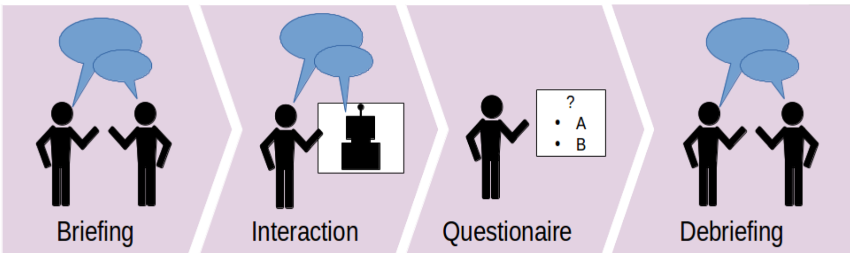


Figure 9.4: General procedure of interaction experiments conducted in this thesis.

BRIEFING During the *briefing* phase, the participants read and accept the personal data protection conditions of the *CSRA* (appendix E). Furthermore, they received general information about the apartment and the interaction study. Naturally, the objective of the experiment is not dis-closed (see participants instructions in appendix D). In some cases, an additional pre-tests is conducted in this phase.

INTERACTION The second phase consists of the *interaction* itself. The participants enter the *CSRA* alone and are monitored via cameras on the ceiling of the apartment from an adjoining room—the control room.

QUESTIONNAIRE After the interaction, all participants complete a *questionnaire* on a computer (alone).

DEBRIEFING In the subsequent *debriefing*, qualitative interviews are conducted, the aims of the experiment are explained, and the participants receive a monetary compensation.

All experiments were ethically approved by the ethic committee of the Bielefeld University[1]. In all studies, I chose a between-subject interaction experiment design to avoid corrie over effects between the conditions. Participants interacting either with the agent with the Attention-Hesitation Model (AHM) (condition *AHM*) or an agent without my model, which is the baseline in all experiments (condition *BASE*).

DATA RECORDING AND ANNOTATION: Interactions were recorded via up to four network-enabled Basler cameras, a webcam facing the user from the agent's perspective and one Rode NT55 omni-directional microphone mounted at the ceiling of the apartment to cover the whole interaction area. Moreover, I collected system events, e.g., generated dialogue acts and if available detailed information about the gaze recognition results. For annotation purposes, the videos, the audio stream and system events were automatically merged into one ELAN [Wit+06] file. One example view is depicted in fig. 9.5 For further information about this process refer to [Hol+16]. The question-

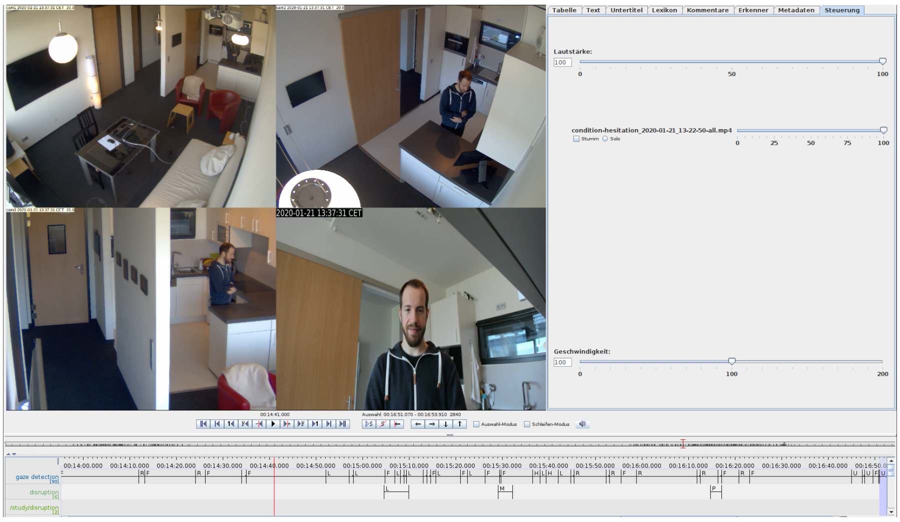


Figure 9.5: Annotation view: Different camera views of the apartment are merged into one file.

naires are conducted with the survey software *LimeSurvey* [Lim19].

DEPENDENT AND INDEPENDENT VARIABLES: To verify my hypothesis, I measured the task performance in each interaction. Since the task is to provide information, I started with post-interaction in-

1 https://www.uni-bielefeld.de/uni/einrichtungen-organisation/zentrale-organisation/kommissionen/ethik

formation recall in EC1. Over the ECs, I developed the assessment of the task performance further, from the post-interaction information recall via questionnaire up to the assessment via a practical task. Table 9.2.1 gives an overview of the measurements for each cycle, which are explained in more detail in the description of the cycles itself. For the task performance, the participants' information recall of the information state by the agent is measure. To this end, I start with a simple questionnaire in the first cycle. Throughout the *ECs*, I am developing the assessment of the task performance further

To measure the side effects, I chose the *Godspeed Questionnaire Series (GQS)* [Bar+09]. The used translation can be found in appendix B.4.This instrument is well-developed for *Human-Robot Interaction (HRI)* and covers important key aspects, such as likability. In addition, participants had in all interaction studies the opportunity to leaf general comments in *free text* form at the end of the questionnaire. From EC4, the subjective ratings of the synthetic voice quality is additionally assessed, based on the results of the comments in the questionnaire. Additionally, I structured the debriefing and conducted semi-structured interviews in the last two interaction studies. Furthermore, the *Visual Focus of Attention (VFoA)* is analyzed in most of the cycles.

To verify that the participants a balanced between the experimental conditions, I collected participants' demographics in terms of age and gender. Furthermore, their average prior experience with technical intelligent systems (see appendix B.5).

Table 9.2.1: Measurements of task performance and side efects for each cycle.

		Task Perfm.			Side effects					
					Subjective Ratings				vFoA	
EC	Autonomy	Information	Prac in Interact.	Prac. post Interact.	Godspeed	MOS	Interview	Free Text	Look Away	Total Time
EC1	WoZ	✓			✓			✓	✓	✓
EC2	Autonomous	✓			✓			✓	✓	✓
EC3	WoZ		✓	✓	✓			✓		
EC4	Autonomous			✓	✓	✓	✓	✓	✓	✓
EC5	Semi-Autonom.	✓		✓	✓	✓	✓	✓	✓	✓

STATISTICAL ANALYSIS: The statistical analysis is influenced by Field et al. [FMF12]. For comparing two means, I will present results of the parametric *Welch two-sample t-tests (T-Test)* [Wel47], or use a *Multivariate Analysis of Variance (MANOVA)* [TF11] for a multivariate analysis of more dependent variables. For the effect size I will present Cohen's d [Coh77] for t-test and η^2 for the *MANOVA*. To check whether the criteria for parametric test are given, I use the Shapiro-Wilk normality test [SW65] and Levene's test for homogeneity of variance [Lev60]. I use the software environment for statistical computing and graphics R [R C19] and the IDE RStudio [All11; RSt15]. Most figures are created with the R-package *ggplot2* [Wic16]. Hypotheses are tested with the *stats*, *psych* [Rev19] and the *effsize* [Tor20] R-package. Due to the study design, the independence of the data can be assumed. Furthermore, dependent data is at least at the interval level. Whenever the assumptions of parametric tests are not met, results of *Wilcoxon Rank Sum Test / Mann-Whitney U tes (WR-Test)*, or *Wilcoxon Signed Rank Test (WSR-Test)* for repeated measures are presented with the corresponding effect size $r = \frac{z}{\sqrt{N}}$ [RGL06]. For the statistical analysis, an alpha level of 0.05 is chosen. In cases of multiple hypothesis testing, the Bonferroni correction [Bon36] in the corresponding R-package is used to adjust the $p - values$.

In addition, I use the usual marks for statistical significance: $*p < .05; **p < .01; ***p < .001$.

9.2.2 *Overview of Evaluation Cycles*

In this section, a short overview of the experiments conducted to investigate different *Attention-Hesitation Models (AHMs)* is given. In total, I performed five *ECs*, consisting of three pilot- and two *HAI* studies in a smart-home environment:

Evaluation Cycle 1: Self-interruptions as Attention-regain Strategy (EC1) In a *WoZ* pilot study, I am investigating whether it is possible to regain the attention of distracted users by applying self-interruptions as a simple attention-regain strategy. Therefore, a simple attention regain strategy in a short interaction scenario within a smart environment is tested. In this case, the agent uses unfilled pauses—simple self-interruptions—as an attention-regain mechanism. It is applied whenever the human looks away from the agent.

Evaluation Cycle 2: Introducing the Focus of Discourse Feature (EC2) In the second cycle, I explore how the model needs to be changed to deal with different *Focus of Discourses (FoDs)*. Therefore, the concept of attention to *Attention on FoD* need to be changed, meaning that the agent applies the attention-regain hesitation strategy when the human attention moves away from the current *FoD* or the agent itself. Here, the hesitation strategy consists again of unfilled pauses. In addition, when the attention guiding to a new *FoD* failed, the agent reacts with repetitions as attention-highlight strategy. Furthermore, I investigate how an *AHM* can be implemented in a fully autonomous agent.

Evaluation Cycle 3: Exploration of a Practical Task during Interaction (EC3) The third evaluation cycle is an excursus to another—more practically—interaction within the smart home. Here, I approach the topic from a different view point. The agent still acts with a hesitation strategy. In contrast to the previous experiments, the agent starts the strategy at predefined points of the interaction. When the human is attentive again, the strategy stops. However, the concept of attention is different in this interaction. It is interpreted as the readiness for the execution of the next (sub-)task.

Evaluation Cycle 4: Introducing the Lengthening Feature and new Evaluation Approach (EC4) In the fourth cycle, I return to my initial scenario. Influenced by the results of the excursus, the task of the participants is adapted. Furthermore, lengthening as a feature is introduced into the hesitation strategy. The agent produces lengthening and unfilled pauses whenever the human looks away from it in this fully autonomous interaction study.

Evaluation Cycle 5: Bringing It All Together (EC5) In the final interaction study, I combine results from the previous studies. The resulting enhanced *AHM* uses the concept *Attention on FoD* and more advanced hesitation strategies, which consist of lengthening, unfilled pauses, hesitation vowels and repetitions.

Table 9.2.2: Overview of different features utilized for the attention model and the corresponding intervention hesitation strategy.

	Attention concept				**Hesitation strategy**				
		Task			**Re-attention**			**Highlight**	
EC	Mutual Gaze	Directed Gaze	Discourse History	Task Progress	Unfilled Pauses	Lengthening	Hesitation Vowels	Repetitions	Lengthening
EC1	✓				✓				
EC2	✓	✓	✓		✓			✓	
EC3	✓			✓	✓			✓	
EC4	✓				✓	✓			
EC5	✓	✓	✓		✓	✓	✓	✓	✓

Table 9.2.2 depicts the features used in the different *AHMs* in each cycle. For the concept of attention, mutual gaze is utilized in all *ECs* and various task related features are explored. As hesitation

intervention strategy, unfilled pauses are utilized in each *EC*. Further features are integrated over the cycles.

10 EVALUATION OF THE ATTENTION-HESITATION MODEL

In this chapter, I carry out five *ECs* to investigate my Attention-Hesitation Model (AHM). In each cycle, I describe briefly the current version of my model in terms of which concept of attention is used and what kind of hesitation strategy as intervention is performed by the agent. Afterwards, the scenario, and its implementation is described, followed by the interaction study design. The results of the experiments are discussed in each cycle regarding my hypothesis towards the task performance, and the metrics for the side effects: subjective ratings and visual attention. To investigate the *AHM*, an information-providing scenario in the *CSRA* is used. The interaction consists of the agent giving the person information about itself and about the intelligent apartment. Present the *CSRA* and its capabilities occurs in each demonstration of the intelligent environment and serves as a suitable interaction test situation as it consists of long information statements by the system. In addition to this information-providing scenario, an interaction with a practical task during the interaction is investigated in the third cycle (section 10.3).

10.1 EC 1: SELF-INTERRUPTIONS AS ATTENTION-REGAIN STRATEGY

In this section, I investigate how an agent can reacquire a user's attention when it drifted away during a human-agent interaction. The goal of this pilot *WoZ* study is to test a simple attention regain strategy in a short interaction scenario. In this case, the agent uses unfilled pauses—a simple self-interruption—as a re-attention mechanism. The results of this study are partly published in [CSW16a].

My hypothesis implies that the *AHM* increase the task performance of the user, as measured by post-interaction information recall. The interaction partner can remember more information from the interaction, because of the intensified attention. This may also be reflected in the behavior of participants. Therefore, the side effects of the self-interruptions of the agent—whenever the visual attention of the human interaction partner is lost—on the interaction are analyzed. More precisely, if the invention strategy is a successful attention regain strategy, participants may less inattentive. This is measured by the gazing behavior of the participants, i.e., the number of look aways and the total time of being inattentive. In addition, the self-interruptions may have an effect on the subjective ratings of the agent—regardless of whether positive or negative nature.

The experiment design I choose is a *WoZ* between-subject method. This way, my hypothesis can be tested without the need to implement the whole scenario. Consequently, this experiment does not inform about whether if it is possible to implement an autonomous system with re-attention capabilities.

10.1.1 *Attention-Hesitation Dialogue Coordination Model*

The dialogue management model has two main responsibilities, *when to (re-)act* and *how to (re-)act*. As stated in chapter 4, the moment of action depends on the attention-state of the human interaction partner and the action itself is a hesitation. In the following, I describe the model chosen in this experiment in more detail.

WHEN TO (RE-)ACT: ATTENTION CONCEPT The agent starts an intervention strategy whenever the human is inattentive. Corresponding the results of [Goo81], missing mutual gaze can be interpreted as missing listeners attentions. In this model, the human is inattentive

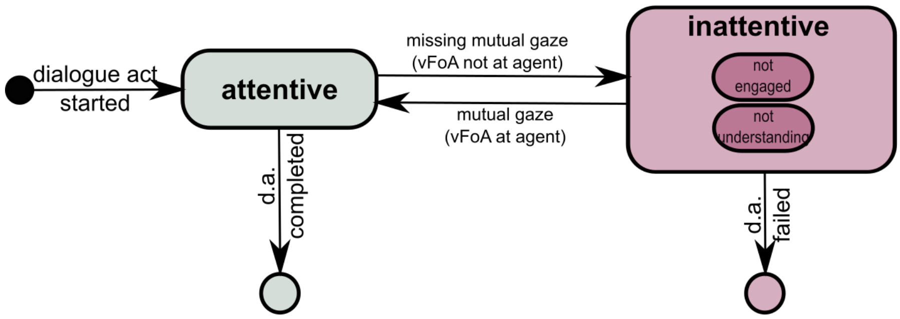


Figure 10.1: Concept of attention for EC1.

when mutual gaze is missing, i.e. the *VFoA* moves away from the agent (see fig. 10.1). As soon as the attention is back—meaning the human interaction partner looks back at the agent—the intervention strategy stops. In this simple model, there is no further distinction between the reason of inattentiveness. The agent reacts with the same intervention strategy, regardless of whether it is due to engagement or understanding problems.

HOW TO (RE-)ACT: HESITATION INTERVENTION STRATEGY Whenever the agent loses the attention of the human interaction partner, it reacts with a hesitation, depicted in fig. 10.2. In this model, the agent uses an unfilled pause—a simple self-interruption—as a re-attention mechanism, as soon as the user looks away. The interaction strategy stops as soon as the user looks back at the agent. In this model, the agent simply continues speaking.

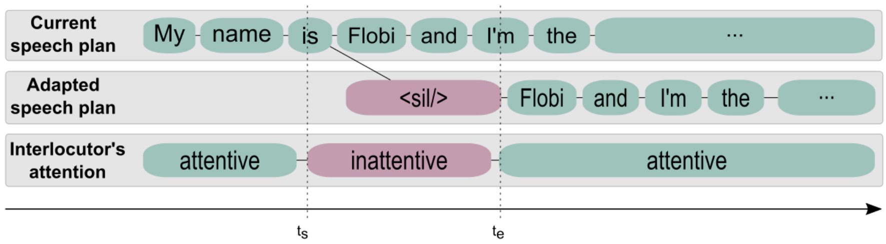


Figure 10.2: Hesitation intervention strategy for EC1.

10.1.2 *Interaction Scenario and Implementation*

The *HAI* takes place in the hallway of the smart home environment *CSRA*. The agent—in this case simulation of the anthropomorphic robot head Flobi (see section 7.1)—is providing information about itself through a sequence of 6 sentences. The *FoD* does not change in this first interaction, since the agent only transmits information about itself. Figure 10.3 shows the interaction setup. The *HAI* has three phases of interaction:

GREETING: The agent welcomes the user. The user has the possibility to great back.

INFORMATION: The agent introduces itself and give information about itself to the human interaction partner (six items).

FAREWELL: The agent says goodbye to the user and requests to fill out the questionnaire at the computer in the room to the right of the participant.

The interaction is realized using the dialogue system presented in chapter 7. Users were facing a tablet, which was shows the virtual Flobi (section 7.1). Through the tablet's camera, Flobi can detect faces in front of it and focus on them, thus establishing shared attention. Flobi's verbalization were predefined. To allow verbal self-interruptions, the incremental speech synthesis module by *Incremental Processing Toolkit (inprotk)* is used, that pause the ongoing speech after the current word (see section 7.3). To be able to start this behavior from another room, a wizard GUI is implemented that sends a start or stop signal to the synthesis module.

Figure 10.3: Experimental setup for EC1. Upper: person interacting with the agent. Lower: ground view of the apartment.

10.1.3 *Evaluation*

I evaluate the effects of verbal self-interruptions of the agent in this first pilot study.

Study Design

The pilot study is designed to measure the effects of my *AHM*. A between-subject interaction study design is chosen to avoid carry-over effects between the conditions. Therefore, participants are randomly assigned to one of two conditions. In the *AHM* condition, the agent reacts with hesitations whenever the *VFoA* moves away from the agent, according to my *AHM*. In the *BASE* condition, the agent does not react to the inattentive interlocutor and kept on speaking. In both conditions, an external distraction is provided in the apartment to the right side of the participant, at an angle of about 90 degrees, to withdraw the user's *VFoA* from the system. The self-interrupting behavior of the system in the *AHM* condition is triggered by the wizard through pressing a button upon perceiving the user's *VFoA* shifting away. This is achieved by observing the whole situation through the different camera views of the apartment presented in section 9.2.1. Based on the overview cameras as well as the webcam picture facing the user from the agent's perspective, the wizard can observe the whole interacting area from the ingoing *control* room. The agent directly stops speaking and continue exactly at the break-off point when the user's *VFoA* returned to the agent—the wizard presses the corresponding button. In the *BASE* condition, the agent continues speaking. The distraction is achieved by a study assistant reentering the room, pretending to bring in some missing documents for the experiment, phrasing a brief verbal apology with an explanation, and leaving.

The experiment design follows the general study design depicted in fig. 9.4. After signing a consent form, the subjects are led to the experiment room. They enter the hallway of the apartment alone through *Door A*. Their instruction is to look at the tablet on the left wall and to fill out a questionnaire on the computer after the interaction. The wizard starts the interaction as soon as the participants stand in front of the tablet, facing it. Phase two of the *HAI* is purposefully disrupted by one of the study assistants, which enters the experiment room through *Door B*. The study assistants disturb the interaction always after the first sentence of the information phase is finished. After the interaction, the participants go to the table and fill out the questionnaire on a computer. Afterwards, a debriefing is performed, and the participants receive a monetary compensation.

The questionnaire consists of three parts: a memory task, subjective ratings about Flobi, and demographics. The memory task consists of six statements, for which the participants had to decide whether this was a statement made by the agent during the information phase (see appendix B.1). In the second part of the questionnaire, the participants have to provide subjective ratings of the agent through a set of adjectives on Likert scales to evaluate five key concepts in human-robot interaction: *anthropomorphism*, *animacy*, *likability*, *perceived intelligence*, and *perceived safety*. Therefore, a translated version of the *GQS* is used, which can be found in appendix B.4. Furthermore, in the last part they have the opportunity to leaf general comments at the end of the questionnaire and demographics and previous experiences accessed. To assess the task performance of the participants, the number of correct answers to the content-related questions of the questionnaire are counted. To obtain a measure for the inattention, the gazes shifts of the user are annotated. The number and duration participants looked away from the agent during the information phase of the interaction are measured. For the subjective ratings, the answers of the second part of the questionnaire are evaluated: the *GQS* and the general comments.

Participants

Participants are recruited at the campus of the University Bielefeld and are mostly students or from the administrative staff. In total, 27 subjects (9 female, 18 male, aged 21-51) took part in the study. The average age is 27.2 with a standard deviation of 5.3. 13 participants are in the *AHM* (11 male, 2 female; $M_{age} = 27.00$, $SD_{age} = 2.55$) and 14 in the *BASE* condition (9 male, 5 female; $M_{age} = 27.36$, $SD_{age} = 7.15$). The study assistants disturbed the *HAI* in the experimental condition (*AHM*) 10.47 seconds in average and in the baseline 10.25 seconds.

Participants in the *AHM* condition have slightly higher average prior experience with technical intelligent systems ($M_{AHM} = 3.86$, $SD_{AHM} = 0.81$; $M_{BASE} = 3.10$, $SD_{BASE} = 1.04$), $W = 136, p = .033$. Participants in both condition has no or very little experience with robotic systems in general or the virtual agent Flobi. Participants in the baseline have some experience with speech systems in general, while in the *AHM* condition they have only little experience with speech systems. In addition, participants in both conditions have programming experiences, Some experiences in the baseline and many in the experimental condition. All participants are very experienced with the use of computers in general.

Task Performance Hypothesis

At first, I explore the task performance, measured as post interaction information recall. Note that all questions are yes/no questions. The percentages of correct answers for each condition for the different memory questions are shown in fig. 10.4.

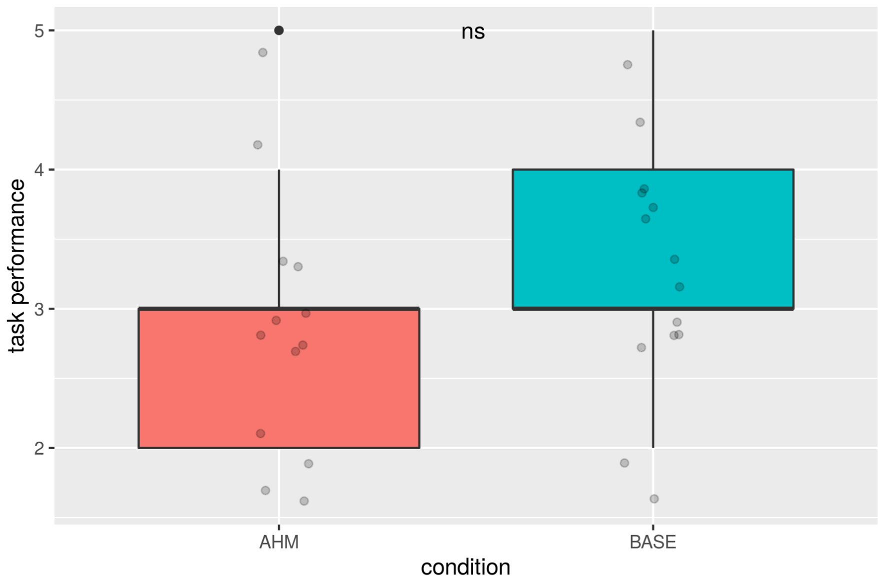


Figure 10.4: Results of the task performance in *EC1*.

The overall percentages of correct answers for the experimental condition (*AHM*) is 48.7% whereas the subjects in the *BASE* condition answered 56.0% correct. The hypothesis that participants achieve higher task performance is tested using a *WR-Test*. There is no statistically significant difference between the task performance in the *AHM* ($M = 2.92, SD = 0.86$) and the *BASE* ($M = 3.36, SD = 0.84$) condition, $W = 63, p - value = .151$. I am thus not able to confirm my hypothesis that the *AHM* has a positive effect on the post-interaction information recall.

Side effects on the Interaction

Next, I investigate the side effects on the interaction, regarding the subjective ratings of the agent and the *VFoA* of the human interaction partner.

SUBJECTIVE RATINGS Figure 10.5 depicts the subjective ratings of the agent for the **five key concepts** *anthropomorphism*, *animacy*, *likability*, *perceived intelligence*, and *perceived safety*. The agent receives high values for likability, perceived intelligence, and perceived safety ($M > 3.1$) but rather low value for anthropomorphism ($M < 2.0$) in both conditions. Exact values can be found in table G.3 in appendix G.1.1. I test the hypothesis that participants rate the agent in the *BASE* condition and the *AHM* condition differently. To this end, I apply the Welch t-sample t-test for the five key concepts. Results show that the key concepts were not rated significantly different after the interaction in the *BASE* or *AHM* condition ($p > 0.1$). The complete test statistics can be found in Table G.2 in appendix G.1.1. The general comments at the end

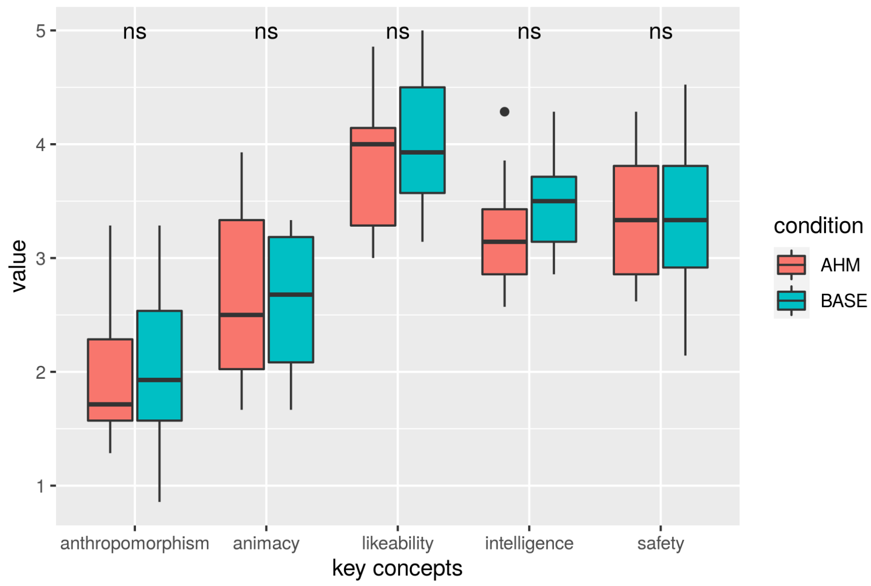


Figure 10.5: Subjective ratings of the agent: results of the *GQS* in *EC1*.

of the questionnaire contained no comment on the self-interrupting behavior of the agent.

VISUAL FOCUS OF ATTENTION Furthermore, I take a closer look at the *VFoA* of the participants to check whether participants in the *AHM* condition show shorter inattentiveness than participants in the *BASE* condition. This is measured as the total time of being inattentive—in this scenario defined as looking away from the agent—or in the number of users looking away from the agent. Figure 10.41 visualize the distribution of the *VFoA* in the different conditions in total time and number. While in the *AHM* condition two-thirds of participants do not look away more than once, in the *BASE* condition more than half of the participants do. The assumptions for a *MANOVA* is not met, therefore the non-parametric *WR-Test* is applied. The number of look aways in the *BASE* condition ($Mdn = 2$) differs significantly from the participants in the *AHM* condition ($Mdn = 1$), $W = 54, p = .027, r = -0.38$. Participants in the control group look away more often. Furthermore, the overall time of participants looking away from the agent—the current *FoD*—is longer in the *BASE* than in the *AHM* condition. This can be seen in the right graph of fig. 10.41. There was a significant difference in the overall time not looking at the agent for the *AHM* ($M = 2.1s, SD = 1.5$) and the *BASE* ($M = 4.5s, SD = 3.5$) condition; $W = 46, p = .016, r = -0.42$.

As seen in fig. 10.7 the average length of the first looking away does not differ between the two conditions. However, in the baseline the variance is larger. The same observation can be made for the second and third time looking away. No participant looked away more than three times.

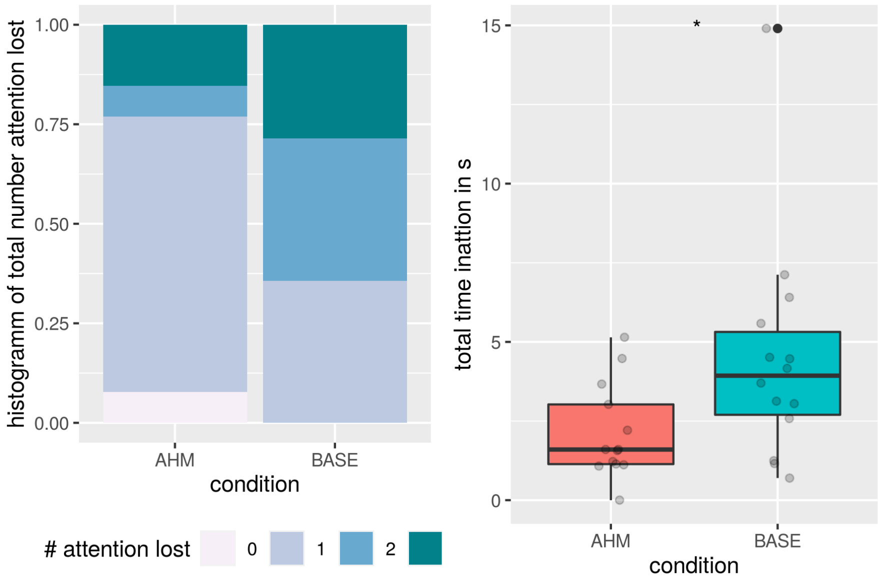


Figure 10.6: Distribution of inattentiveness: (left) the number of look away (NA_{number}). Whenever the user look away from the agent, this number increase by one. (right) the total time (NA_{totoal}) participants are inattentive.)

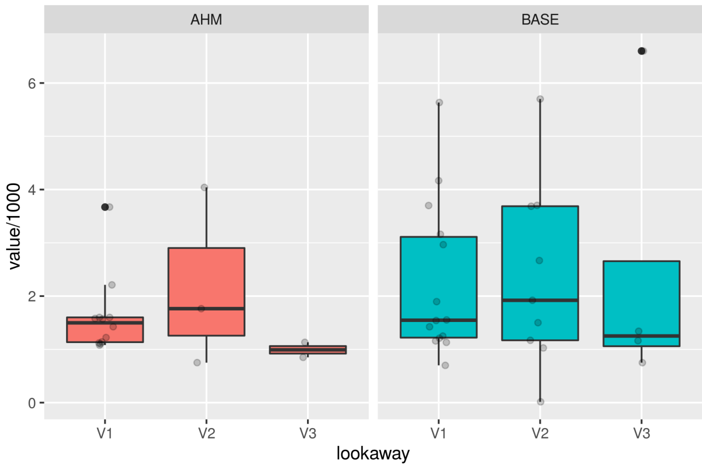


Figure 10.7: The time of *VFoA* moves away separated for the individual looking away.

10.1.4 *Discussion and Lessons Learned*

With this first pilot study, I investigated whether a simple *AHM* affects the attention of the human interaction partner. The looking behavior measurements in this experiment suggest that the self-interrupting of the agent has a significant effect on the *VFoA* of the human interaction partner. This effect manifests in the overall time and the number of times participants looked away from the current *FoD*. These results indicate that the self-interrupting agent has an effect on looking behavior of the human interaction partner. Specifically, it indicates that unfilled pauses—simple self-interruptions of the agent—are an effective intervention strategy to regain the attention of the interaction partner. It could have turned out the other way. The interaction partner could have taken the additional time and pay more attention to something else, but interestingly, this was not the case. The unfilled pause leads to less inattentiveness. However, this effect was not reflected in the task performance. Participants in the baseline and *AHM* condition show no significant differences in the memory task. Furthermore, no significant difference in the subjective ratings of the agent could be measured.

The general experiment procedure was suitable for this investigation at hand. However, improvements can be made. First, the interaction time should be longer, to increase the amount of attention shifts and the possibility to apply the intervention strategy. In addition, the interaction questionnaire to measure the task performance need to be improved. The fact that the median in both condition is three (of six) indicates that participants simply guessed in both conditions and the simple yes/no questions were to difficult. Furthermore, the interaction should be extended to different *FoD*. In this scenario, the agent only speaks about itself, it is questionable how the *AHM* affect the participants in more complex interaction scenarios. Finally, this experiment does not inform about whether if it is possible to implement an autonomous system with re-attention capabilities. It is still questionable, if such a system works autonomously.

10.2 EC 2: INTRODUCING THE FOCUS OF DISCOURSE

The results of the first evaluation cycle in section 10.1 indicated, that unfilled pauses can be used as a simple intervention strategy to regain the visual attention of the human interaction partner. However, it gave no insights about the realization of such systems. Furthermore, more complex interactions with different foci of discourses should be investigated. In this section, it is depicted how the *AHM* need to be enhanced to deal with different *FoD*. To this end, the integration of the required modalities is explained. In addition, the interaction study is conducted with the resulting autonomous system. Parts of the work, presented in this section are published in [CSW16b].

10.2.1 *Attention-Hesitation Dialogue Coordination Model*

Whereas in the first experiment the *FoD* simply is on the agent itself, it changes over time in more complex interaction scenarios. In these interactions, it is possible to guide the attention of the user while changing the *FoD*. The following model accounts for these requirements.

WHEN TO (RE-)ACT: ATTENTION CONCEPT This model observe the current *VFoA* as well as the current *FoD*, as depicted in fig. 10.8. It distinguishes two different situations (a) the current *FoD* changes and (b) the *FoD* is ongoing. In (a) the user is inattentive, if the attention guiding strategy fails. I classify this state as an understanding problem because the interaction partner did not follow the attention shift. This means, the user does not look once at the new *FoD* within a time frame. For an ongoing *FoD* (b), the user is inattentive whenever the *VFoA* neither matches the current *FoD* nor the agent itself. I classify this state as an engagement problem because the agent misses joint attention throughout the current *FoD* and without a discourse change.

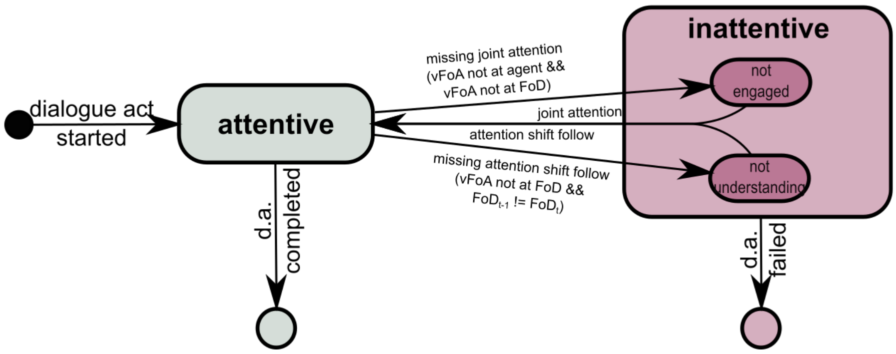


Figure 10.8: Concept of attention for EC2.

HOW TO (RE-)ACT: HESITATION INTERVENTION STRATEGIES Whenever the agent looses the attention of the human interaction partner, it reacts with a hesitation (see fig. 10.9). In this model, the agent uses unfilled pauses—simple self-interruptions—as attention regaining mechanism, as soon as it looses the user's attention during an ongoing *FoD*, meaning an engagement problem occurs (b). It stops this intervention strategy, as soon as the user looks back at the current *FoD* or the agent and continues speaking.

Whenever the attention guiding strategy fails, meaning an understanding problem occurs (a), the agent reacts with a highlight attention strategy: a repetition of the multi-modal guiding strategy. This is repeated, with a short pause in between, until the user is attentive—looked at the new *FoD* at least once.

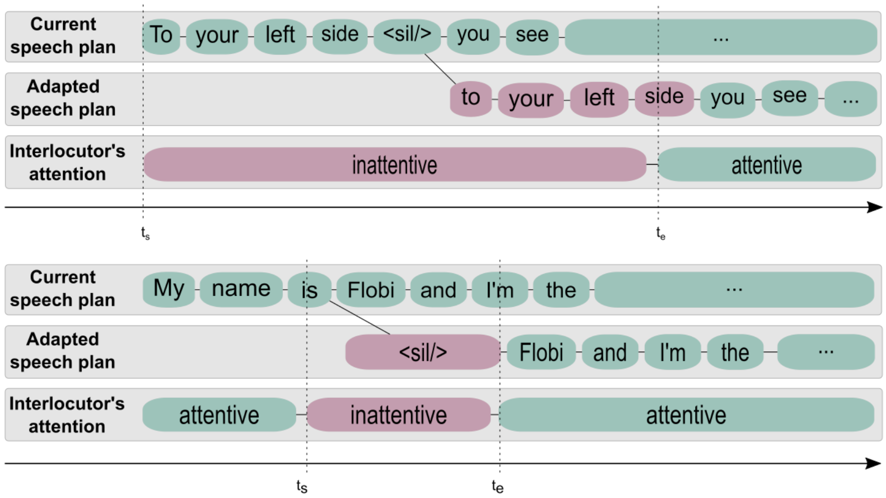


Figure 10.9: Hesitation intervention strategies for *EC2*: (a) highlight (b) re-attention.

10.2.2 *Interaction Scenario and Implementation*

The *AHM* is implemented as a *HAI* within the *CSRA*. The scenario is almost the same as in the first pilot study section 10.1, besides some extensions to the interaction. In contrast to the first pilot study, the *FoD* changes over time. Whereas in the first interaction the *FoD* only is on the agent itself, in this interaction Flobi also talks about its environment and explains parts of the smart environment through a sequence of 23 sentences. The scenario contains four different *FoD*: (1) the agent itself, (2) the kitchen unit, (3) the living room with an interactive table and (4) the ceiling. This allows to test the attention guiding strategy of the agent. Accordingly, different areas for the *VFoA*: on the agent, kitchen, living, ceiling and other. In addition to the verbalization of objects, the agent uses for some information support by corresponding non-verbal actions of the apartment (in the following referred as embodied information). This is explained in more detail later in this section. To be able to talk about the different *FoD* the interaction area is switched from the hallway to the kitchen.

To account for the 'how' in this model, the agent produces unfilled pauses or repeats its attention guiding strategy as necessary. This means, whenever the interaction partner does not react to an attention drift—did not look at the new *FoD* at least once—the agent repeats its strategy. Whenever the user does not look at the *FoD* or the agent, it stops speaking.

The interaction setup is depicted in fig. 10.10. The users face a monitor, which is showing the virtual agent Flobi. Using a camera on top of the monitor, Flobi can detect faces in front of it and focus on them to establish shared attention. As in the previous study, the *HAI* has three phases of verbal action by the agent:

GREETING: The agent welcomes the user. The user has the possibility to great back.

INFORMATION: : The agent gives information about itself, the kitchen, the living room and the ceiling of the intelligent apartment.

FAREWELL: : The agent says goodbye to the user and request to move on to fill out a questionnaire at the computer in the living room to the right of the participant.

In the second phase, Flobi gives information about itself and the intelligent apartment. It talks about different objects within the apartment, thus a specific point of interest can be defined for each *FoD*. In addition to the verbalization of objects, some information are supported by corresponding non-verbal actions of the apartment (embodied information). The following examples illustrate the difference between embodied and non-embodied information.

Example 10.2.1: Embodied information: Additionally presented via a corresponding actor within the apartment

Verbalization: "Der Griff der Schranktür - links neben mir - leuchtet blau auf, wenn ich dir dort etwas zeigen möchte."
(The handle of the cabinet door to my left lights up blue if I want to show you something there.)

Action: The cabinet door handle lights up blue once.

Example 10.2.2: Non-embodied: Without any support from additional actors within the apartment.

Verbalization: "Der Tisch im Wohnzimmer ist interaktiv. Mann kann sich auf ihm die Karte der Wohnung anschauen."

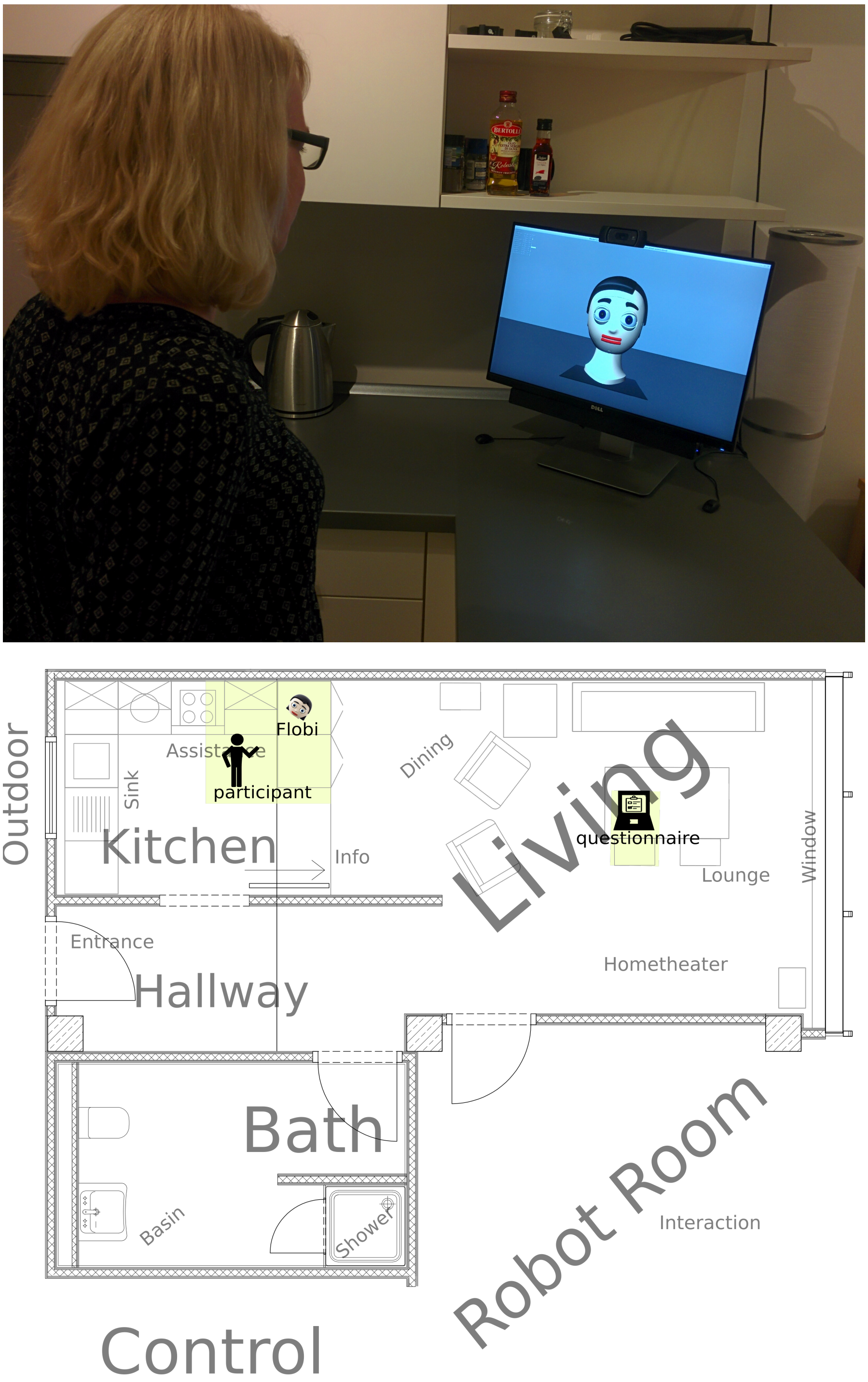


Figure 10.10: Experimental setup for EC2. Upper: person interacting with the agent. Lower: ground view of the apartment.

(The table in the living room there is interactive. You can look at him the map of the apartment.)

Action: None. The table stays inactive.

While in example 10.2.1 the verbal explanation is supported by the light flashing briefly, in example 10.2.2 the table does not present the current information. The light flashing is automatically initiated by the dialogue system via a system task within in *CSRA*.

ATTENTION GUIDING: Flobi uses a multi-modal highlight if the current *FoD*. Although the agent Flobi does not have a body that points to certain objects, it can highlight the current *FoD* using multi-modal communication signals presented in section 7.3. It is possible to verbalize attention shift, e.g., with the phrase "to your left side". Furthermore, it uses the highlight service of the *CSRA* presented in section 7.3.3. In the following experiment, I only use utterances with gaze behavior. Two highlight strategies are repeatable:

Listing 10.2.1: Configuratuion for the repeatable highlight strategies.

```
verbalization text="Links von dir ist die Kueche."
attention target_id="kitchen" modality="GAZE"
    duration="3000000"

verbalization text="Rechts von dir siehst du das
    Wohnzimmer."
attention target_id="living" modality="GAZE"
    duration="3000000"
```

ATTENTION MONITORING: The attention module distinguishes between ongoing *FoDs* and a change of the *FoD* with a simple rule-based model. If the attention guiding was not successful—meaning the human was inattentive during an attention shift—the agent reacts with a multimodal repetition of the current attention guiding strategy. To monitor the attention of the user, the gaze detector by Schillingmann and Nagai is used [SN15]. It provides information about the current *VFoA*, presented in section 7.3.1. In combination with the information about the current *FoD* from the dialogue management, the attention module (section 7.3.2) is possible to assess the attention state of the user. Furthermore, it distinguishes two different situations (a) the current *FoD* changes and (b) the *FoD* is ongoing. In (a) the user is inattentive, if the attention guiding strategy fails, which is classified as an understanding problem because the interaction partner did not follow the attention shift. This means, the user does not look once at the new *FoD* within a time frame. For an ongoing *FoD* (b), the user is inattentive whenever the *VFoA* neither matches the current *FoD* nor the agent itself. This state is classified as an engagement problem because the agent misses joint attention throughout the current *FoD* and without a discourse change.

INTERVENTION STRATEGIES: Furthermore, Flobi reacts with an unfilled pause when loosing attention during an ongoing *FoD*, as in the previous study. This is initiated whenever the attention monitoring module recognizes inattention based on engagement errors. To this end, the synthesis module pauses the current speech output after the ongoing word (see section 7.3.3). The speech resumes when the attention monitoring module recognize attention again. For a repetition, the synthesis module pauses the current speech output and change the speech plan by adding the repetition of the highlight strategy.

10.2.3 *Evaluation*

This *HAI* scenario was tested in an interaction study.

Study Design

The experimental procedure is similar to the first pilot study and follows the general study design depicted in fig. 9.4. I conducted a between-subject human-agent interaction study. After signing a consent form, the subjects are led to the experiment room. They enter the room alone, only with the instruction to go into the kitchen, look at the agent, listen carefully and fill out a questionnaire on the computer afterwards. The interaction starts as soon as the participant stand in front of Flobi. The disturbances are triggered at predefined points of the interaction. At the end of the interaction, the participants go to the table and filled out the questionnaire on a computer. In contrast to the first pilot-study, this interaction is fully autonomous.

I designed three audio-visual external distractions in the apartment to distract the user's attention from the system. The first is visual and achieved through blinking lights, the second consists of a sound played in the apartment, while the third was achieved by the experimenter assistant re-entering the room. All disruptions happen at the same three points of the interaction and are not randomized, to get comparable results between the participants.

The questionnaire consists of three parts: a memory task, subjective ratings about Flobi, and demographics. The memory task consists of ten multiple choice questions about the information stated by Flobi during the *information phase*. In contrast to *EC1*, this consists not of simple yes or no questions, but of a choice of four possible answers (including "I don't know."). Four of the ten questions address information which was not only verbally described by Flobi, but additionally presented *embodied* via a corresponding actor in the apartment. In the second part, the participants had to provide subjective ratings of the agent to evaluate five key concepts in *HRI*: anthropomorphism, animacy, likability, perceived intelligence, and perceived safety. In the last part, they have the opportunity to leaf general comments at the end of the questionnaire and demographics and previous experiences assessed. The experiment ends with the *debriefing*.

Participants

Participants were recruited from the campus of the Bielefeld University and mostly students or from the administrative staff. They are between 19 and 40 years old. I recorded 30 trials with 14 female and 16 male participants in total. 15 participants were in the baseline (six female, 9 male; $M_{age} = 24.73$, $SD_{age} = 5.22$) and 15 in the *AHM* condition (8 female, 7 male; $M_{age} = 25.47$, $SD_{age} = 3.40$). There is no statistical difference between the ages of the two groups, $W = 136.5, p = .326$.

The participants are balanced regarding their average prior experience with technical intelligent systems ($M_{AHM} = 2.85$, $SD_{AHM} = 0.83$;$M_{BASE} = 3.07$, $SD_{BASE} = 0.75$), $W = 118.5, p = .815$. Participants in both conditions have no or very little experience with robotic systems in general, the virtual agent Flobi or programming. They have little experience with speech systems in general. All participants are very experienced with the use of computers in general.

Task Performance Hypothesis

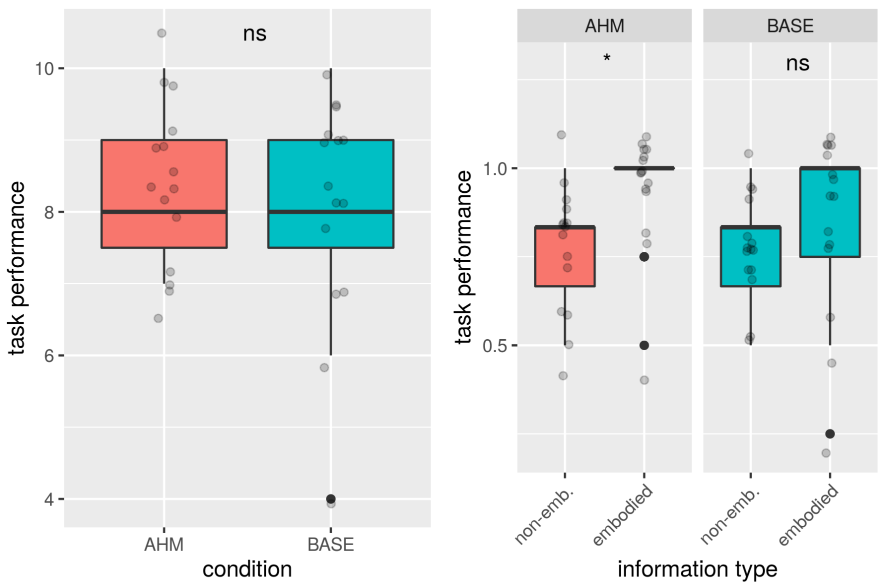


Figure 10.11: The task performance in total (left) and divided for *embodied* and *non-embodied* information (right).

The assumptions for a parametric test are not met, therefore the non-parametric *WR-Test* is applied. Figure 10.11 visualized the total task performance (TP_{All}) in each condition (left) and divided for *embodied*($TP_{embodied}$) and *non-embodied* ($TP_{non-embodied}$) information (right). The participants did not achieve significantly different scores for $\mathrm{TP}_{\mathrm{All}}$ between *AHM* ($M = 8.40, SD = 1.12$) and the *BASE* condition ($M = 8.00, SD = 1.51$); $W = 123.5, p = .652$. Thus, the hypothesis of positive memory effect of self-interruption cannot be confirmed. In general, the relatively high degree of task performance indicate that the difficulty of the questions was appropriate.

For exploration purpose, the task performance divided for *embodied* ($\mathrm{TP}_{\mathrm{embodied}}$) and *non-embodied* ($\mathrm{TP}_{\mathrm{non-embodied}}$) information is shown on the right of Figure 10.11. The *WSR-Test* with Bonferroni correction shows a significant effect of the information type in the *AHM* condition only. In the *AHM* condition, the *embodied* information was recalled significantly better ($M = 0.93, SD = 0.15$) than the *non-embodied* information ($M = 0.78, SD = 0.16$); $V = 10, p_{adj} = .048, r - 0.51$. Interestingly, in the *BASE* condition, the difference between *embodied* ($M = 0.83, SD = 0.24$) and *non-embodied* information ($M = 0.78, SD = 0.16$) was not statistically significant; $V = 40.5, p_{adj} = .924$.

Side effects on the Interaction

Next, I investigate the side effects on the interaction, regarding the subjective ratings of the agent and the *VFoA* of the human interaction partner.

SUBJECTIVE RATINGS Figure 10.12 visualizes the subjective ratings of the five key concepts between the baseline and the condition *AHM*. As is the previous study, the agent Flobi receives high

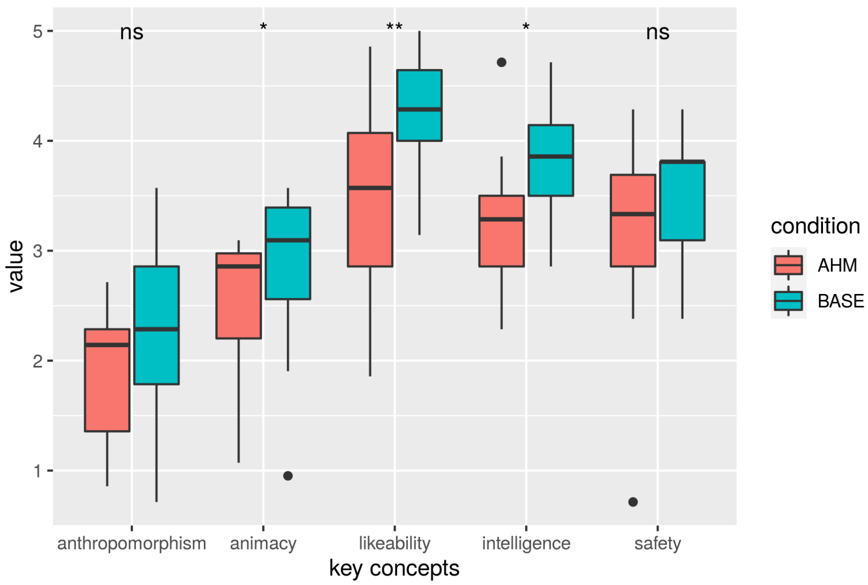


Figure 10.12: The subjective ratings of the agent.

values for the key concepts *likability* and *perceived intelligence* and *perceived safety* ($M > 3.1$) and rather low values for anthropomorphism ($M < 2.3$) in both conditions. The two sample Welch t-test suggests that the self-interrupting agent has an effect on *likability* and the *perceived safety* of the agent. The agent in the *AHM* condition is rated significantly less likable ($M = 3.44, SD = 0.87$) than the agent in the *BASE* condition ($M4.29, SD = 0.58$); $t(24.46) = -3.14, p = .004, CI = [-1.40, -0.29], d = -1.15$. In addition, it is also rated less intelligent in the *AHM* condition ($M = 3.27, SD = 0.59$) than the agent in the *BASE* condition ($M = 3.82, SD = 0.57$); $t(27.96) = -2.59, p = .015, CI = [-0.99, -0.11], d = -0.95$. For the other key values, I perform the Wilcoxon test because the assumptions of parametric tests are not met. Flobi's animacy was rated slightly lower in the *AHM* ($M = 2.47, SD = 0.69$) than in the *BASE* condition ($M = 2.86, SD = 0.74$); $W = 64.5, p = .048, r = -0.37$. The key concepts *anthropomorphism* and *perceived safety* do not differ significantly (anthropomorphism $W = 89.5, p = .34$; safety $W = 84, p = .24$)[1].

In the post-hoc questionnaire, the participants had the opportunity to leave general comments. In the baseline, four participants give negative feedback concerning the voice quality, as can be seen in example 10.2.3. They perceive the voice as monotonous and choppy.

Example 10.2.3: Example comment by participant in the *BASE* condition

VP04: *"[...] durch die fehlende bzw wenige Betonung schnell monoton"*

VP04: "[...] due to the lack of or little emphasis, its monotonous"

In the *AHM* condition, five participants give negative feedback concerning the voice quality, as can be seen in example 10.2.4. The critique was mostly about the self-interrupting behavior, such as the statement by participant VP09. One participant perceived the repetitions as rude (stated in the debriefing). Two additional participants noted the adaptive behavior of Flobi without any judgment, e.g., in the comment by VP22.

Example 10.2.4: Example comment by participant in the *AHM* condition

VP09: *"Sprachausgabe hat teilweise arg 'gehackt'. Längere Pausen mitten im Satz zB."*

VP22: *"Flobi sprach nur bei 'Augenkontakt', außer Klobi [sic] forderte zum zur Seite schauen auf"*

VP09: "Speech output has 'chopped' badly sometimes. Long pauses in the middle of a sentence, for example"

1 Test results can be found in table G.5 and exact values in table G.4.

> VP22: "Flobi only spoke when there was 'eye contact', unless Klobi [sic] asked to look to the side"

VISUAL FOCUS OF ATTENTION A closer look at the *VFoA* of the participants is taken, to investigate if participants in the *AHM* condition are less inattentive than participants in the *BASE* condition. Figure 10.13 visualizes the distribution of the *VFoA* of the different con-

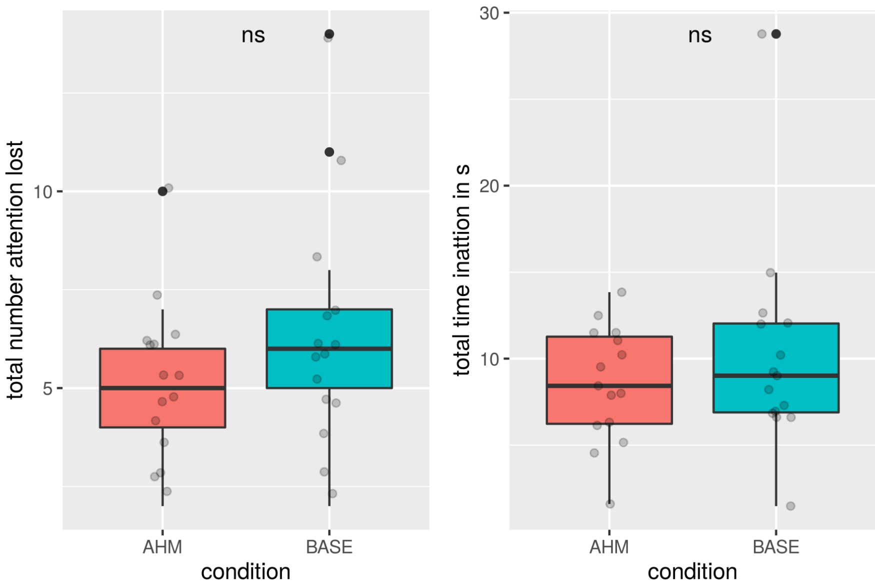


Figure 10.13: Distribution of inattentiveness: (left) shows the total number of look away NA_{number}. (right) visualizes the total time participant did not look at the agent or the current *FoD*

ditions. The left graph shows the number of look away (NA_{number}). Whenever the user look away from the agent or the current *FoD*, this number increase by one. The assumptions for a *MANOVA* is not met, therefore the non-parametric tests with Bonferroni correction are applied. To test, whether participants in the *BASE* condition look away more often than in the *AHM* condition, a *T-Test* is performed. The number of look aways in the *BASE* condition ($M = 6.33$, $SD = 0.85$) does not differ significantly from the participants in the *AHM* condition ($M = 5.13$, $SD = 2.99$); $t(23.88) = -1.31, p_{adj} = .408, CI = [-3.09, 0.69], d = -0.48$. Participants in the control group do not look away significantly more often. Furthermore, there is no significant difference in the overall time people look away from the agent or the current *FoD* between the *AHM* ($M = 8.5s, SD = 3.4$) and the *BASE* ($M = 10.2s, SD = 6.1$) condition, $W = 98, p_{adj} = .924, r = -0.71$.

10.2.4 *Discussion and Lessons Learned*

In this section, I investigated how my *AHM* can be implemented in an autonomous system. I demonstrated that three things can be implemented and integrated: (1) a multimodal attention guiding, (2) an attention monitoring system, and (3) intervention strategies to react on inattentive interaction partner. Everything together forms the implementation of my *AHM*. Thereby, I investigated how my *AHM* need to be adapted to deal with different *FoDs*. This manifests itself in two ways. At first, multimodal attention guiding possibilities were enabled. Besides human-like behaviors, such as pointing, smart-home specific attention guiding capabilities are integrated, e.g., the use of lights. Second, the dialogue history was included into my *AHM*. This allows the model to differentiate between an ongoing topic and an *FoD* change. Thus, different hesitation strategies as intervention strategies are possible. On the one hand, the agent can with unfilled pauses as re-attention strategy and repetitions to emphasize the current discourse change.

Furthermore, I conducted an interaction study with the resulting system. In contrast to the first study, this system worked fully autonomously. This experiment had not the aim to evaluate the quality of the individual modules, but rather to demonstrate that it is possible to implement such an autonomous system and to further investigate my research hypothesis. Nevertheless, we can draw a few conclusions about the quality of the system from the results. The multi-modal attention guiding worked, which can be seen in (1) the low error rate in the task performance, i.e., the post interaction memory recall in both conditions and (2) the low values participants are inattentive. In addition, the *embodied* information results in less memory errors in the *AHM* but not in the *BASE* condition. This could be due to two reasons. Firstly, an additional modality was used for communication. But this was the case in both conditions, so that cannot be the only reason. In addition, the agent performs in the *AHM* condition the two hesitation strategies to react on and guide the user's attention. A better memory recall for the *embodied* information in the *AHM* condition is therefore plausible, as in the *embodied* information case the *VFoA* is an important requirement to gather this relevant additional information. However, this needs to be further investigated.

Beside the fact, that *embodied* information is recalled better in the *AHM* condition than *non-embodied* information, no significant effect on the general task performance between the two conditions could be found. Furthermore, the positive effect of the *AHM* on the participants looking behavior found in 10.1 could not be confirmed in this interaction study. This could have several reasons. Firstly, in could depend on the interaction itself. This effect may only occur in simple interactions without any discourse changes. More complex interactions could be itself demands more attention from the participants. The different discourse changes could attract the attentiveness of participants. Another explanation could be, that the *AHM* changed. Both, the moment of intervention as well as the intervention itself changed. Lastly, the implementation could be not accurate enough, e.g., due to detection errors.

Unfortunately, the agent with the *AHM* perceived lower values for *likability*, *animacy* and *perceived intelligence*. In the debriefing, it turned out, that the number of repetitions of the highlight strategy has a big impression on the participants, which may influence the subjective ratings of the agent. This is also reflected by some negative comments in the questionnaire. In this interaction, Flobi repeated the highlight strategy so often until the participant looked at the new *FoD*. This behavior could have a big impact on the subjective rating of the agent's likeability, especially if the participant does not look – or even want to look – at the new *FoD* and the agent tries to enforce the visual attention repeatedly. The consequence for the attention model is that it should include a threshold for insistence on achieving the desired user attention. Note that this threshold can be learned, and can be dependent on user preferences as well as on information type and relevance. In addition, the validity of the modules should be further investigated. Errors in the tracking could lead to false-positive triggering of the repetitions. Furthermore, the participants give negative feedback concerning the voice quality. Therefore, a more adaptable hesitation strategies may can provide a more variable prosody to counteract a perceived rudeness of self-interruptions.

Compared to the first pilot study, I changed the study design in the following ways. At first, I moved the interaction from the wardrobe to the kitchen. This allows to talk about different foci of discourses. Furthermore, the interaction length was slightly extended. In this way, participants had more time to be inattentive. To increase this possibility further, the number of disruptions is also increased. The results of the *VFoA* shows, that this change of the experiment design was successful. However, the interaction length can expand further.

In addition, I changed the measurement of the task performance, by adding the answer "I don't know". This may improve the quality of task performance measurement. Additional, the questionnaire were appropriate, which can be seen by the high task performance. However, a better way of measuring the task performance should be considered.

To sum up, this section shows that it is possible to implement an autonomous *AHM*, even if the individual modules can be improved.

10.3 EC 3: EXPLORATION OF A PRACTICAL TASK DURING INTERACTION

In the previous sections, the interaction consisted off one or more monologues of the agent without the possibility to directly perform the task during the interaction. Instead, participants had to perform the task–recall the presented information—after the interaction in the questionnaire. In this section, I explore a more practical task and investigate how such an interactive scenario can be formalized and modeled. Therefore, a scenario in which the interactive agent supports a user in an ongoing practical task in the smart environment is presented. Cooking poses a suitable interactive interaction, which is highly multi-modal and situated.

The research questions in this section target two areas. From the user perspective it is important to know what strategies for timing the verbal instructions regarding the physical actions should be used (1) and from the perspective of the system architecture how both the dialogue and the incremental task can be modeled and systematically coordinated (2). An initial discussion of the resulting model is published [CCW17a].

In the following, I first present the practical task, and its current implementation in the *CSRA*. Afterward, a dataset is presented with analyses regarding the timing of information presentation. Based on this, I draw conclusions for a possible *AHM* to present information incrementally and evaluate it in an interaction study.

10.3.1 *Practical Task: Preparation Phase of Cooking*

One interaction scenario in the *CSRA* is cooking. In this *EC* we are focussing on the preparation phase of cooking. Example 10.3.1 shows an excerpt of the instruction for the preparation phase of cake baking. The human wants to prepare necessary ingredients with the help of Flobi. Thereby, s/he asks which ingredients are needed from different locations within in the apartment's kitchen.

Example 10.3.1: Example instructions for the preparation phase of cooking.

H:	*What do I need to bake the cake?*
A:	For the cake, we need ingredients from four locations: the fridge, the cupboard near the microwave, the cupboard to my right and the cutlery drawer.
H:	*What do I need from the fridge?*
A:	Out of the fridge, we need milk and cream, eggs, butter, margarine, cheese and lemon.
H:	*What do I need from the cupboard near the microwave?*
A:	From the cupboard near the microwave, we need the big mixing bowl, the small mixing bowl, muffin tin, dessert bowls, parchment paper, the mixer and salt.
H:	*What do I need from the cupboard to your right?*
A:	From the cupboard to your right, we need flour, sugar, baking powder, vanilla sugar, cocoa, chocolate and the sugar pearls.
H:	*What do I need from the cutlery drawer?*
A:	From the cutlery drawer, we need the tablespoon, teaspoon, stir bars, scissors, scraper, whisk and the cutlery knife.
H:	*Anything else?*
A:	That's it! Now let's make a cake.

Current Implementation

The current scenario is implemented with *Pattern Based Mixed Initiative Interaction Toolkit (Pamini)*. For the design of this interaction, mostly the *Human Information Request* interaction pattern is used, which is depicted in fig. 10.14. The human can ask Flobi what kind of ingre-

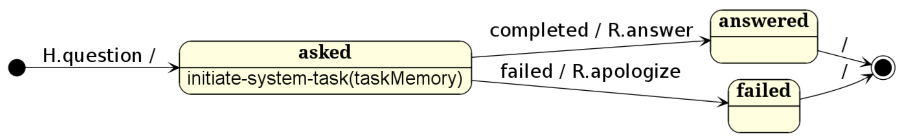


Figure 10.14: The *Human Information Request* interaction pattern of Pamini.

dients are needed from the special locations. Example 10.3.2 depicts an excerpt of the preparation phase, configured with four *Human Information Request* interaction patterns.

Example 10.3.2: Interaction patterns of the preparation phase of cooking.

1	***H:***	*Information request*	*What do I need to bake the cake?*
	A:	*Answer*	For the cake, we need ingredients from four locations: the fridge, the cupboard near the microwave, the cupboard to my right and the cutlery drawer.
2	***H:***	*Information request*	*What do I need from the fridge?*
	A:	*Answer*	Out of the fridge, we need milk and cream, eggs, butter, margarine, cheese and lemon.
		...	
3	***H:***	*Information request*	*What do I need from the cutlery drawer?*
	A:	*Answer*	From the cutlery drawer, we need the tablespoon, teaspoon, stir bars, scissors, scraper, whisk and the cutlery knife.
4	***H:***	*Information request*	*Anything else?*
	A:	*Answer*	That's it! Now let's make a cake.

In the following, I present a dataset, containing six *Human-Human Interaction (HHI)* of the preparation phase of this cooking scenario.

HHI Dataset

In a master thesis supervised by me, Chromik recorded a dataset about the preparation phase of a cooking scenario in *HHI* [Chr16]. The recording took place in the *CSRA*, with the goal to figure out how humans structure the information in this preparation phase. In total, 12 subjects (9m, 3f) took part in six teams **T1-T6** of two subjects interacting with each other per trial. **T1**, **T4**, and **T6** knew each other, the other teams were totally unfamiliar to each other before the study.

The task for the participants was to search all necessary ingredients for backing a cake in the kitchen. One participant per team had a list of ingredients (*reader*), while the other had to fetch these objects (*fetcher*). The *fetcher* should stay in the kitchen, whereas the other should stay in front of the kitchenette. They could both see and hear each other from their positions The list was divided into four different kitchen locations (i.e., cupboards and drawers) and each location contained seven objects. In total, there were 28 baking ingredients and accessories. No further instructions were given on how to coordinate this task. Figure 10.15 visualizes the various kitchen locations and the

Figure 10.15: Experimental setup in EC3. Upper: Different kitchen locations used in this study. Lower: Ground view of the apartment with the position of *reader* and *fetcher*.

positions of the two participants (*reader* and *fetcher*).

The recorded dataset gives some insights into the coordination of such a cooperative, practical task. The teams have shown a set of strategies for solving the task, but also some similarities within these

Team	Location	Chunking	Feedback
T1 (familiar)	Fridge	1,1,1,1,1,1,1	fetched each item separately and put it down in front of the reader
	Cabinet l	2,-,2,-,1,2,-	
	Cabinet r	1,1,1,1,1,1,1	
	Drawer	2,-,1,1,1,1,1	
T2 (unfam.)	Fridge	2,-,2,-,1,2,-	verbal acknowledge (ger: "ja","ok")
	Cabinet l	2,-,1,1,1,2,-	
	Cabinet r	2,-,2,-,1,2,-	
	Drawer	2,-,1,1,1,2,-	
T3 (unfam.)	Fridge	1,1,3,-,-,2,-	verbal repetition of each item gaze acknowledge (look at reader)
	Cabinet l	2,-,3,-,-,2,-	
	Cabinet r	3,-,-,3,-,-,1	
	Drawer	2,-,1,1,1,2,-	
T4 (familiar)	Fridge	1,1,1,1,1,1,1	verbal repetition of each item verbal acknowledge (ger: "ja","ok")
	Cabinet l	2,-,2,-,1,2,-	
	Cabinet r	3,-,-,1,3,-,-	
	Drawer	2,-,1,1,1,1,1	
T5 (unfam.)	Fridge	1,2,-,1,1,1,1	verbal acknowledge (ger: "ja")
	Cabinet l	1,1,1,1,1,2,-	
	Cabinet r	5,-,-,-,-,2,-	
	Drawer	7,-,-,-,-,-,-	
T6 (familiar)	Fridge	2,-,1,1,1,2,-	verbal acknowledge (ger: "jawohl")
	Cabinet l	1,1,2,-,1,1,1	
	Cabinet r	1,1,1,1,2,-,1	
	Drawer	2,-,1,2,-,1,1	

Table 10.1: Chunking and feedback behavior in the human-human interactions of the fetch task.

strategies (depicted in table 10.1). I take a closer look at how the *reader* divides the task into subtasks (chunking) and what timing they choose for presenting the next information chunk. Additionally, I do a first analysis of the feedback signals of the *fetcher*.

DIVIDING THE TASK INTO SUBTASKS: For the dividing of the task, all participants worked through the kitchen locations one after another. For each location, the *reader* started with a brief introduction, providing information about the relevant location. This was followed by a listing of the ingredients, the *reader* again split into sets. One exception to this approach was used by **T4**. In this case, the *reader* started the first location with the item "milk" without the location. This may be attributed to the *readers'* expectation that this information is not necessary. Milk can be usually found in the fridge.

Table 10.1 shows the chosen chunking of the information for each team. A frequently chosen strategy was grouping the first and last two objects of each location and requesting the objects in between separately (i.e., 2 1 1 1 2). One reason for grouping the first two objects could be, that they often categorically matched (e.g., a *big bowl* and a *small bowl*). However, the last two items are also partly grouped together, even if these objects categorically did not belong together (e.g., *blender* and *salt*). A different explanation for this approach could be to signalize the completion of the fetching of ingredients from the current location.

TIMING FOR THE NEXT INFORMATION CHUNK: Besides the chunking of the information the timing plays an important role in such a cooperative interaction scenario. The teams found different techniques to choose an appropriate moment for the presentation of the next information chunk. Most *fetchers* looked at the corresponding kitchen location, while the *reader* stated the required items. The *reader* from team **T1**, **T3**, and **T5** presented the next information as soon as their partner had put down the current object. In team **T2** and **T4**, the next chunk was presented as soon as the current object was fetched. *Reader*$_{T6}$ mostly presented the next chunk only after *fetcher*$_{T6}$ gave a verbal feedback.

FEEDBACK FROM THE FETCHER: All *fetchers* provided feedback to the reader. *Fetcher*$_{T3}$ and *fetcher*$_{T4}$ repeated each item verbally. *Fetcher*$_{T1}$ and *fetcher*$_{T3}$ brought each object to their *reader* individually and put it down in front of them. All *fetchers*—except in **T1**—gave some sort of verbal acknowledgement after fetching the current item(s), usually by saying "yes" or "ok". Additionally, *fetcher*$_{T3}$ looked at the *reader* after each object. *Fetcher*$_{T1}$ and *fetcher*$_{T5}$ asked for repetitions a few times.

RESULTING DESIRED INTERACTION: Based on the insights, posed by the analysis of the *HHI*, a model for information chunking and timing in a similar *HAI* scenario can be created. The following example dialogue excerpt presents a desired interaction between an agent the human for the dialogue act **R.answer** of the *Human Information Request* interaction pattern.

Example 10.3.3: Desired interaction for the dialogue act R.answer of one *Human Information Request* interaction pattern.

Nr.	*Verbal Actions*		*Nonverbal Actions*
1	**A:**	*From the fridge we need milk and cream...*	
2			Human goes to the fridge and finds the ingredients.
3	**H:**	*yes, I have them.*	
4	**A:**	*eggs*	
5			Human fetches cream and eggs and places milk, cream, and eggs in front of the agent.
6	**A:**	*butter*	
7			Human finds the butter and looks at the agent.
8	**A:**	*margarine*	
9	**H:**	*What did you say?*	
10	**A:**	*margarine*	
11			Human nods, fetches the butter and margarine and brings both to the agent.
12	**A:**	*cheese and lemon.*	
13	**H:**	*Cheese and lemon, let's see...*	
14			Human fetches the remaining ingredients.
15	**H:**	*Done!*	
16	**A:**	*That's it! Now let's make a cake.*	

Currently, it is not intended to model such a cooperative instruction with *Pamini*. The single dialogue act—the answer of what kind of objects are located in the fridge—is presented at once. However, with the *AHM*, the system speech can be coordinated with the human attention for this dialogue act and make the information presentation incrementally.

10.3.2 *Attention-Hesitation Dialogue Coordination Model*

The desired incremental information presentation of the dialogue act **R.answer** of one *Human Information Request* interaction pattern can be realized with my *AHM*. In contrast to the previous models, the starting point of the hesitation strategy—the delay of the presentation of the next information chunk—is predefined by the task description and not initiated by the human behavior.

WHEN TO (RE-)ACT: ATTENTION CONCEPT In contrast to the previous *AHMs*, the concept of attention is different in this interaction. Attentive is interpreted as the readiness for the execution of the next (sub-)task, as depicted in fig. 10.16. The interaction partner is inatten-

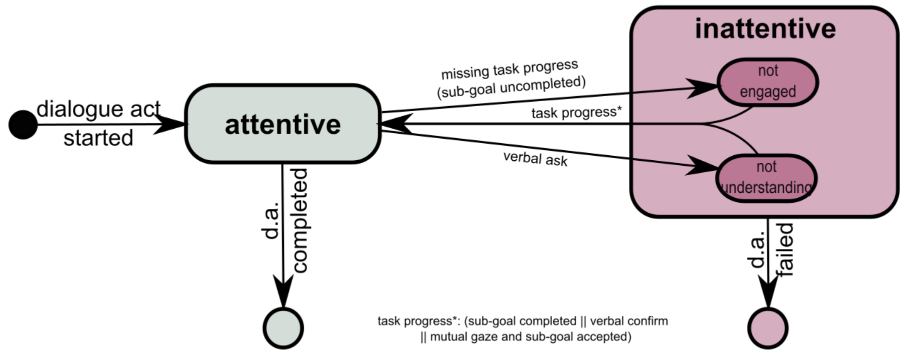


Figure 10.16: Concept of attention for EC3.

tive when task progress is missing, i.e. the current sub-goal of the previously presented subtask is not completed. In this model, this is interpreted as an engagement problem in the task—the interaction partner has not (yet) fulfilled the task. Of course, it can also rely on understanding problems. In this model, such highlight strategies to repair this state are only initiated after a verbal response indicating understanding problems, such as "What did you say?". Further features are of course conceivable, such as monitoring the task progress and detecting errors in their execution, and are further discussed in section 10.3.5.

HOW TO (RE-)ACT: HESITATION INTERVENTION STRATEGIES Whenever the agent looses the attention of the human interaction partner, it reacts with one of the hesitation strategies visualized in fig. 10.17. The hesitation strategies are the same interaction strategies presented in *EC2*. Whenever an understanding problem occurs, the agent reacts with (a) a repetition of the multi-modal guiding strategy as an attention highlight strategy, more precisely a repetition of the last information chunk. As soon as the task progress is missing, meaning an engagement problem in the task occurs, the agent uses (b) unfilled pauses—simple self-interruptions—as re-attention strategy.

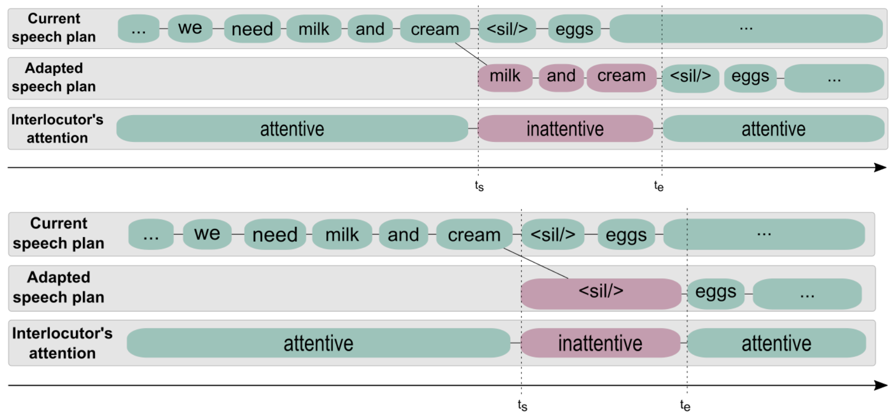


Figure 10.17: Hesitation intervention strategies for EC3: (a) highlight (b) re-attention.

10.3.3 *Implementation*

The model is formalized as an extended form of a finite state machine, augmented with internal state actions. This approach is a variant of the concept of interaction patterns, I discussed in section 6.1.3. *Pamini* has the concepts of task to deal with time intensive system actions and can provide information about the current state of these actions. However, the concept of task is only applied for system actions. In this scenario, the search for the corresponding ingredients is not a system action, but rather a human action. Therefore, a special interaction pattern is designed, dealing with the concept of such human tasks, visualized in fig. 10.18. This interaction pattern is initiated by the agent and has to be configured for the specific interaction scenario. The main idea is to have two levels of coordination, which can be configured separately but influence each other. I divided the pattern into a broad *dialogue level* and a fine *task level* coordination. The *dialogue level* controls the main interaction flow and depends on the achievement or non-achievement of the sub-goals defined in the *task level*.

The interaction pattern has multiple phases. The first phase is the preparation phase. In this phase, the agent can introduce the task. After an optional verbal confirmation, the state changes from initial to prepare and the *Dialogue Management (DM)* initiates the monitor task. This system task is responsible for monitoring the human's task progress. Additionally, during this phase, the first subtask can be presented by the agent. In the pattern, the task information is specified as a list with the corresponding chunking of the information for each subtask. The agent hesitates as re-attention strategy between these chunks of information.

In the nextInfo state, the human should perform the current subtask. Simultaneously, the agent observes the human and provide additional

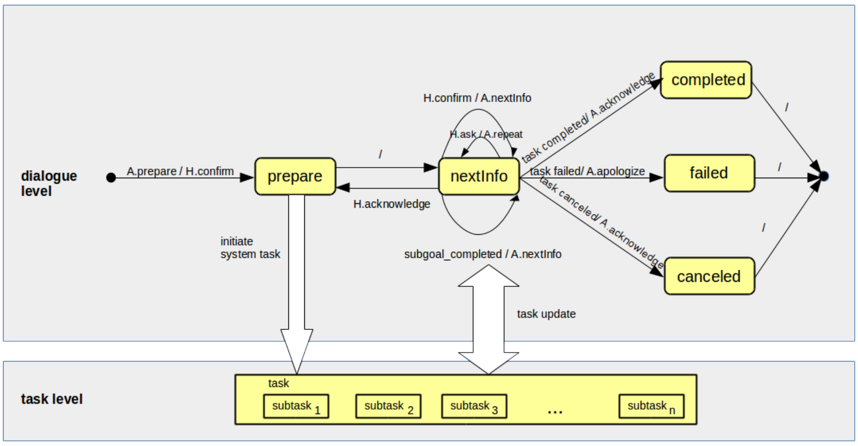


Figure 10.18: Interaction model for incremental information presentation and task representations with different subtasks. The achievements of sub-goals influence the overlying interaction model

feedback when necessary or presents the next information chunk. In the current configuration, the agent repeats the previous stated chunk of information as attention highlight strategy. The model distinguishes between dialogue and task level in this state. In the *dialogue level* the human can ask for repetitions or verbally confirm the current completion of the subtask, resulting in the agents presenting of the next subtask.

Apart from the explicit verbal confirmation by the human interaction partner in the dialogue level, the presentation of the next information chunk can also be triggered from the *task level*. The system task monitors the human task progress and is responsible for the timing strategy of the presentation of the next chunk in addition to the demand of the human.

Reminding the example 10.3.3, the same dialogue can be formalized as follows:

Example 10.3.4: Desired interaction with corresponding levels and events.

Nr.	*Level*	*Event*	*Verbal Actions*		*Nonverbal Actions*
1	*dialogue*	*A.prepare*	**A:**	*From the fridge we need milk and cream...*	
2	*task*	*task update*			Human goes to the fridge and finds the ingredients.
3	*dialogue*	*H.confirm*	**H:**	*yes, I have them.*	
4	*dialogue*	*A.nextInfo*	**A:**	*eggs*	
5	*task*	*sub-goal complete*			Human fetches cream and eggs and places milk, cream, and eggs in front of the agent.
6	*dialogue*	*A.nextInfo*	**A:**	*butter*	
7	*dialogue*	*sub-goal complete*			Human finds the butter and looks at the agent.
8	*dialogue*	*A.nextInfo*	**A:**	*margarine*	
9	*dialogue*	*H.ask*	**H:**	*What did you say?*	
10	*dialogue*	*A.repeat*	**A:**	*margarine*	

11	*task*	*sub-goal complete*			Human nods, fetches the butter and margarine and brings both to the agent.
12	*dialogue*	*A.nextInfo*	**A:**	*cheese and lemon.*	
13	*dialogue*	*H.acknowledge*	**H:**	*Cheese and lemon, let's see...*	
14	*task*	*sub-goal complete*			Human fetches the remaining ingredients.
15	*dialogue*	*H.confirm*	**H:**	*Done!*	
16	*dialogue*	*A.acknowledge*	**A:**	*That's it! Now let's make a cake.*	

The example 10.3.4 visualizes the dialogue act R.answer as an incremental information presentation with the corresponding events initiating the next information chunk. Line 3 depicts a verbal confirmation on the dialogue level, whereas in line 5, 11, and 14 the monitoring task recognize a sub-goal completion. In line 7, the human signals its readiness for the next information chunk through initiating mutual gaze.

TIMING STRATEGIES: Optimally, the agent should have the necessary sensors and capabilities to figure out the right moment for presenting the next information chunk. However, the necessary sensors or capabilities are not always available. In these cases, simplifications can be made. I propose the following trigger strategies to recognize the moment when the interaction partner is attentive again to hear the next chunk of information:

Dialogue level:

- *verbal:*
- *non-verbal:* monitor the human interaction partner and observe non-verbal turn-taking cues, e.g., mutual gaze or head nodding.

Task level:

– *sub-goal completion:* monitor the human interaction partner and observe the progress of each subtask to recognize sub-goal completion.
– *sub-goal failure:* monitor the human interaction partner and observe the progress of each subtask to recognize sub-goal failures.
– *task progress fallback:* use learned average or configured durations for each subtask. Adapt the timing between the information chunks based on (non-)verbal acknowledgement in previous interactions.

Of course, also combinations of these strategies are possible. If it is not possible to monitor the task progress correctly, the agent has the possibility to rely on verbal or non-verbal features, or to use a *fallback* strategy. For the *task-progress* strategy, I implemented an approach that monitors the opening and closing state of the cupboards and drawers, whereas the *non-verbal* strategy presents the next information chunk after mutual gaze is detected. The *fallback* strategy is configured with a timeout. To this end, the average durations from the previous interactions were used.

10.3.4 *Evaluation*

Based on the observations gathered in the *HHI* dataset, I formulate the model for incremental information presentation, based on the attention of the human interaction partner. To examine whether such an incremental information presentation is helpful or not, we carried out a *WoZ HAI* experiment with a between-subject design. There, we investigated a first model of incremental information presentation. The interaction toke place in the kitchen area of the *CSRA*, and as interaction partner the virtual agent Flobi is used. Some results of this investigation are published in [CCW17b].

Study Design

For the evaluation, we compared two conditions. In the *AHM* condition, the items for each location are presented with the chunking and timing observed in the *HHI* dataset. More precisely, the agent groups the first and last two items together and presents the objects in between separately (2 1 1 1 2). Furthermore, Flobi presents the next information chunk either upon an attentive signal from the participant (mutual gaze or verbal feedback) or when the agent could draw conclusions to the attention based on reaching a sub-goal. This is defined as the moment when an object is put down. In the *BASE* condition, the agent present the information about all seven objects per location at once without waiting of feedback from the user (verbal, non-verbal, or task progress). Participants had to fetch the same 28 backing ingredients and accessories from the same locations as in the *HHI* setup, visualized in fig. 10.15. Figure 10.19 depicts the experimental setup in this study. The human can ask for repetition in both conditions. In the

Figure 10.19: Experimental setup. Upper: Person interacting with Flobi. Lower: Ground view of the apartment with the position of participant and the agent.

BASE condition, the agent repeated all items from the current location, whereas it only repeated the current chunk of information in the *AHM* condition. After the fetching, the participants are asked to clean up again and put the ingredients back to their locations. Thereby, they can ask where the ingredients belong to. To identify the correct cabinet or drawer, the corresponding handle light up in both conditions.

The task performance is measured to evaluate whether these differences have an effect on the interaction. To this end, we counted the errors during the object fetching. An error can be the missing of an object, the fetching of a wrong object, or the asking for repetition. In addition, we measured the errors in the cleaning phase, by annotating whether the objects were put back into the correct locations. The error rate for the cleaning is normalized with the actual number of objects, fetched by the corresponding participant. In contrast to the previous studies, I do not assess the visual attention in this experiment. Since the participants move around a lot in the apartment during the task, a continuous measurement is difficult.

To access the memory capacity of the participants, a pre-test is conducted in the *briefing*. To this end, participants listened to a pre-constructed audio file, containing ten words produced by a synthetic voice, i.e., a Mary TTS's German female *Hidden Markov Model (HMM)* voice without further modification. Each word depicts a profession, food, sport, building, or city. In total two words per category. After the participants listened to the audio file, they are requested to repeat many of the words as they remember. The predefined verbalization of Flobi are triggered by the wizard, pressing a corresponding button in the control room.

Participants

In total, 30 subjects took part in the study. Two participants had to be excluded because of data loss. From the remaining 28 participants, 15 participants (7 male, 8 female; $M_{age} = 22.93$, $SD_{age} = 2.87$) were in the *AHM* and 13 (6 male, 7 female; $M_{age} = 24.46$, $SD_{age} = 2.99$) in the *BASE* condition.

Furthermore, the participants are balanced regarding to their average prior experience with technical intelligent systems ($M_{AHM} = 2.4$, $SD_{AHM} = 0.81; M_{BASE} = 2.62$, $SD_{BASE} = 0.94$), $t(35.72) = -0.23, p = .817$. Participants in both condition have no or very little experience with programming, robotic systems in general, speech systems in general, or the virtual agent Flobi. However, all participants are very experienced with the use of computers.

Figure 10.20 visualizes the distribution of task performance in the memory pretest ($TP_{pretest}$) for each condition. The *T-Test* shows no significant differences in the mean memory performance in the pretest between the conditions ($M_{AHM} = 6.07$, $SD_{AHM} = 1.35; M_{BASE} = 6.92$, $SD_{BASE} = 1.43$), $t(24.95) = -1.63, p = .12$.

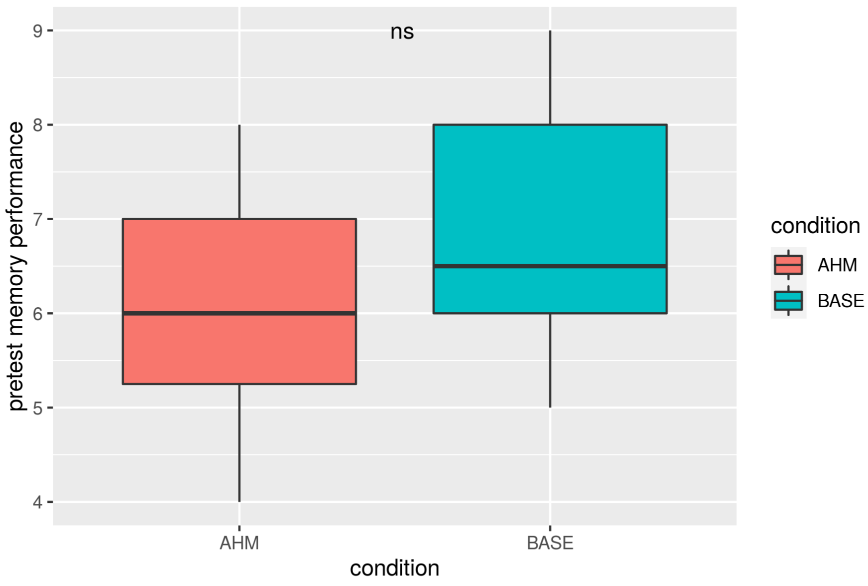


Figure 10.20: Results of the memory pretest for the *BASE* and *AHM* condition.

Task Performance Hypothesis

To evaluate the task performance, the amount of mistakes the *fetcher* made. Therefore, we counted the errors during the object fetching and annotated whether the objects were put back into the correct locations.

FETCHING PERFORMANCE: Figure 10.21 visualizes the errors during the object fetching task. The graph highlights that participants in the *AHM* condition ($M = 6.00$, $SD = 5.82$) made fewer errors than participants in the *BASE* condition ($M = 11.15$, $SD = 5.05$), $W = 42.5, p = .012, r = 0.48$. This indicates that such an incremental information presentation is useful in *HAI*, and therefore should be modeled and further investigated.

Figure 10.22 gives an overview of the different error types. As can be seen in the first row, the *AHM* condition results in fewer errors, than the *BASE* condition. The remaining rows depict the type of error. The

Error type	Condition	Mean	Median	SD
missing object	AHM	1.20	1	1.37
missing object	BASE	3.00	3	2.04
wrong object	AHM	1.07	1	1.10
wrong object	BASE	0.31	0	0.48
ask for repeat	AHM	1.33	1	1.63
ask for repeat	BASE	1.85	1	1.91

Table 10.2: Values for the different error types in EC3.

values in table 10.2 shows, that the number of missed objects is higher in the *BASE* condition. Interestingly, the number of wrong objects is sightly higher in the *AHM* condition. Furthermore, the participants in the *BASE* condition asked more questions and therefore heard the description more often. However, as can be seen in table 10.3 only the

Error type	group 1	group 2	p.adj	p.signif
missing object	BASE	AHM	.041	*
wrong object	BASE	AHM	.051	ns
ask for repeat	BASE	AHM	.460	ns

Table 10.3: Results of the Welch's t-test with Bonferroni correction.

difference for the missing object is significant.

CLEANING PERFORMANCE: The effect on the error rate in the fetch phase cannot be measured in the cleaning phase. Figure 10.23 visualizes the measurement for the cleaning phase. It can be seen that the presentation style has no significant effect on the post interaction recall of the object positions. Participants in the *AHM* condition ($M = 0.17$, $SD = 0.17$) made not more or less errors than participants in the *BASE* condition ($M = 0.28$, $SD = 0.53$), $t(14.10) = -0.73, p = .480$. In both conditions, a similar distribution of error points can be observed.

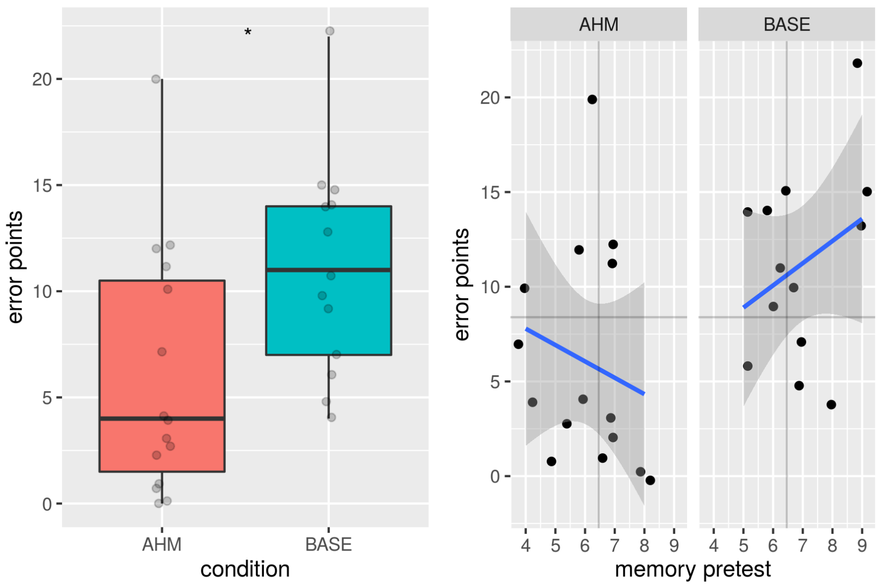


Figure 10.21: Errors during the object fetching task as a function of memory performance in the pretest. The horizontal line represents the median of the error points in total, whereas the vertical line visualizes the median of the memory performance of all participants in the pretest. The blue line representing the regression line with a 95% confidence interval[2].

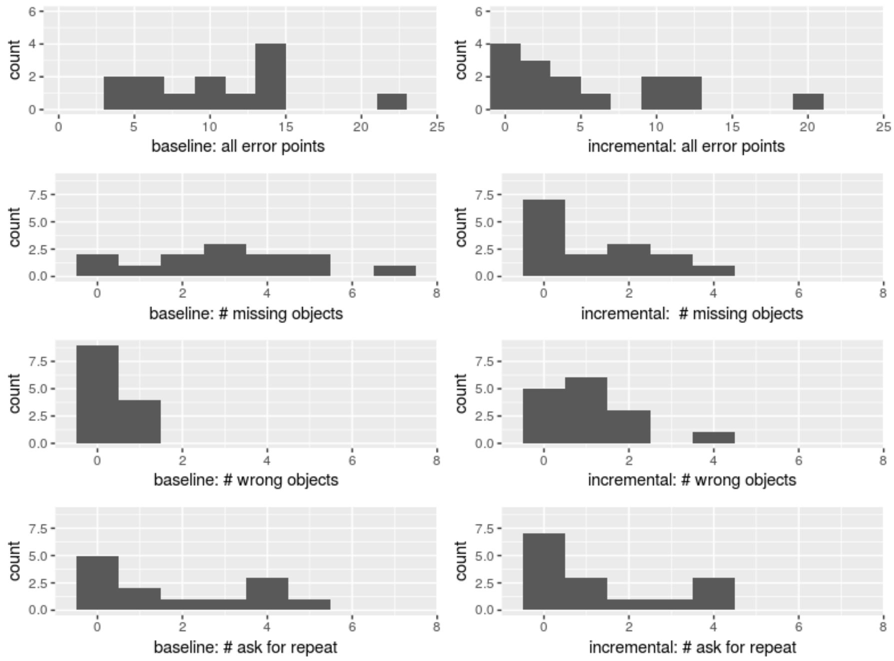


Figure 10.22: Total number of error points (first row) and the error points for each type of error (remaining rows) shown as a histogram over the participants for the baseline (left) and incremental *AHM* condition (right).

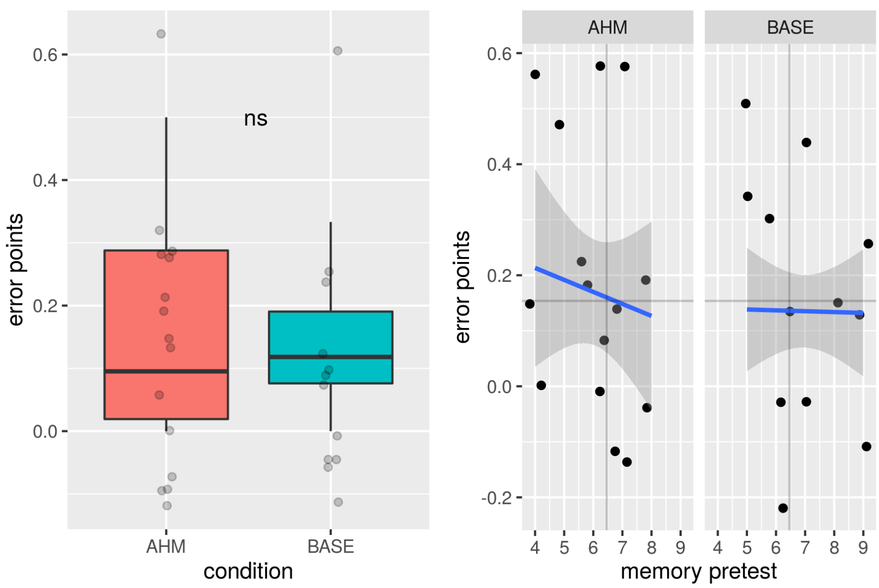


Figure 10.23: Normalized errors during the cleaning task as a function of memory performance in the pretest. The horizontal line represents the median of the error points in total, whereas the vertical line visualizes the median of the memory performance of all participants in the pretest. The blue line representing the regression line with a 95% confidence interval

Side effects on the Interaction

Next, I investigate the side effects on the interaction, regarding the interaction time and subjective ratings of the agent. The following table gives an overview of the timing. In the *AHM* condition, every

Location	Baseline	Inc. Condition
fridge	0:32 m (σ=27s)	0:25 m (σ=19s)
cabinet l	1:04 m (σ=15s)	0:59 m (σ=20s)
cabinet r	0:53 m (σ=15s)	0:51 m (σ=22s)
drawer	0:47 m (σ=13s)	0:45 m (σ=17s)
total	3:17 m (σ=57s)	3:01 m (σ=62s)

Table 10.4: Duration and standard deviations of the (sub-)tasks in the *AHM* and *BASE* condition.

subtask required less time on average than the baseline. However, the difference between the total duration is not significant, $t(14) = -0.73, p = .225$

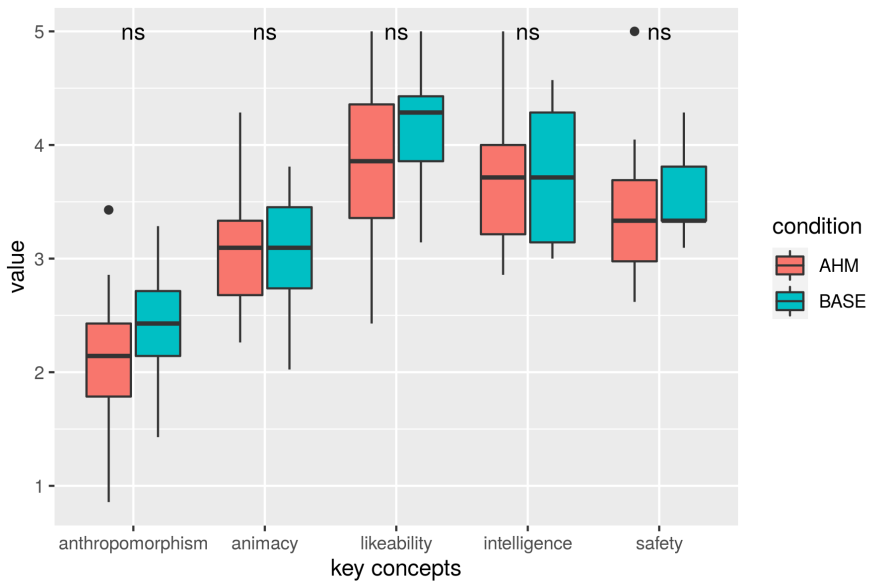


Figure 10.24: Evaluation of the godspeed questionnaire for the baseline and incremental condition.

SUBJECTIVE RATINGS A visualization of the results of the *GQS* for the two conditions can be found in fig. 10.24. It can be seen that only the key concepts of anthropomorphism and likability the mean values differ between the conditions. Interestingly, for both concepts, the *BASE* condition receives higher values. The exact means and standard deviations for the different key concepts can be found in appendix G.3.1. However, these differences are not significant ($p > .05$).

The participants rated the appropriateness of Flobi's statements regarding their length (four items) and timing (tree items), and the interaction itself (six items). The visualization is depicted in fig. 10.25, and the individual items for the appropriateness are presented in appendix B. The Cronbach's α for the four items regarding the tim-

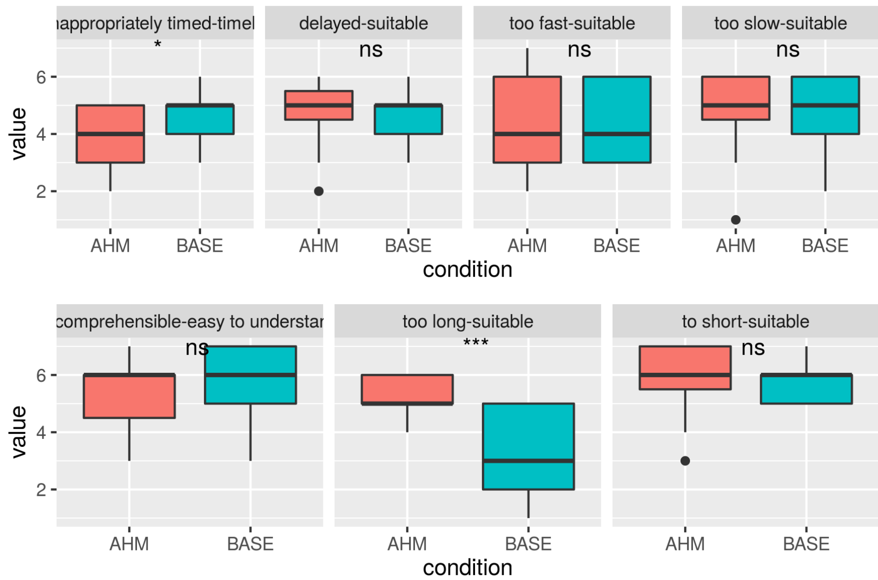


Figure 10.25: Subjective ratings of the appropriateness of the statements in timing and length.

ing and the three items regarding the length is each unsatisfactory to calculate an average value ($\alpha_{time} = .69$, $\alpha_{length} = .49$). However, the results of the *MANOVA* shows that there is a statistically significant difference between the conditions on the combined dependent variables (ratings of the appropriateness of the statements in timing and length), $F(7, 20) = 6.86, p < .001, \eta^2 = 0.51$. The post-hoc test shows, that there is significant difference in the rating of "inappropriately timed" vs. "timely" and "too long" vs. "suitable". The participants in the *AHM* condition rated Flobi's statement less timely ($F(1, 26) = 4.63, p < .041, \eta^2 = 0.15$), but more suitable regarding length ($F(1, 40) = 27.50, p < .001, \eta^2 = 0.71$). A further question was what kind of chunking the participants would have liked. In the *BASE* condition, the preferred chunk size was four, whereas the participants in the *AHM* condition prefer a grouping of three items ($\tilde{x}_{BASE} = 4, \sigma^2_{BASE} = 0.58$; $\tilde{x}_{AHM} = 3, \sigma^2_{AHM} = 0.31$).

In the post-hoc questionnaire, the participants had the opportunity to leave general comments. Some of them regarding the voice or Flobi's behavior are shown in example 10.3.5 and 10.3.6.

Example 10.3.5: Example comment by participant in the *BASE* condition

VP09: *"Schade, dass nur der virtuelle und nicht der echte Flobi anwesend war. Wenn er einen vollständigen Körper hätte, könnte er auch auf Gegenstände zeigen."*

VP21: *"Die Anweisungen waren besonders mit den optischen Signalen gut umsetzbar. Ich denke bei mehrfacher Benutzung ist es einfacher sich die Dinge zu merken."*

VP09: "It's a shame that only the virtual and not the real Flobi was present. If it had a full body, it could also point to objects."

VP21: "The instructions were particularly easy to implement with the optical signals. I think it's easier to remember things when you use it multiple times."

Participants in the baseline commented on its possibility to highlight the current *FoD*. One participant complains about Flobi's missing arms to point at objects, the other liked the lighted handles.

Example 10.3.6: Example comment by participant in the *BASE* condition

VP04: *"Die Stimme war teils ein bisschen mechanisch und dadurch abgehakt."*

VP08: *"Man musste sich kurz an Flobis Rhythmus gewöhnen."*

VP20: *"Angenehme Stimme - freundlich, man fühlte sich gleich willkommen."*

VP26: *"Viele Anweisungen waren etwas zu schnell, ansonsten alles gut."*

VP04: "The voice was a bit mechanical sometimes, which made it choppy."

VP08: "It took a moment to get used to Flobi's rhythm."

VP20: "Pleasant voice - friendly, you immediately felt welcome."

VP26: "Many instructions were a little too fast, otherwise everything was good."

In the *AHM* condition, participants left one negative and one positive statements regarding the voice, and two additional participants struggled with the information presentation rhythm. In the *debriefing*, some participants noted, that they had issues to interpret Flobi's pauses. Sometimes they weren't sure if the last ingredient for the location was already listed.

10.3.5 *Discussion and Lessons Learned*

In this section, I explored a more practical task and investigated how such an interactive scenario can be formalized and modeled. Therefore, I improved an existing interaction scenarios in the *CSRA* by applying the *AHM* on one dialogue act of the information exchange between the human and the agent. In this case, the agent provide information about the needed ingredients for backing a cake from a specific location in the apartment. The *AHM* hesitate the next chunk of information within this dialogue act until the human's attention is back, meaning s/he is ready for it. This is realized through the agent's monitoring of the human to draw conclusions about the corresponding task progress.

The evaluation of the *AHM* shows that the participants in the *AHM* condition perform significantly better in the task during the interaction, measured by a lower error rate. However, there are differences in the type of error. Participants in the *BASE* condition had a significant higher number of missed objects. Interestingly, the most participants in the baseline waited until Flobi stated all eight object per location, before they started fetching the ingredients. To this end, they had to remember more items at once and therefore needed to be attentive longer. Participants in the *AHM* condition could perform the task in their speed, as the agent waited with the next information until the participants completed the subtask or verbalized their readiness for the next information. Interestingly, the number of wrong objects is sightly higher in the *AHM* condition, e.g., fetching vanilla sugar instead of sugar. This could be explained due to the fact, that in the *BASE* condition all objects—also the similar ones—are mentioned together, whereas in the *AHM* condition these are stated successively. In addition, participants in the *BASE* condition heard the description more often. Even though the total number of asking for repetition does not significantly differ between the two conditions, the participants in the *BASE* heard all eight objects on demand, whereas the participants in the *AHM* condition just heard the last information chunk. This may be one explanation for the fact that the task performance in the cleaning phase not differ. The positive effect on the task performance in the interaction cannot be observed in the task performance after the interaction. However, this could have several additional reasons. The participants had their object directly in front of them and therefore know how many objects the need to put back to their location. Furthermore, the participants which made more errors in the fetching task—in terms of missing an object—had an easier cleaning task, due to the fewer ingredients they fetched.

Interestingly, the *AHM* had no effect on the interaction time, but on some other side effects. The participants in the *AHM* condition rated Flobi's statement less timely, but more suitable regarding length. This is interestingly, as the *AHM* provided the next information "just in time", whenever the subtask was completed or the participants demand the next chunk. This shows that the smaller the blocks of information, the more important the moment when the information is presented, which is also reflected by the comments, stated by the participants at the end of the questionnaire and the debriefing.

Further features for the *AHM* are of course conceivable, such as monitoring the task progress and detecting errors in their execution. Especially the recognition of errors during the task can detect understanding errors. For example, if Flobi detects that attention is paid to the wrong cabinet—recognized by the *VFoA* or the opening of a wrong cabinet—it could repeat its highlight strategy. In addition, further—more intrusive reactions are conceivable, such as a correction.

However, with the current *AHM*, it is possible to improve the performance in a practical task during the interaction. It is still questionable, if the *AHM* can also improve the performance measured after the interaction. To investigate this further, a better way of measuring the task performance should be considered, or more precisely another task is needed. Furthermore, to get more profound insights into the participants subjective ratings, the debriefing should be expanded. The experience from the last studies show that participants are willing to talk about their experience in the debriefing. This should be formalized by applying the method of semi-structured interviewed.

10.4 EC 4: INTRODUCING LENGTHENING AND NEW EVALUATION APPROACH

The results of the last two pilot studies (in section 10.2 and 10.3) shows, that a more human-like hesitation strategy is needed to improve differentiation of the unfilled pause with a turn end. In the examples by Goodwin in section 2.2.3 sometimes a lengthening in the word for the unfilled pause is observed. Furthermore, Betz et al. propose a model based on *HHI* research, which main feature is the use of lengthening [BWV16]. In their corpus study, they found that humans in *HHI* often use lengthening as a time-buying strategy. To this end, the feature of lengthening is integrated into the *AHM*. In this section, the resulting model and the interaction study in detail is presented and further analyses to investigate whether the use of lengthening in the hesitation intervention strategy improves the participants subjective ratings regarding the interaction and the agent is conducted. Focussing on the lengthening feature, the attention concept is kept simple in this experiment.

Furthermore, the study design is improved by introducing a new measurement of task performance. Inspired by the results of the last interaction scenario (section 10.3), a practical task for the users after the information-providing is integrated. To this end, the information recall—the task performance—can be measured by the performance in the practical task after the interaction and need not to accessed by a questionnaire.

In a cooperative effort, my colleague S. Betz and I realized the feature of lengthening in the context of my *AHM* and the information presentation scenario. We published preliminary results of the corresponding experiment in [Bet+18] and the implementation of the hesitation strategy in [Car+18].

10.4.1 *Attention-Hesitation Dialogue Coordination Model*

In the previous pilot studies, the *AHMs* were quite differently. Because of the negative subjective feedback in the second pilot study, we choose the *AHM* from the first pilot study with a lengthening extension in the hesitation strategy.

WHEN TO (RE-)ACT: ATTENTION CONCEPT In this model, the human is inattentive when the *VFoA* moves away from the agent, like in the *AHM* condition of the first pilot study. Therefore, when the user looks away from the agent, the intervention strategy starts and as soon as the user looks back at the agent the strategy stops.

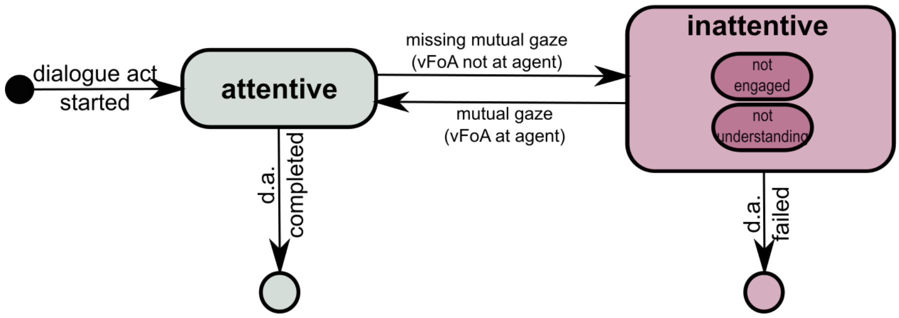


Figure 10.26: Concept of attention for EC4.

HOW TO (RE-)ACT: HESITATION INTERVENTION STRATEGY Whenever the agent loses the attention of the human interaction partner, it reacts with the hesitation strategy depicted in fig. 10.27. In this model,

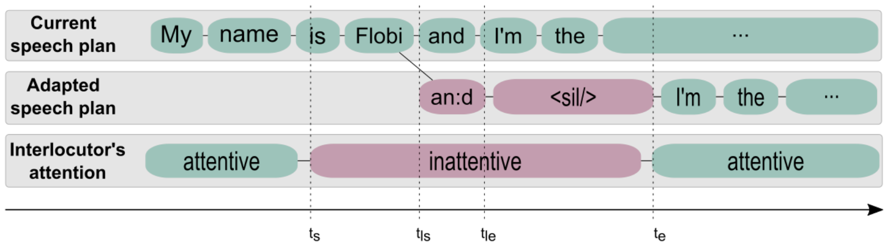


Figure 10.27: Hesitation intervention strategy for *EC4*.

the agent reacts by lengthening a syllable at the next appropriate point, followed by an unfilled pause. After the lengthening, the agent produces an unfilled pause. The strategy itself can be stopped (t_e) at several points:

BEFORE THE LENGTHENING STARTS ($t_e < t_{ls}$): If the strategy is interrupted before t_{ts}, it has no effect on the synthesis. The current speech plan will be pursued further.

BEFORE THE LENGTHENING ENDS ($t_{ls} < t_e < t_{le}$) The lengthening will be produced, but afterwards the agent directly continues speaking.

AFTER THE LENGTHENING ENDS ($t_{le} < t_e$) The agent produces the lengthening followed by an unfilled pause. When the interaction strategy stops, the agent simply continues speaking.

10.4.2 *Interaction Scenario and Implementation*

In this scenario, the agent provides a background story and instructs the participants to look for hidden sweets at seven different locations in the apartment. Each participant is asked to search for sweets that have allegedly been hidden in various places in the *CSRA*. The task is embedded in a story about construction workers who have just left the apartment and caused confusion in the agent's sensors because of the dust they stirred. The agent lists all potential hiding spots and asks the participant to memorize them and examine later. After the explanation, the participants can search for these sweets at the mentioned places. The interaction consists of five phases, which are listed briefly.

GREETING: The agent welcomes the user, who has the possibility to great back.

INFORMATION: The agent introduces itself, the apartment, and the task at hand: search for sweets that are hidden at different places in the apartment.

SWEETS: Flobi lists all seven potential hiding places in the smart environment, four in the kitchen and three in the living room.

SEARCH: The agent requests to search for the sweets. Participants performing this task without the help of Flobi.

FAREWELL: The agent thanks, says goodbye to the user, and requests to fill out the questionnaire at the computer in the living room.

To realize this scenario, the dialogue system presented in chapter 7 and section 10.2.2 is used. Furthermore, a new implementation of the hesitation strategy is used. This is explained in more detail below. The system is completely autonomous.

Figure 10.27 shows an example of the hesitation strategy, which is implemented as a separate module into *inprotk* (see 7.3.2). At the moment t_s, the module receives the event start hesitation. The hesitation module takes the *Incremental Units (IUs)* from the *left buffer* of the synthesis module, in this case a list of wordIUs, each representing a single word. It searches for the best entry point and lengthens the most appropriate segment, according to the research results by [BWV16]:

for word **in** words-to-say **do**
  if type of word $\in$ [determiner, preposition, conjunction] **then**
    if type of last syllable of word $\in$ [long-vovel, nasal] **then**
      append word to best
    else
      append word to acceptable
    end if
  else if type of last syllable of word $\in$ [long-vovel, nasal] **then**
    append word to appropriate
  end if
end for
for entrypoints **in** [best, appropriate, acceptable] **do**

if entrypoints is not empty **then**
return first element from entrypoints
end if
end for

In this example, the synthesis module already played back the first two wordIUs "my" and "name". The rest of the current phrase ("is Flobi and I'm the ...") is still in the playback pipeline. According to the proposed strategy, the best entry point for the hesitation strategy is a determiner, a preposition or a conjunction (function word), which contains a long vowel or a nasal. If the function word does not contain a log vowel or nasal, it is marked as an acceptable entry point. In this example, this matches the wordIU "and" (t_s). If no function word with an appropriate entry point is found, the strategy searches the next appropriate entry point in the other words. As a maximum search, we have inserted a look ahead with a limit of 5 words. This means that the hesitation does not start too late after a shift in attention. The order is as follows:

1. Function words with a long vowel or a nasal.
2. Other words with a long vowel or a nasal.
3. Function words without a long vowel or a nasal.

If none of these entry points are found, then the strategy will not apply the lengthening before the unfilled pause. Otherwise, the last syllable which contains a nasal or a long vowel (or another segment if this does not contain a long vowel or nasal) is then stretched by the so-called stretch factor. Further information about this can be found at [Bet+18]. Afterward, the synthesis module will be paused until the attention is back on the agent.

In the case the dialogue management wants to stop the interaction strategy earlier (e.g., the estimated attention state of the human interaction partner changed to *attentive*), the strategy can be interrupted at several points:

BEFORE THE LENGTHENING STARTS ($t_e < t_{ls}$): If the strategy is interrupted before t_{ts}, it has no effect on the synthesis. The current speech plan will be pursued further. The stretch factor from the corresponding syllable will be removed.

BEFORE THE LENGTHENING ENDS ($t_{ls} < t_e < t_{le}$) The lengthening will be produced, but afterward the agent directly continues speaking. The strategy does not initiate the pausing of the synthesis.

AFTER THE LENGTHENING ENDS ($t_{le} < t_e$) The agent produces the lengthening followed by an unfilled pause. When the interaction strategy stops, the agent simply continues speaking. This is initiated by a resume of the speech synthesis.

10.4.3 *Evaluation*

We evaluate this hesitation strategy in an *HAI* study. S. Betz and I carried out an interaction study with 17 participants interacted with the baseline system, and 14 with the hesitation system. Four additionally participants were recorded in a third condition for exploratory purposes, see [Bet+18] for more information. In addition, I have carried out a follow-up study to get further participants afterwards. The presented data is a combination of this first data collection (without the third condition) and my follow-up survey to receive the 20 participant per condition. The experimental design, implementation, and conditions are the same in both data collections and take place in the *CSRA*.

Study Design

The experiment design follows the general study design depicted in fig. 9.4 and applied in the proceeding pilot studies. As in the polit studies, a between-subject design was chosen. Each participant interacts with our system in either the *AHM* or the *BASE* condition. However, a few adaptions had to be made. To access the memory capacity of the participants, again the short memory pre-test is conducted beforehand (see section 10.3.4). The resulting task performance of the pretest ($\mathsf{TP}_{pretest}$) is used to check, whether the participants' memory performance differs between the conditions.

Figure 10.28 shows the experimental setup. As in the second *EC* in section 10.2, we used three kinds of disruptions during the *Sweet instruction* phase. First, a visual disruption, i.e., a light that turns on in the visual focus. The second is an auditive disruption, music in the living room, played for a few seconds. The last disruption is multimodal, a human enters the apartment shortly. All disruptions happen at the same three points of the interaction and are not randomized, to keep the results comparable between the participants. For the third phase (the searching), participants are asked to call out each place before looking at it, to ensure that they remember the places and do not search the entire place and find things by chance. The interaction is monitored audiovisually from the adjacent control room. The number of sweets retrieved by each participant is recorded as task performance TP_{sweets}.

After the interaction, the participants fill out the questionnaire to assess personal data and prior experience, subjective evaluation of the agent, and collect general comments about the interaction and the study. In addition to the godspeed questionnaire, we collected information about the perceived quality of the synthetic voice via the 5-point MOS questionnaire. This scale was chosen for maximum comparability with traditional MOS-based synthesis evaluation (example 10.4.1).

Figure 10.28: Experimental setup. Upper: a person interacting with the agent. Lower: a ground view of the apartment. The red sweets display the hiding places.

Example 10.4.1: Traditional MOS-based synthesis evaluation.

Wie beurteilen Sie die Qualität der Stimme des Agenten:
sehr schlecht ○○○○○ *sehr gut*

How do you rate the quality of the agent's voice:
very bad ○○○○○ very well

The experiment ends with the *debriefing*, which contains a short *semi-structured interview* to get further insights and impressions from the participants. In particular, we are interested if they felt that the agent adopted its behavior to their behavior. We choose the method of semi-structured interview to have the opportunity for further inquiries and for more qualitative evaluation. The questions guiding the interview are shown in example 10.4.2.

Example 10.4.2: Questions for the semi-structured interview.

1. *Hattest du das Gefühl, dass Flobi sich dir angepasst hat?*
2. *Wenn ja, wie und warum?*
3. *Wie fandest du das?*
4. *Hast du sonst noch irgendwelche Anmerkungen oder Kommentare?*

---

1. Did you feel like Flobi adapted to you?
2. If so, how and why?
3. How did you like that?
4. Do you have any other remarks or comments?

Participants

In total, 48 participants were recruited at the Bielefeld University campus to take part in this study. Two participants in the first had to be excluded because their language competence did not suffice to follow the instructions correctly. Furthermore, four participants were recorded for a third condition for exploratory purposes (see [Bet+18] for more information) and are also excluded in this evaluation. In addition, another two trails had to be excluded, because of missing data recordings. From the remaining participants (n=40), 20 were in the baseline (12 female, 8 male; $M_{age} = 23.30$, $SD_{age} = 5.03$) and 20 in the *AHM* condition (12 female, 8 male; $M_{age} = 24.95$, $SD_{age} = 3.14$).

Figure 10.29 visualizes the distribution of the task performance in the memory pretest ($TP_{pretest}$) for each condition. There is no statistical significant difference between the pretest performance in the *AHM* condition ($M = 6.70, SD = 1.45$) and the baseline ($M = 6.85, SD = 1.39$), $W = 189.5, p = .775$. Furthermore, participants

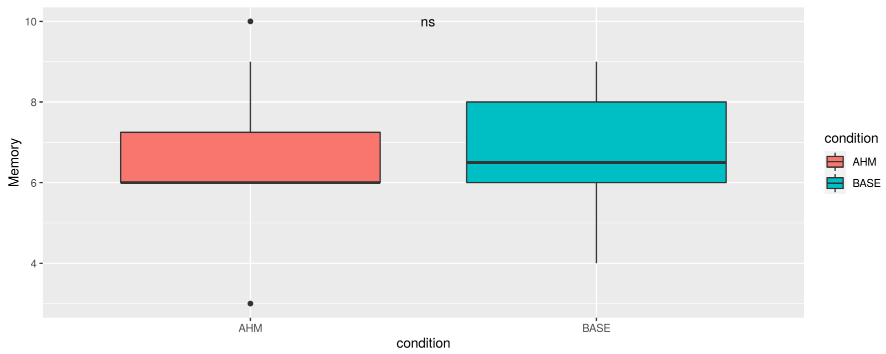


Figure 10.29: Results of the memory pretest task performance ($TP_{pretest}$) for the *BASE* and *AHM* condition.

are balanced regarding their average prior experience with technical intelligent systems ($M_{AHM} = 2.96$, $SD_{AHM} = 1.08$;$M_{BASE} = 2.64$, $SD_{BASE} = 0.79$), $W = 229, p = .439$. Participants in both conditions have no or very little experience with robotic systems in general and the virtual agent Flobi in particular. They have little experience with speech systems in general. However, in addition, they have some programming experiences and are very experienced with the use of computers in general.

Task Performance Hypothesis

Figure 10.30 depicts the performance in the finding task of sweets TP_{sweets} for the different conditions. First, I test my hypothesis, that participants in the *AHM* condition achieve higher TP_{sweets} values than participants in the *BASE* condition. Figure 10.30 visualize the distribution of the TP_{sweets} for the different conditions. The *WR-Test* shows a significant difference in the scores of TP_{sweets} for the *AHM* ($M = 6.30, SD = 0.86$) and *BASE* condition ($M = 5.50, SD = 1.28$), $W = 273, p = .020, r = 0.33$. This confirms my hypothesis.

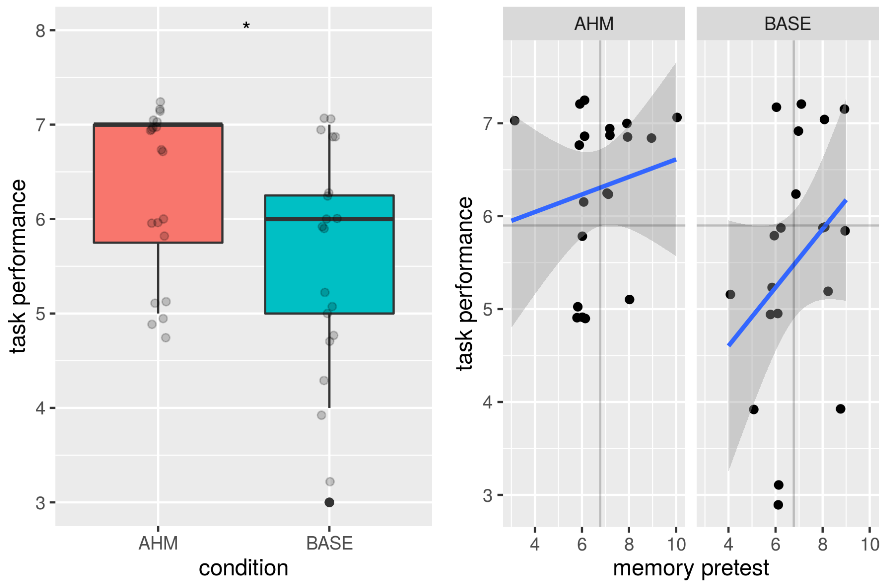


Figure 10.30: Left: Distribution of TP_{sweets} for the two conditions *BASE* and *AHM*. Right: Performance of TP_{sweets} for $TP_{pretest}$ in the different conditions. The horizontal line represents the mean of the points in total, the vertical line visualizes the mean memory performance of all participants in the pretest. The blue line represents a linear regression with a 95% confidence interval.

Side effects on the Interaction

Next, I investigate the side effects on the interaction, regarding the subjective ratings of the agent and the *VFoA* of the human interaction partner.

VISUAL FOCUS OF ATTENTION Next, I will take a closer look at the *VFoA* of the participants. Figure 10.31 visualizes the distributions of attentiveness of the human interaction partner. The assumptions for a *MANOVA* is not met, therefore the non-parametric *WR-Test* with Bonferroni correction is applied. The number of look aways in the *BASE* condition ($Med = 17.0$) does not differ significantly from the participants in the *AHM* condition ($Med = 14.5$); $W = 239.5, p_{adj} = .580, r = -0.17$. Participants in the control group do not look away significantly more often. Furthermore, there is no significant difference in the overall time people look away from the agent or the current *FoD* between the *AHM* ($M = 58.48s, SD = 41.76$) and the *BASE* ($M = 39.89s, SD = 16.376$) condition, $W = 246, p_{adj} = .442, r = 0.20$.

SUBJECTIVE RATINGS AND QUALITATIVE EVALUATION The subjective ratings were assessed through several measures: the *GQS*, MOS ratings, general comments at the end of the questionnaire, and through the semi-structured interview. I examine the effect of the re-attention strategy on the *GQS*, the subjective ratings of the agent,

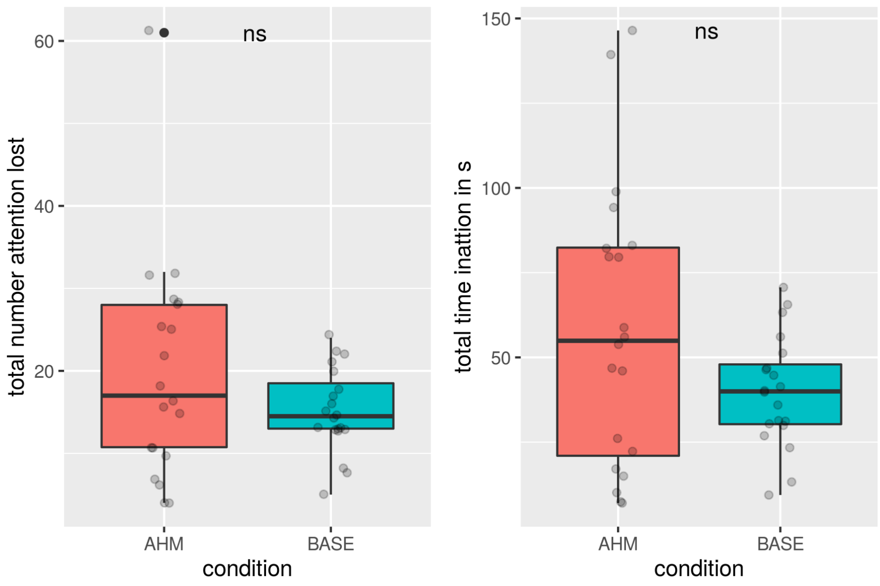


Figure 10.31: Distribution of *VFoA* moves away.

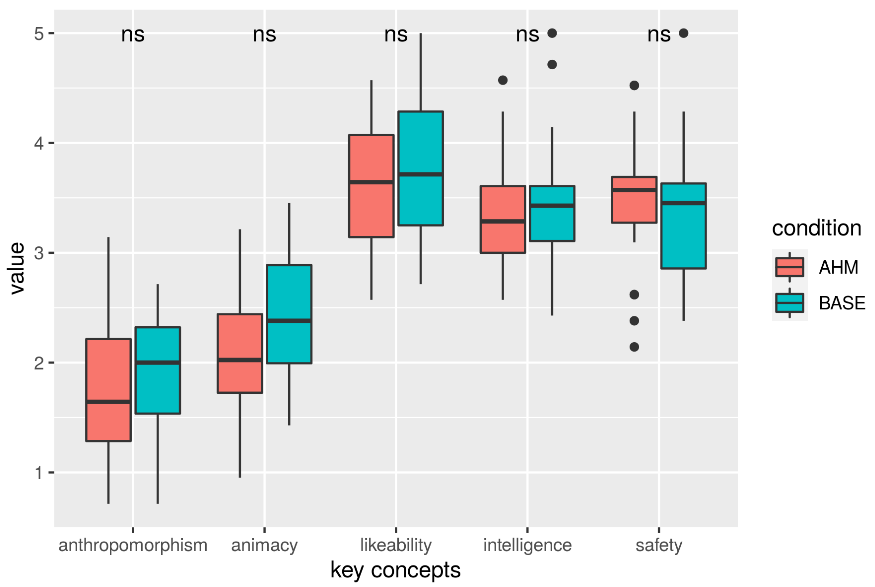


Figure 10.32: Subjective ratings of the agent for the five key concepts *anthropomorphism*, *animacy*, *likability*, *perceived intelligence*, and *perceived safety*.

regarding the five key concepts *anthropomorphism*, *animacy*, *likability*, *perceived intelligence*, and *perceived safety*. In contrast to the second EC in section 10.2, the condition has no significant effect on any of these attributes ($p > .1$). The exact statistics can be found in appendix G.4.3. As in the previous studies, the agent receives high values for *likability*, *perceived intelligence*, and *perceived safety* ($M > 3.3$), but rather low values for *anthropomorphism* and *animacy* ($M < 2.5$) in both conditions.

Figure 10.33 visualizes the user feedback in the 5-point MOS scale on synthesis quality (1 = "very bad", 5 = "very good"). The following

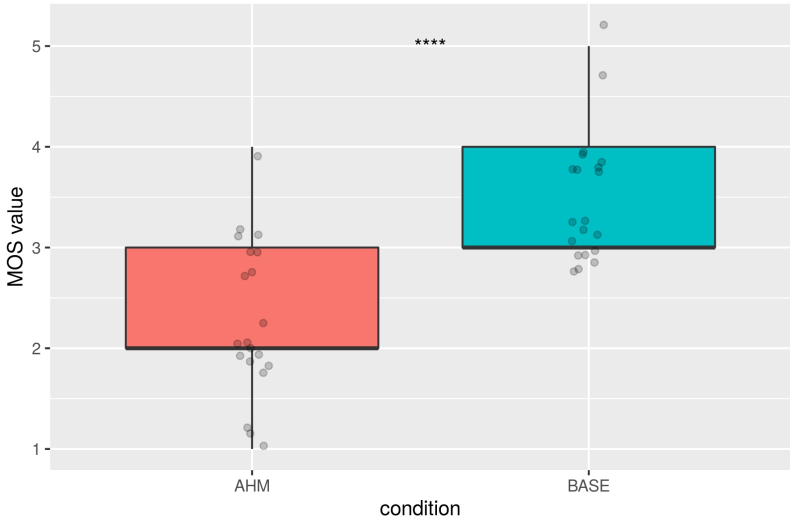


Figure 10.33: MOS results for the different conditions in the interaction study.

results can be derived from these graphs. Users rate the synthesis quality worse in the *AHM* condition ($M = 2.30, SD = 0.80$) compared to the *BASE* condition ($M = 3.55, SD = 0.69$). The *WR-Test* shows that there is a significant effect of the *AHM* condition on users' perception of synthesis quality, $W = 53, p < .001, r = -0.67$. This effect on the synthesis quality is also reflected by the general comments of the study participants.

The participants had two opportunities to leave some general comments (1) in the last part of the questionnaire and (2) at the semi-structured interview after the interaction, which we conducted to figure out if they noticed the adaption of the agent and which behavior they would prefer. In the general comments of the questionnaire, 13 out of 20 participants in the *AHM* conditions, reported negative voice quality. In the baseline, only two participants criticized monotonous voice without emphasis.

In the interview, some people reported that they liked the adaptive behavior of the agent. In the *BASE* condition, only three of the 20 participants had the feeling that Flobi had adapted to them. Two of them thought that Flobi looks in the same direction as the participants. This was indeed sometimes the case, but because the participants followed the view of Flobi and not vice versa. Only one participant noted that Flobi looked at them most of the time and also turned its head toward the participants.

In the *AHM* condition, 11 of 20 participants had the feeling that Flobi had adapted to them. All the 11 participants noticed that it paused when they looked away. Four of the 11 subjects liked this behavior. Example statements can be seen in example 10.4.3.

Example 10.4.3: Example answeres by participant in the *BASE* condition on the question how they like this behavior.

+ VP24: *"Ich fand es hilfreich."*

+ VP34: *"Das war gut bei den Störungen."*

o VP22: *"Bin mir unsicher."*

o VP25: *"Weiß ich nicht."*

- VP32: *"Stocken hat gestört."*

- VP39: *"Doof! Erst referenziert und dann hat er aufgehört darüber zu reden."*

+ VP24: "I found it helpful."

+ VP34: "It was good, during the disturbances."

o VP22: "I am unsure."

o VP25: "I do not know."

- VP32: "Stocken hat gestört"

- VP39: "Stupid! First referenced and then stopped talking about it."

They found it rather helpful, e.g., during the disturbances [VP25, VP34]. Reasons for this included, e.g., the additional time for looking. Two of 11 participants were uncertain, whether they like this behavior [VP22, VP25]. Four of 11 participants found it rather disturbing [V32, VP39].

In the interview, three additional participants said, that the lagging of the speech was disturbing. They even did not notice, that Flobi only speaks whenever they look at it. In general, these participants reported in the interview, that they like the idea of self-interrupting and adaptable agents and would prefer such behavior—under the premise of improvements to the technical realization.

10.4.4 *Discussion and Lessons Learned*

This interaction study produced a set of results for the evaluation of my *AHM*. My hypothesis that participants in the *AHM* condition perform better—measured as post-interaction information recall—can be accepted. In contrast to the first and second EC (in section 10.1 and 10.2), this is not measured via a questionnaire, but rather via the finding rate of sweets hidden in the intelligent apartment. The better task performance could result in theses different measurements or more precisely in the different tasks participants had to perform. In both studies, the task for the short interaction was to provide information from the agent to the human interaction partner. Whereas in *EC1* the task performance was measured through a post-hoc questionnaire, in this experiment the participants had to find sweets. This practical task could be more understandable for the participants and a better reason to remember Flobi's statements. Another explanation can be that the improved hesitation strategy leads to the positive effect in the task performance independently of the measurements. Here, further research is necessary. The results of the task performance also show a ceiling effect in the *AHM* condition. When comparing the two regression lines, the fig. 10.30 indicates that especially participants which score badly in the pretest, benefit from the hesitation condition. For statistical analysis, this sample is not large enough. But it would be an interesting further research question, if the task performance in the memory pretest has an impact of the effect of the *AHM*. It would be interesting to test, whether the effect is bigger or not, when the total number of points that can be achieved is increased.

Furthermore, in this study, the voice quality was rated significant differently in the two conditions. This is also reflected in the post-hoc statements and interviews. However, the *AHM* did not influence the ratings of the five key concepts. The ratings of the agent itself were not affected in this study. It remains to be decided, if the negative ratings regarding the agent's voice quality are due to the intervention strategy itself or due to actual sound interferences. An investigation of the recording shows that the synthesis sometimes produces crackling noises within the interaction. Another interpretation is that the strategy perceived in interaction as not appropriate. The intervention strategy itself could be judged as appropriate and desirable, but turn out as annoying in the interaction itself. Obviously, the intervention strategy itself could be not appropriate and require improvements.

In addition, the effect on the inattention could not be reproduced, participants in the *AHM* condition did not look away significantly more often or longer in total. This could be due to the following reasons: The interaction was longer than in the first pilot study and had several foci of discourse. I measured the inattention, whenever the human did not look at the agent or the current *FoD*, but the agent itself only spoke when the human looked at the agent. This mismatch of attention measurement and speaking strategy could be the reason for this effect. Furthermore, in this experiment the agent did not stop speaking instantly after it lost the attention of the user, but rather at specific points of speech. This behavior could be more difficult to understand for the participants because the reaction of the agent did not follow immediately after the action of the human.

These findings have important implications for developing my *AHM*. The moment to react—the start of the hesitation strategy—should reflect the current *FoD*. Several participants were irritated that the agent had stopped speaking when they looked at the current *FoD*. Even if it was a good strategy in the first pilot study to hesitate whenever the human looks away, in a scenario with different *FoD* this attention strategy is not sufficient. The lengthening seems to counteract the negative ratings of the agent from the second pilot study in section 10.2. However, the rating of the voice quality was poor. The reason for this is unclear and need further investigations.

The experimental design was suitable for this investigation at hand. The change of the task performance measurements was successful, participants can find sweets in both conditions. However, due to the ceiling effect, the number of sweets should be extended. In addition, it would be interesting if participants are performing better also better in an information questionnaire at the end of the interaction with the current *AHM*. The extended measurements of the subjective ratings gave further valuable insights. The method of semi-structured as well as the MOS-scale allow a differentiated picture of the participants' subjective ratings.

To sum up, this study set out with the aim of assessing whether an enhanced hesitation strategy—the use of lengthening—can improve the participants' subjective ratings of the agent. The results show, that the godspeed values are not rated significantly lower in the *AHM* condition than in the baseline. However, the results of the extended measurements via the MOS-scale made clear that participants rated the voice quality lower in the *AHM* than in the *BASE* condition. The second aim was to investigate whether my *AHM* can increase the task performance in more practical task. I could reproduce the positive effect on the task performance from *EC3* and show that my *AHM* can increase the task performance in such practical tasks.

In addition, this study raises several further research questions: (1) It is also possible to increase the task performance in a questionnaire? (2) Do characteristics of participants have an influence on the effect of the *AHM*? (3) It is possible to reduce the negative subjective ratings of the synthesis quality? (4) Which task performance can be expected when using an advanced attention model, which also include the changes of the *FoD*? Some of this research questions are addressed in the next sections.

To conclude this chapter, I could again that the *AHM* can work autonomously. Furthermore, the interactive behavior of my *AHM* lead to better task performance measured by information recall of the hiding places of sweets. This effect is, however, accompanied by lower subjective ratings of the agent's voice quality, although the use of lengthening counteract the lower subjective ratings of the agent itself.

10.5 EC 5: BRINGING IT ALL TOGETHER

In this last evaluation cycle, I combine several aspects from my previous models and evaluate this enhanced AHM in a final interaction study. The previous studies showed that my AHM can be enhanced in several ways. Therefore, I combine the differentiation of two states in the attention concepts (used in EC2 and EC3) with the enhanced hesitation strategy of the last experiment (EC4). Especially in the last experiment in section 10.4, it turned out that the attention concept should depend on the current *FoD*. Furthermore, hesitation vowels are introduced as additional re-attention mechanism feature, according to the results by Goodwin [Goo81]. With the resulting enhanced AHM, I investigate whether it is possible to increases the task performance without negative side effects on the interaction.

10.5.1 *Attention-Hesitation Dialogue Coordination Model*

The last experiment in EC4 showed that the moment to react should reflect the current *FoD*. Therefore, I combine the AHM from EC3 in section 10.2 with the enhanced hesitation strategy of the last experiment in section 10.4. Especially lengthening tuned out to be an appropriate feature to *buy time* in the last EC4. Therefore, I included this behavior additionally in the repetition of my highlight strategy. This less monotonous behavior should counteract the negative ratings of the repetition. Furthermore, I include a hesitation vowel as an additional re-attention mechanism, if participants are inattentive over a longer period of time.

WHEN TO (RE-)ACT: ATTENTION CONCEPT As in EC2, this model includes the current VFoA as well as the current *FoD*. It also distinguishes two different situations (a) the current *FoD* changes and (b) the *FoD* is ongoing. In (a) the user is inattentive, if the attention guiding strategy fails. The strategy succeeds when the user looks at the new *FoD* at least once within a time frame. For an ongoing *FoD* (b), the user is inattentive whenever the VFoA neither matches the current *FoD* nor the agent itself.

HOW TO (RE-)ACT: HESITATION INTERVENTION STRATEGIES Whenever the agent losses the attention of the human interaction partner, it reacts with a hesitation strategy (see fig. 10.35). This model combines intervention strategies from the previous studies with some modifications. (a) When the attention guiding strategy fails, the agent reacts with a repetition of the multi-modal guiding strategy. In addition, this repetition includes lengthening of orientation giving words, such as "on your le:ft side". This is repeated up to three times, until the user is attentive—looked at the new *FoD* at least once. (b) The agent

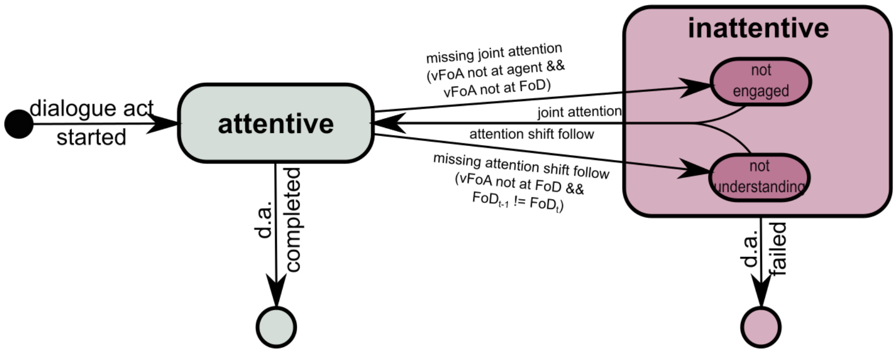


Figure 10.34: Concept of attention for EC5.

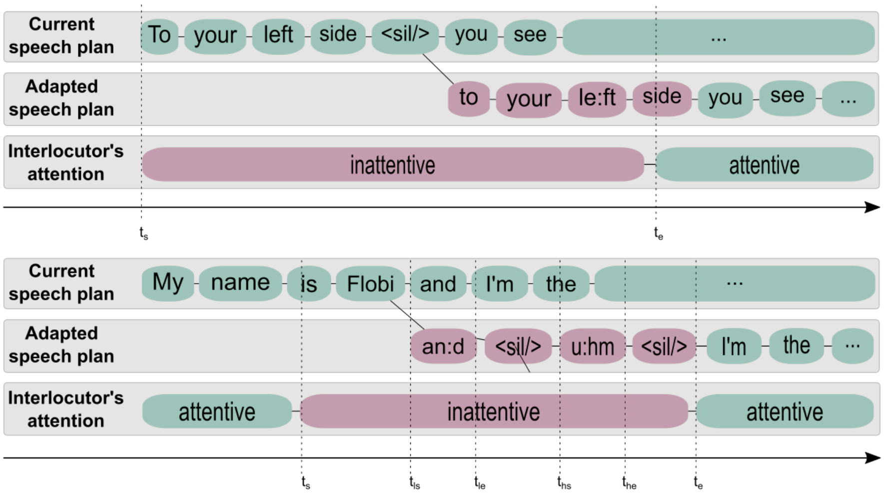


Figure 10.35: Hesitation strategies for *EC5*: (a) highlight (b) re-attention.

uses the re-attention hesitation strategy, as soon as the user looses the attention during an ongoing *FoD*. This hesitation strategy starts with lengthening at the next useful entry point. After the lengthening, the agent produce an unfilled pause, followed by a hesitation vowel (with lengthening), again followed by an unfilled pause. The strategy itself can be stopped (t_e) at several points:

BEFORE THE LENGTHENING STARTS ($t_e < t_{ls}$): If the strategy is interrupted before t_{ts}, it has no effect on the synthesis. The current speech plan will be pursued further.

BEFORE THE LENGTHENING ENDS ($t_{ls} < t_e < t_{le}$) The lengthening will be produced, but afterwards the agent directly continues speaking.

IN THE FIRST UNFILLED PAUSE ($t_{le} < t_e < t_{hs}$) The agent produces the lengthening followed by an unfilled pause. When the interaction strategy stops, the agent simply continues speaking.

BEFORE THE HESITATION VOWEL ENDS ($t_{hs} < t_e < t_{he}$) The agent produces the lengthening followed by an unfilled pause and the hesitation vowel. Afterwards, the agent directly continues speaking.

IN THE SECOND UNFILLED PAUSE ($t_{he} < t_e$) The agent produces the lengthening followed by an unfilled pause, a hesitation vowel, and again followed by an unfilled pause. When the interaction strategy stops, the agent simply continues speaking.

10.5.2 *Interaction Scenario and Implementation*

The interaction takes place in the kitchen of the *CSRA*. In this interaction, the agent provides a background story and instructs the user to look for hidden sweets. Before this task, the agent gives some general explanations about the intelligent apartment. Then, Flobi lists all potential hiding places, asking the participant to memorize and later investigate these. Afterwards, the user searches for the sweets. The interaction consists of five phases, which I list briefly in the following.

GREETING: The agent welcomes the user, who has the possibility to great back. The agent introduces the task at hand: search for sweets that are hidden at different places in the apartment.

INFORMATION: Flobi introduce the apartment and gives information to ten objects/locations within the smart environment. Afterwards, the user is requested to fill out the first part of the questionnaire at the computer in the living room (to the right of the participant).

SWEETS: Flobi lists all ten hiding places for sweets in the smart environment.

SEARCH: The agent requests to search for the sweets. Participants performing this task without the help of Flobi, but it provides feedback through nodding to the user whenever they say where to look.

FAREWELL: The agent thanks, says goodbye to the user, and requests to fill out the questionnaire at the computer in the living room.

To realize this scenario, the dialogue system presented in chapter 7 is uses. The user is facing a monitor in the kitchen, which is showing the virtual Flobi (presented in section 7.1). Using the web camera on top of the monitor, Flobi can detect faces in front of it and focus them, thus establishing shared attention. Flobi's utterances are predefined. To allow verbal self-interruptions, whenever the user is inattentive, the hesitation strategy presented in section 10.4.2 in improved using some modifications and enhancements. The hesitation module takes the *IUs* from the *left buffer* of the synthesis module, in this case a list of wordIUs, each representing a single word. It searches for the best entry point and lengthens the most appropriate segments as described in section 10.4.2. In contrast to the previous algorithm, I modified the maximum search. The look ahead limit is reduced from 5 to 4 words to allow a faster reaction on the attention shift. In this example, the synthesis module already played back the first two wordIUs "my" and "name". The rest of the current phrase ("is Flobi and I'm the ...") is still in the playback pipeline. According to the proposed strategy, the best entry point for the hesitation strategy is the wordIU "and" ($t_l s$). Furthermore, the strategy of section 10.4.2 has been changed in the following points: After the lengthening, the synthesis module will be paused up to 2000ms (< sil/ >). If this is not enough time, the module inserts a filler ("ähm" or "äh"), also applied with lengthening, followed by a second pause until the dialogue management stops the hesitation strategy (t_e). The modification of the speech stream is realized by a stopping of the synthesis, adding the filler, and again adding the not synthesized words to the synthesis module. After the filler is produced, the strategy pauses the synthesis. In the case the dialogue management wants to stop the interaction strategy earlier (e.g., the estimated attention state of the human interaction partner changed to attentive), the strategy can be interrupted at several points

BEFORE THE LENGTHENING STARTS ($t_e < t_{ls}$): If the strategy is interrupted before t_{ts}, it has no effect on the synthesis. The current speech plan will be pursued further. The stretch factor from the corresponding syllable will be removed.

BEFORE THE LENGTHENING ENDS ($t_{ls} < t_e < t_{le}$) The lengthening will be produced, but afterwards the agent directly continues speaking. The strategy does not initiate the pausing of the synthesis.

IN THE FIRST UNFILLED PAUSE ($t_{le} < t_e < t_{hs}$) The agent produces the lengthening followed by an unfilled pause. When the interaction strategy stops, the agent simply continues speaking. This is initiated by a resume of the speech synthesis.

BEFORE THE HESITATION VOWEL ENDS ($t_{hs} < t_e < t_{he}$) The agent produces the lengthening followed by an unfilled pause and the hesitation vowel. Afterwards, the agent directly continues speaking. This is initiated by not pausing again the speech synthesis.

IN THE SECOND UNFILLED PAUSE ($t_{he} < t_e$) The agent produces the lengthening followed by an unfilled pause, a hesitation vowel, and again followed by an unfilled pause. When the interaction strategy stops, the agent simply continues speaking. This is initiated by a resume of the speech synthesis.

To figure out the right moment for stopping and starting of the strategy, an attention module is implemented. This fuses again the information of the *DM* and the gaze detection with a simple rule-based decision presented in section 7.3.2 and section 10.2.2. Since previously observed gaze detections errors besides the recognition of mutual gaze, I decided to use a wizard for the gaze input. Therefore, I implemented a wizard for sending gaze input via the middleware. This is done via keyboard signals from the control room. It sends example gaze hypotheses for looking straight, to the left, right, up, or down.

In addition, the agent provides autonomous feedback through nodding to the user whenever they say where to look during searching of sweets. This is detected via a simple voice activity detection.

10.5.3 *Evaluation*

The scenario with my enhanced *AHM* is evaluated in a human-agent interaction study within the *CSRA*. To investigate my research question, whether the enhanced *AHM* can improve the task performance of information as well as a more practical task, the experimental design was modified.

Study Design

The experiment design follows the general study design. Figure 10.36

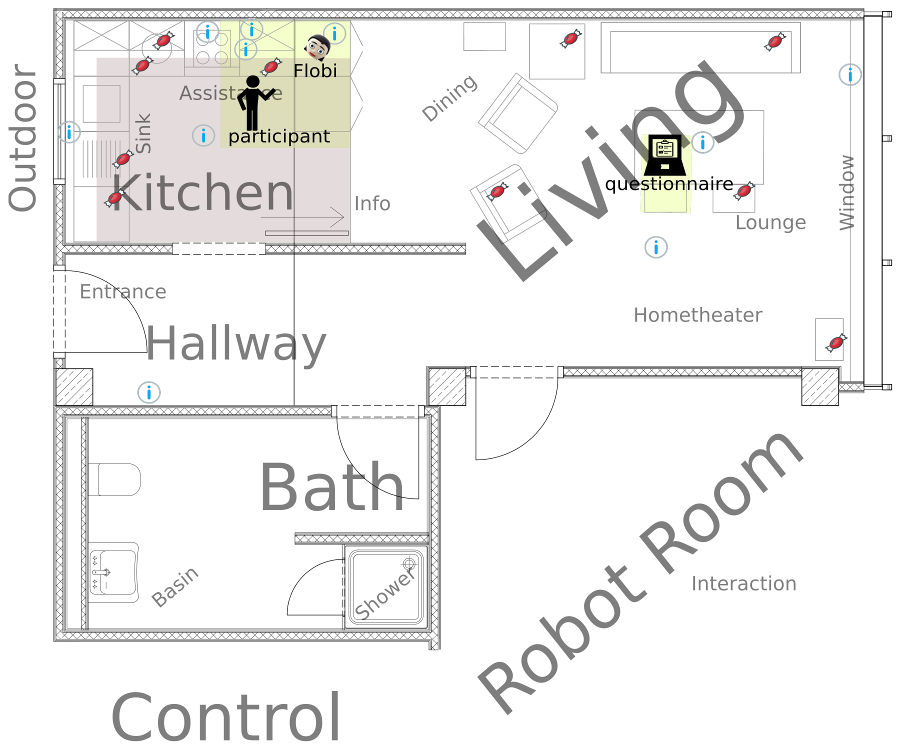


Figure 10.36: Experimental setup: a ground view of the apartment with the corresponding locations of the information items (blue) and sweets (red).

visualizes the interaction area and the corresponding locations of the information items (blue) and sweets (red). I use a between-subject study design to avoid corrie over effects between the conditions. Each participant interacts in one of the two conditions: *BASE* or *AHM*. The interaction is semi-autonomous, the information normally provided by the gaze tracker is replaced by a wizard to reduce detection errors. I perform the wizard in the *AHM* condition after several testing runs.

To access the memory capacity of the participants, the memory pre-test, introduced in *EC3*, is conducted in the *briefing*. During the interaction the participants are disturbed three times each in the *info* part as well as the *sweets* part, with different flashing lights, sounds, and human distractions (opening the door, knocking on the window). All disruptions happen at the same points of the interaction and are not randomized, to keep the results comparable between the participants. The interaction is monitored audiovisually from the adjacent control room.

The total task performance TP_{total} is calculated as the addition of performance in the *information* and *sweets* part:

$$\mathsf{TP}_{total} = \mathsf{TP}_{info} + \mathsf{TP}_{sweets} \tag{10.1}$$

The information task performance TP_{info} is measured in the first part of the questionnaire. The participant is asked to write down each object that was explained by Flobi. These replies are annotated by two different annotators separately. Uncertainties are addressed and discussed in a second stage. The same annotation procedure is used to assess the task performance in the *sweets* part (TP_{sweets}).

In addition, the participants fill out a questionnaire two times. The first one after the *info* part of the interaction. Beside the question for the information task, the godspeed questionnaire, and the MOS-based synthesis evaluation are assessed in this first questionnaire. After searching the *sweets*, participants filled out a second questionnaire to assess again the subjective rating of the agent, and its voice quality. In addition, personal data and prior experience are requested, and the participants had the opportunity to leaf general comments about the interaction and the study.

Furthermore, during the *debriefing*, I conduct a semi-structured interview to collect further information, depicted in example 10.5.1.

Example 10.5.1: Questions for the semi-structured interview.

1. *Hattest du das Gefühl, dass Flobi sich dir angepasst hat?*
2. *Wenn ja, wie und warum?*
3. *Hat Flobi bei dir*
   - *Pausen gemacht?*
   - *Wörter lang gezogen?*

  - *Häsitationen benutzt (ähm)?*
  - *etwas wiederholt?*
4. *Wie fandest du das Verhalten? Hat es dich gestört? Hat es dir geholfen? Keines von beiden?*
5. *Hast du sonst noch irgendwelche Anmerkungen oder Kommentare?*

---

1. Did you feel like Flobi adapted to you?
2. If so, how and why?
3. Did you realize that Flobi used
   - unfilled pauses?
   - lengthening?
   - hesitation vowels (uhm)?
   - repetitions?
4. How did you like this behavior? Did it bother you, did it help you, or neither of them?
5. Do you have other remarks or comments?

During the interview, the questions start very broadly and get more specific over time. The questions 3 to 5 are only asked in the *AHM* condition. This allows the participants to talk about the interaction strategy without alerting them of this specific behavior. I use a semi-structured interview and not a questionnaire to get first qualitative impressions. The advantage of being able to explain what *pausing, lengthening, hesitating, repeating* means and to ask for more details about the participants' impression outweighs the drawbacks, e.g., influencing the interviewee through own verbal comments and non-verbal cues[3]. Finally, the gaze is annotated afterwards to assess the visual focus of attention.

Participants

43 participants were recruited at the Bielefeld university campus to take part in this study. Three participants had to be excluded because of data loss. From the remaining participants (n=40), 20 were in the *BASE* (10 female, 10 male; $M_{age} = 23.50$, $SD_{age} = 2.42$) and 20 in the *AHM* condition (10 female, 10 male; $M_{age} = 23.95$, $SD_{age} = 3.65$). Figure 10.37 visualizes the distribution of task performance in the memory pretest ($TP_{pretest}$) for each condition. The *T-Test* shows no significant differences in the mean memory performance in the pre-test

3 For an overview of how using interviews as research method see [Kaj05]

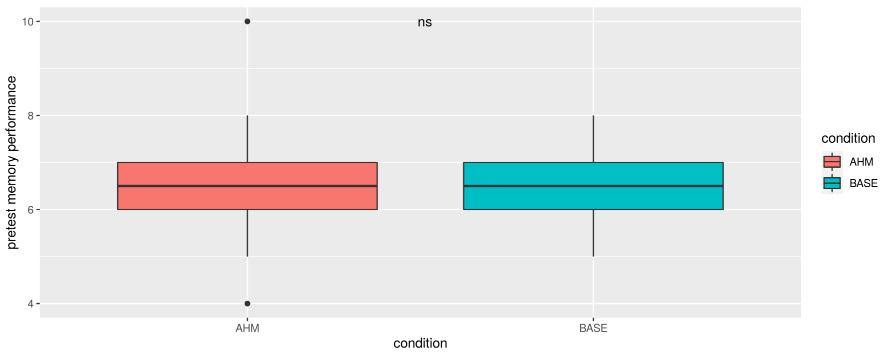


Figure 10.37: Results of the memory pretest for the *BASE* and *AHM* condition.

between the conditions, $t(33.95) = -0.14, p = 0.89$. The participants in the baseline remember on average $M_{pretest} = 6.5$ and in the hesitation $M_{pretest} = 6.55$ points.

Furthermore, the participants are balanced regarding their average prior experience with technical intelligent systems ($M_{AHM} = 2.77$, $SD_{AHM} = 0.76$; $M_{BASE} = 2.82$, $SD_{BASE} = 0.59$), $t(35.72) = -0.23, p = .817$. Participants in both condition have no or very little experience with robotic systems in general or the virtual agent Flobi. They have little experience with speech systems in general. In addition, they have some programming experiences, but all participants are very experienced with the use of computers in general.

Task Performance Hypothesis

To test my hypothesis, that the *AHM* leads to higher task performance—measured as post-interaction information recall—I compare the total task performance TP_{total} between conditions. As shown in equation 10.1, TP_{total} consists of the performance in the information part TP_{info} and the number of found sweets TP_{sweets}. A visualization of the task performance for the *information* part, the *sweets* part, and in total for each condition can be seen in fig. 10.38. The *T-Test* shows a significant difference in the mean of the overall task performance between the conditions, $t(37.95) = -2.22, p = .032, CI = [0.09, 2.00], d = .70$. The participants in the *AHM* condition ($M = 17.00, SD = 1.52$) achieved more points than participants in the *BASE* condition ($M = 15.95, SD = 0.47$). Thus, my hypothesis can be accepted.

A closer look at the two scenario parts shows that this effect only occurs in the *sweets* part and not in the *information* part (see fig. 10.38). The *T-Test* with Bonferoni correction shows no significant differences in the mean of the **information task performance** TP_{info} between the conditions, $t(37.23) = -0.64, p_{adj} = .94, CI = [-1.03, 0.54], d = .21$. The participants in the *AHM* condition ($M = 7.65, SD = 1.14$) achieved not statistically more points than participants in the *BASE* condition ($M = 7.40, SD = 1.31$). As can be seen in the plot (fig. 10.38), the

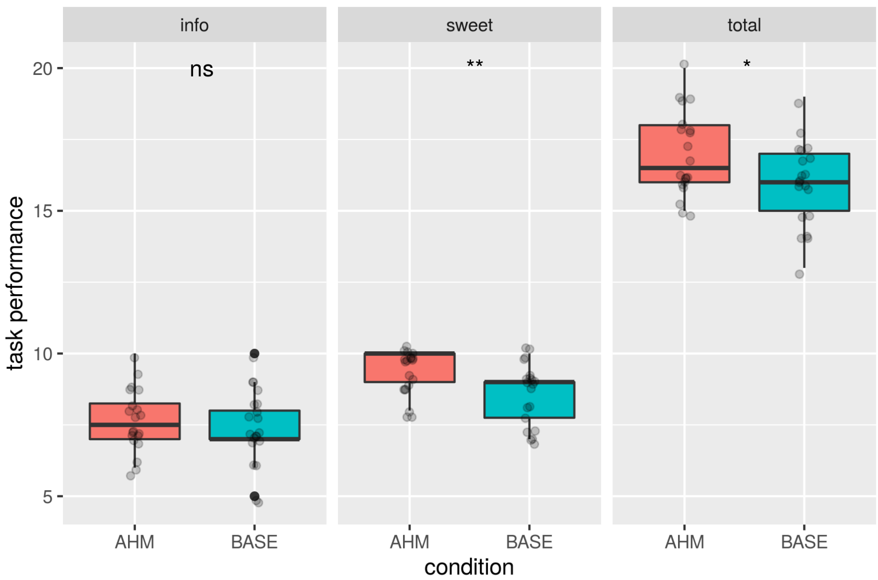


Figure 10.38: Task performance during the information part, the finding rate of sweets, and the performance in total for the *BASE* and *AHM* condition.

sweets task performance is not normally distributed. The assumptions of the parametric test are not met. Therefore, the *WR-Test* with continuity and Bonferoni correction is used. It indicates that scores in the *AHM* condition are significantly higher than in the *BASE* with a medium effect $W = 110.5, p_{adj} = .022, r = .40$. The participant in the *AHM* $(M = 9.40, SD = 0.75)$ condition achieved on average statistically more points in the sweets part than participants in the *BASE* condition $(M = 8.55, SD = 1.10)$.

Side effects on the Interaction

Next, the side effects on the interaction, regarding the subjective ratings of the agent and the *VFoA* of the human interaction partner are investigated.

SUBJECTIVE RATINGS In this section, I evaluate the subjective ratings, consisting of the Godspeed questionnaire, MOS ratings of speech quality and the qualitative interview after the interaction. First, I take a closer look at the **five key concepts** of the *GQS*: anthropomorphism, animacy, likability, perceived intelligence, and perceived safety. Figure 10.39 visualizes the scores after the *information* and the *sweets* part. The agent receives in both conditions high values for likability $(M > 3.7)$ and intelligence $(M > 3.6)$ but rather low values for anthropomorphism $(M < 2.7)$ and safety $(M <= 2.9)$. The differences in the scores in the five key concepts between the *AHM* and the *BASE* condition are not significant $(p > 0.5)$ for both interaction parts. The exact test statistics can be found in table G.15.

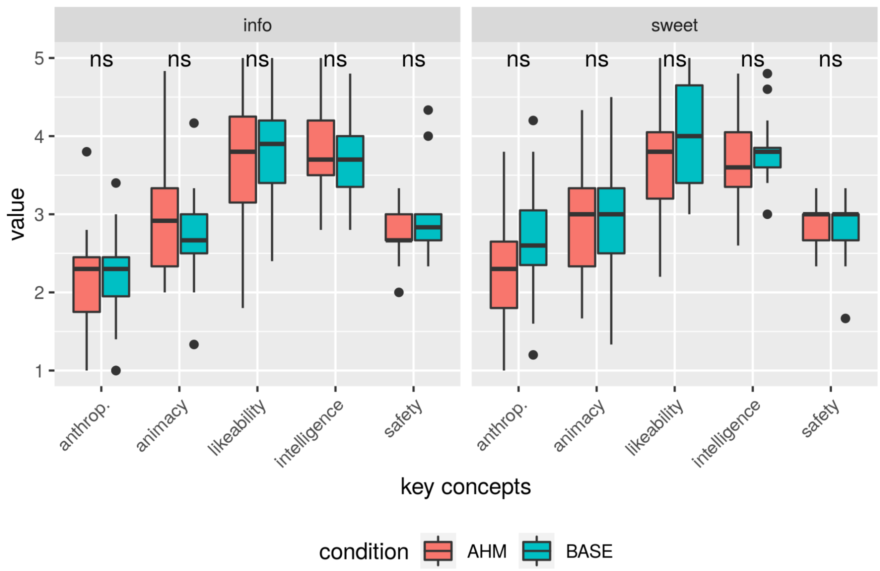


Figure 10.39: Subjective ratings in the five key concepts between the baseline and the hesitation condition after the *information* and *sweets* part.

Figure 10.40 allows a closer look at the ratings of the **voice quality**. It visualizes the MOS-values after the *information* and *sweet* part for each condition. The MOS-values in the *AHM* condition ($M_{info} = 2.9, M_{sweets} = 3.35$) do not significant differ from *BASE* condition ($M_{info} = 3.35, M_{sweets} = 3.35$), neither after the *information*, $W = 256.5, p = 0.107$, nor after the *sweets* part, $W = 197.5, p = 0.954$. In

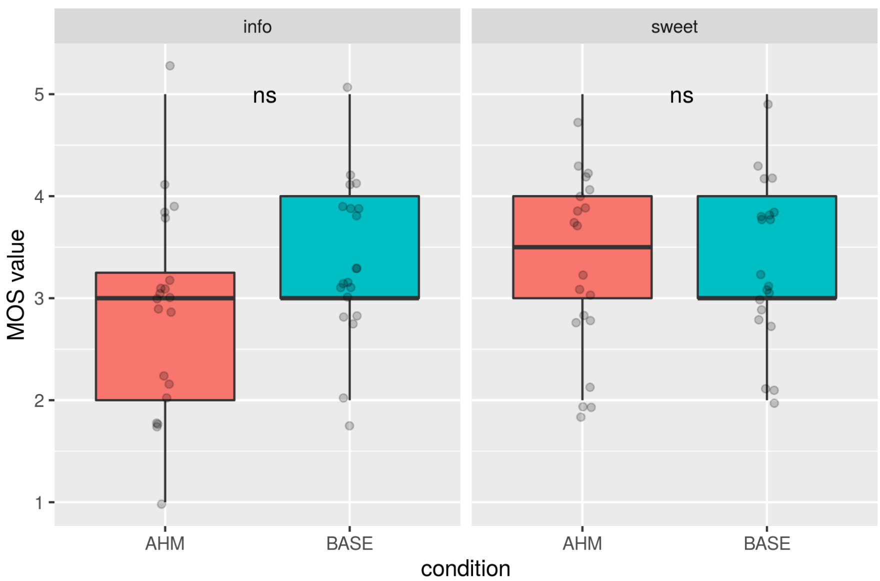


Figure 10.40: Subjective ratings of the voice quality on a Likert scale (5-very good; 1 - very bad).

contrast to the previous experiment in *EC4*, the participants rate the voice quality not significant different between the two conditions after the interaction.

Next, I discuss the qualitative evaluation of my *AHM* and the results of the **semi-structured interview**. In the comments at the questionnaire, only one participant in each condition give negative feedback regarding the voice quality. Both participants found the voice very monotone and staccato. At the end of the study, I asked the participant five questions, to receive feedback of the intervention strategy (example 10.5.1). The main question was whether they have the impression that Flobi adapted its behavior to them or not. In the *BASE* condition, eight of 20 participants had the feeling that Flobi adjusted its behavior to them. Example comments are shown in example 10.5.2. Seven of them noticed that Flobi follows them with its gaze, for example participant *VP12*. The last participant [*VP17*] hat the impression that Flobi adapted its language to them. The subject could not further explain what kind of language adaptions she means. The other 13 participants did not have the feeling that Flobi adapted its behavior. Several participants commented that Flobi was not responsive, exemplary depicted by participant *VP04*.

Example 10.5.2: Example answeres by participant in the *BASE* condition on the question if and how Flobi adapted its behavior.

VP12: *"Blick gezielt ausgerichtet, dahin geguckt, wo ich bin."*

VP17: *"Im Sprachlichen hat er sich mir angepasst und verständlich und ruhig erklärt. Das führte zu einer entspannte Atmosphäre"*

VP04: *"Nein, nicht wirklich, er hat nicht auf mich reagiert. Er hat einfach seinen Text erzählt und ich hab zugehört. Er ist nicht auf mich eingegangen."*

VP12: "Targeted gaze, looked where I am."

VP17: "It adapted its language to me. It explained clearly and calmly. Thad led to a relaxed atmosphere."

VP04: "No, not really. It did not respond to me. It just told its text and I listened. It did not cater to me."

In the *AHM* condition, the participants answered differently. All participants in the *AHM* condition had the feeling that Flobi adapted its behavior to them. Each subject noticed that Flobi used repetitions without any specific inquiries. Most subjects realized, that they had to look at specific points in the apartment in these situations, e.g., the comment by participant *VP59* in example 10.5.3. Ten of the 20 participants noticed that Flobi used pauses. Four of the participants commented on it without any specific questions, e.g., participant *VP59*.

Example 10.5.3: Example answeres by participant in the *AHM* condition on the question if and how Flobi adapted its behavior.

VP59: *"Flobi hat wiederholt, wenn ich irgendwo bestimmtes hingucken sollte, was sehr cool war [...]"*

VP59: *"[...] Pausen waren hilfreich, um zu merken, dass etwas wichtig ist."*

VP59: "Flobi repeated when I was supposed to look somewhere, which was very cool [...]"

VP59: "[...] pauses were helpful in realizing that something was important."

Six additional participants reported that they noticed pauses when explicitly asked about it. While the participants understood the moment of repetitions very well, they could not formulate such a "rule" for the pauses. No participant mentioned lengthening or hesitation vowels. At request, only three participants reported to have heard a hesitation vowel.

Regarding the subjective rating of these interventions, only four of the 20 participants stated that they dislike this behavior in general and that it rather bothered them. Exemplary comments can be found in example 10.5.4 by participants *VP43* and *VP53*.

Example 10.5.4: Example answeres by participants in the *AHM* condition on the question how they like this behavior.

- VP43: *"Freundlich, aber wie ein Roboter. Er hat gestockt zwischendurch und Wiederholungen haben gestört."*

- VP53: *"Wiederholungen waren sehr auffordernd und teilweise überflüssig."*

o VP56: *"Wenn ich nicht hingeguckt habe, wo hin er es wollte, hat er wiederholt, das hat mich sehr verwirrt; Gleiches beim 'Ähm', nur die Pausen haben geholfen, dass ich mir Dinge besser merke."*

o VP48: *"Wiederholungen haben mir im ersten Moment geholfen, mich an die Dinge zu erinnern, danach waren sie unnötig; 'Ähm' und Pausen sind mehr wie eine echte Interaktion und gut."*

\+ VP42: *"Flobi hat sich wiederholt und das hat dabei geholfen, den Fokus auf die Dinge zu finden, er war sehr freundlich."*

+ VP49: *"Flobi hat sich angepasst, wie es ein Mensch tun würde. Mit seiner Stimme hat er pausiert oder Dinge wiederholt, was geholfen hat, es sich etwas besser zu merken."*

+ VP45: *"sehr natürliche Interaktion, es kam nicht steif rüber; Wiederholungen waren sehr natürlich, 'ähm' hab ich gehört und es war sehr natürlich"*

- VP43: "Friendly, but like a robot. It halted from time to time and the repetitions disturbed."

- VP53: "Repetitions were very demanding and sometimes unnecessary."

o VP56: "If I didn't look where it wanted, he repeated, that confused me very much; Same with the 'Uhm', only the breaks helped me remember things better"

o VP48: "Repetition helped me remember things in the first moment, after that, they were unnecessary. 'Uhm' and pauses are more like real interaction and good."

+ VP42: "Flobi was repeating himself and that helped me to focussing, he was very friendly."

+ VP49: "Flobi has adapted like a human would do. With his voice he paused or repeated things, which helped to remember it a little better."

+ VP45: "Very natural interaction, it didn't come across as stiff; Repetitions were very natural; the 'uhm' - I heard it, and it was very natural."

Another five participants gave a more nuanced opinion, such as participants *VP56* and *VP48*. All but one negative feedback was regarding the repetitions, which was perceived as annoying over time. One participant gave negative feedback about the pauses and the hesitation vowel. For him, Flobi's behavior was unclear and not understandable. The other participants (11 of 20) gave rather positive feedback, e.g., by participants *VP42* and *VP49*. Most participants liked the adaptive behavior. Reasons against it were mostly regarding the insisting perception of the repetitions. Interestingly, several participants did not notice the pausing or the hesitation vowel.

Several participants had the impression that there was a difference in Flobi's behavior between the first (*information*) and the second (*sweet*) part. Upon request, they could not explain it further, but some noticed that Flobi adapted more to them in the second part.

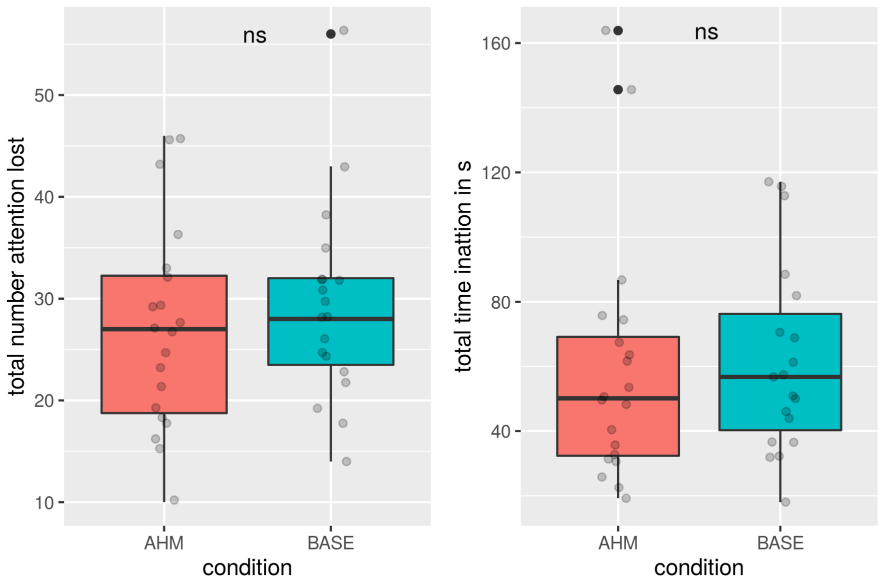


Figure 10.41: Distribution of *VFoA* moves away.

VISUAL FOCUS OF ATTENTION Figure 10.41 visualizes the distribution of the *VFoA* of the different conditions. As can be seen in the plots, participants in the control group did not look away significantly more often or longer in total. The number of look aways does not differ significantly between the *BASE* ($M = 29.26$, $SD = 9.59$) and the *AHM* condition ($M = 27.05$, $SD = 10.17$), $t(37.01) = -0.70, p = .489, CI = [-8.62, 4.19], d = -0.22$. Also, there is no significant difference in the overall time participants look away from the agent or the current *FoD* between the *AHM* ($M = 50.1s, SD = 38.0$) and the *BASE* condition ($M = 56.8s, SD = 29.5$); $W = 164, p = .478, r = -0.71$.

10.5.4 *Discussion and Lessons Learned*

The results of this experiment have implications for my *AHM* and the corresponding hypotheses.

TASK PERFORMANCE HYPOTHESIS: My hypothesis that participants perform better in the task with the *AHM*, measured as post-interaction information recall, is confirmed again. Interestingly, this effect is only observed in the *sweets* and not the *information* part. An explanation for this could be that these tasks are slightly different. While participants in the *information* part only have to remember Flobi's statements, they can work through the task more practically in the *sweets* part. Especially in the *AHM* condition, the task performance in the *sweets* part has a negative skewness. This can lead to a ceiling effect because the independent variable of the condition may no longer have an effect on the task performance. Participants simply can't find more than ten sweets. Remembering the hiding places of sweets seems to be easier for the participants in both conditions, which can be seen on the higher task performance in the second part regardless of the condition. Therefore, an explanation could be that the AHM only leads in such more practical tasks to better performance. Another explanation could be the awareness of the task. In the first part, participants were not explicitly advised that they should remember Flobi's statements, even if they may be primed by the memory test at the beginning. In the *sweets* part they may be more aware of their task to remember the places. So, the awareness of their current task might influence the task performance itself and the positive effect of the AHM. Furthermore, the participants could be more motivated in the second part due to the reward of finding sweets. Another reason might be, that the participants had time to adapt their behavior to Flobi's re-/actions and familiarize themselves with it. After this acclimatization phase, the user may benefit better from the adaptive behavior of the agent. In this case, the better task performance would also occur in the *information* part, if the order of the parts would be swapped. It is still questionable, why the positive effect on the task performance only occurs in the *sweets* part.

SIDE EFFECTS: To investigate the difference between the *information* and the *sweets* part further, I additionally analyze the subjective ratings Figure 10.42 visualizes the same data grouped by condition. The means and standard deviations for all values can additionally be found in table G.14. This way, the different ratings of the same

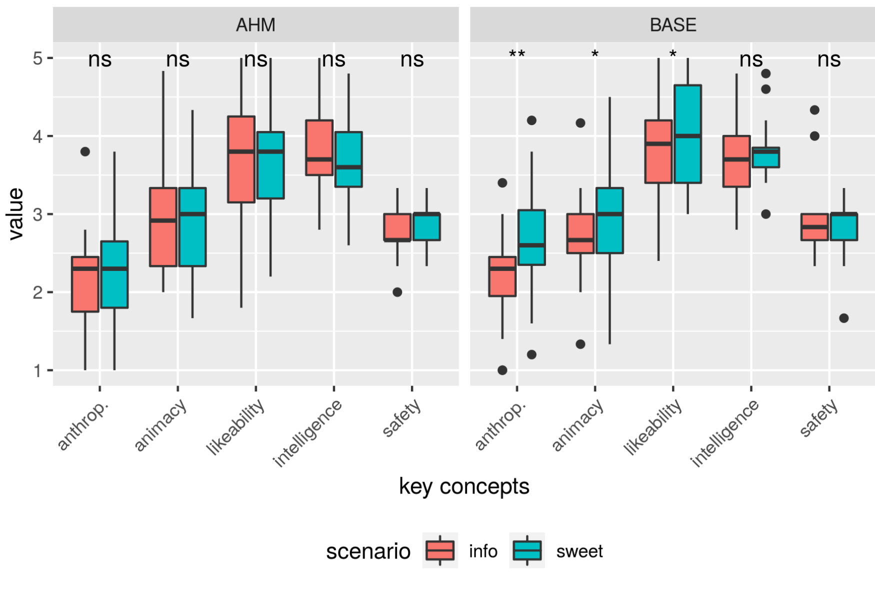


Figure 10.42: Subjective ratings of the five key concepts between the *information* and *sweets* part for each condition.

agent between the two interaction parts can be better observed. Participants in the baseline rated the agent significantly different between the *information* and *sweets* part in the key values *anthropomorphism*, *animacy* and *likability*. The agent perceived higher *anthropomorphism* scores after the *sweets* part (M=2.69, SD=0.76) than the *information* part $(M = 2.20, SD = 0.61)$, $t(19) = -4.31, p_{adj} > .001, CI = [-0.73, -0.25]$. Furthermore, it perceived higher *animacy* scores after the *sweets* part $(M = 3.01, SD = 0.763)$ than the *information* part $(M = 2.71, SD = 0.57)$, $t(19) = -2.61, p_{adj} = .034, CI = [-0.54, -0.05]$. And it also perceived higher *likability* scores after the *sweets* part $(M = 4.01, SD = 0.70)$ than the *information* part $(M = 3.81, SD = 0.69)$, $t(19) = -2.41, p_{adj} = .048, CI = [-0.41, -0.03]$. There was no significant difference for the key values *intelligence* and *safety* between *information* and *sweets* for the *BASE* condition. Interestingly, for the *AHM* condition, there was no significant difference in the scores for any key values between *information* and *sweet* part $(p > .05)$. All exact statistics can be found in the appendix table G.16.

To evaluate the difference between the ratings after the *information* and *sweets* parts, the non-parametric *WSR-Test* is used. It shows that the user in the *BASE* condition do not rate the voice quality significantly different after the *information* and *sweets* part $(M_{info} = 3.35, M_{sweets} = 3.35)$, $V = 45.5, p_{adj} = 1.0$. However, in the *AHM* condition, the voice quality is rated significantly higher for the *sweets* part $(M_{sweets} = 3.35)$ than in the *info* part $(M_{info} = 2.9)$, $V = 10, p_{adj} = .033, r = -0.25$. Interestingly, the user in the *AHM* condition rated the

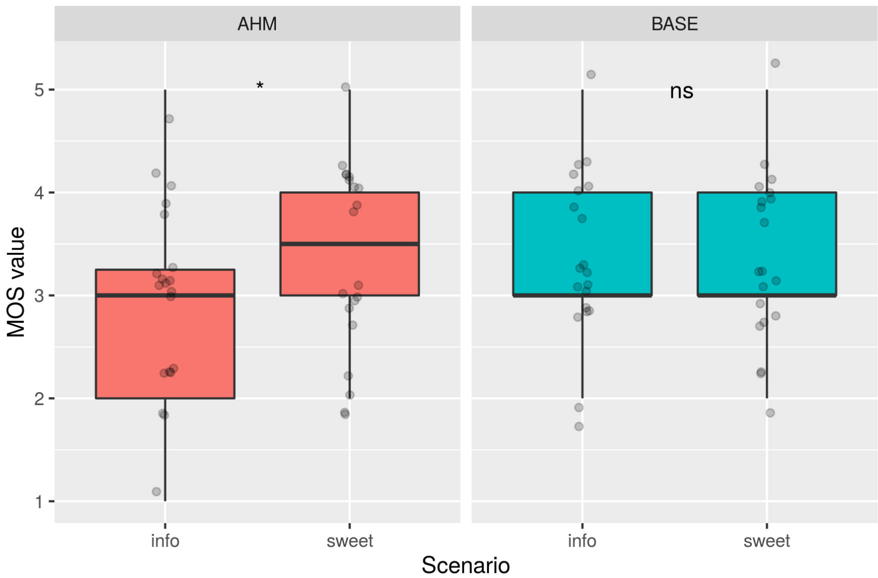


Figure 10.43: Subjective ratings of the voice quality on a Likert scale (5-very good; 1 - very bad).

voice quality slightly worse after the first interaction part.

Reasons for this could again be that participants had more time to familiarize themselves with the agent, their motivation may was higher in the second part, or the task itself could be more enjoyable. In addition, Flobi was indeed a little more interactive in the *sweets* phase. It reacted to speech with head nodding during the finding phase. Surprisingly, this difference could not be observed for the *AHM* condition. For the ratings of the voice quality, the results were inverse. Participants rated the voice quality in the *AHM* condition differently for the *information* and *sweets* part. The users in the *AHM* condition rated the voice quality slightly worse after the first interaction part. This could be because they needed some time to familiarize with this behavior. It is possible that, after some time, the participants habituate to this behavior and don't notice it anymore. Otherwise, it is also possible that they adapt their behavior, which no longer requires further hesitation actions.

In addition, the positive effect on the inattentiveness of participants from the first pilot study could not be reproduced. It is still questionable whether it is because of the different hesitation strategies or because of the more complex interaction scenario. In contrast to the first pilot, this scenario had various *FoD*. A possible explanation for these results may be that the positive effect on the inattentiveness only occur in very static interactions without various discourse changes, or throughout longer phases with the focus on one object or location. It is possible that the discourse changes itself are regaining the attention of the interaction partner. In general, all participants were very attentive and followed the instructions of Flobi very well, which can be seen on the high task performance and low inattentiveness values.

ATTENTION-HESITATION MODEL: These results have implications for my *AHM* itself. The moment to react seems to be appropriate. Participants did not complain inappropriate pauses, like in the previous studies. In addition, the hesitation strategies were well-received. Most participants in the *AHM* condition liked the adaptive behavior, even though not each participant even noticed all characteristics of the strategy. This itself is positive, we do not always consciously perceive hesitations in *HHI*. Only the repetitions are sometimes perceived as annoying. There are several possible explanations for this result. Firstly, the strategy of repetitions itself may be annoying, regardless of the current performance. Some users could find such behavior patronizing in general. Another option may be that the timing was not appropriate. The pause between the repetitions could not have been long enough for participants to react appropriately. Apart from this, there is always the possibility of wizard errors. These could be manifold, in terms of pressing a button at the wrong time or misinterpret the participant's current focus of attention.

To answer my research questions for this experiment, the chosen study design was appropriate. I could investigate my hypothesis whether my enhanced *AHM* can increase the task performance in more practically task during the *sweets* phase. Furthermore, the investigation of the task performance in an information task was possible. Only an effect on the task performance in the *sweets* and not in the *information* part could be observed. Even though the two parts are kept as similar as possible, the task itself, the previous experience with Flobi and the *AHM*, and the locations and objects differ between the parts. The comparison of the task performances in these different parts need to interpreted with caution, because of these differences. Furthermore, the subjective ratings of the agent and the interaction are examined. Through the different questionnaires and the qualitative interviews, I could get further insights into the effect of my model.

FURTHER RESEARCH QUESTIONS: This research has thrown up many questions in need of further investigation, e.g., how the participant's memory performance in the pretest influenced the task performance. To get a better understanding of the influence of the general memory performance, fig. 10.45 visualizes the different task performances over the participant's memory performance in the pretest. From this data, we can see that especially in the *AHM* condition the

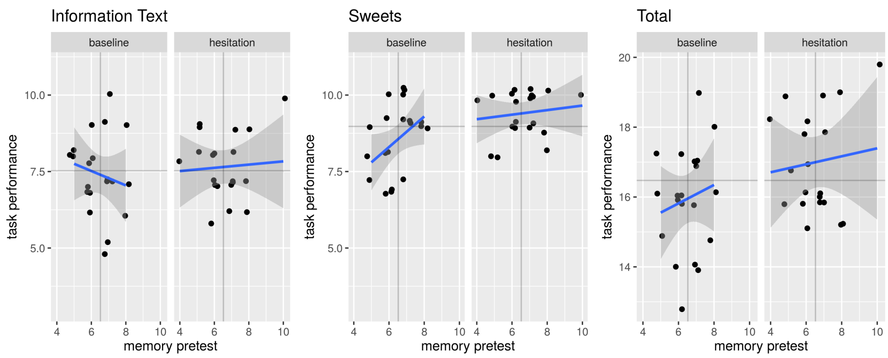


Figure 10.44: Task performance during the information phase, the finding rate of sweets, and the performance in total for the *BASE* and *AHM* condition.

task performance of *sweets* seems to be independent of the results of the pretest. A median-split using participants memory pretest is conducted. This results in two subgroups, participants with memory performance in pretest $TP_{pretest} <= 6$ are marked as *low* and participants with $TP_{pretest} > 6$ marked with *high memory*[4]. The density plot for these subgroups in fig. 10.45 indicates, that especially participants with low memory may benefit from the *AHM*. This leads to interesting further research questions, if participants with lower memory performance can benefit more from the *AHM* than participants with higher memory performance. This need to be further investigated. Other questions, which arise are (1) Is there a difference between the performance in a questionnaire and finding sweets as measurements

4 Such median-splits are not without controversy. Therefore, I only used it for further hypothesis generation and exploration.

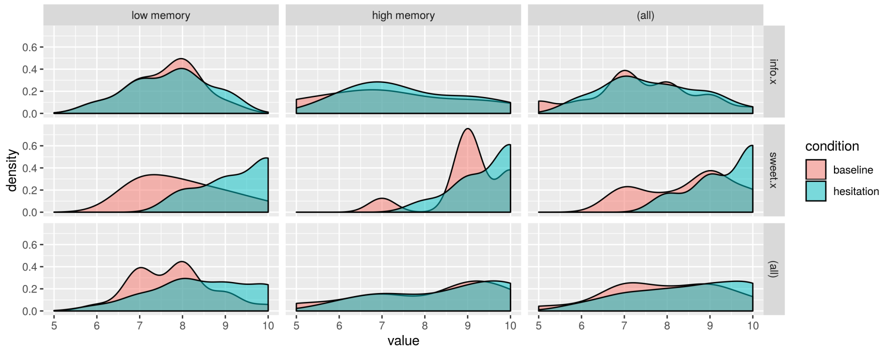


Figure 10.45: Task performance during the information phase, the finding rate of sweets and the performance in total for the *BASE* and *AHM* condition.

of task performance? (2) Which characteristics of participants have an influence on the effect on the task performance of the *AHM*? (3) Which characteristics of participants influence the subjective ratings of the *AHM*?

To sum up, I could show that participants perform better in the task, with no worse subjective ratings of the agent's voice quality or the agent itself. The most interesting finding was that the task performance only differ in the *sweets* part and not in the *information* part.

11 COMPARISON OF THE FIVE EVALUATION CYCLES

In the last chapter, five *EC* of the *AHM* are conducted. Throughout these iterations, I investigated my hypothesis that the Attention-Hesitation Model (AHM) increases the task performance in human-agent interaction. Therefore, the *AHM* was iteratively improved. In this chapter, the results of the five *ECs* are compared and further discussed. To this end, the used features in each *EC* are summarized (section 11.1). Afterwards, the results regarding the task performance hypothesis are discussed in section 11.2, followed by a discussion about the corresponding side effects (section 11.3). The chapter concludes with a final *AHM* (section 11.4) and a summary of the third part of this thesis (chapter 12).

11.1 FEATURES OF ATTENTION CONCEPT AND HESITATION STRATEGY

Table 11.1.1: Overview of different features utilized for the attention model and the corresponding intervention hesitation strategy.

	Attention concept				Hesitation strategy				
		Task			Re-attention			Highlight	
EC	Mutual Gaze	Directed Gaze	Discourse History	Task Progress	Unfilled Pauses	Lengthening	Hesitation Vowels	Repetitions	Lengthening
EC1	✓				✓				
EC2	✓	✓	✓		✓			✓	
EC3	✓			✓	✓			✓	
EC4	✓				✓	✓			
EC5	✓	✓	✓		✓	✓	✓	✓	✓

Table 11.1.1 depicts again the features used in the different implementations of my *AHM* in each cycle.

WHEN TO (RE-)ACT: ATTENTION CONCEPT The moment **when** the intervention strategy starts differs between the cycles. While in *EC1* and *EC4* the agent only starts the intervention when the participant looks away from it, in *EC2* and *EC5* the agent starts the intervention when the human does not look at the agent *or* the *FoD*. Furthermore, beside the mutual gaze and task related feature of directed gaze, the discourse history is considered to distinguish between inattentiveness due to engagement or understanding problems. The exploration in *EC3* used a different attention concept. Through the underling task of the interaction—fetching the ingredients during the interaction and not the information recall afterwards—the *AHM* can use the task progress itself as a feature for the attention of the interlocutor. This is not possible in the other cycles, based on the tasks at hand.

HOW TO (RE-)ACT: HESITATION INTERVENTION STRATEGIES In addition, the **how** of the intervention strategy also differs. The *AHM* in *EC1* only uses unfilled pauses (silence) as an intervention strategy, regardless of why they are inattentive. The *AHM* in *EC2* additionally with uses repetitions when applied after the discourse changes. In this model, unfilled pauses are used as a repair mechanism for inattentiveness based on engagement errors, whereas the repetitions deal with inattentiveness based on understanding problems. Whereas in both studies—*EC4* and *EC5*—the agent produces a lengthening followed by an unfilled pause, in *EC5* the agent additionally reacts with hesitation vowels and repetitions.

As a measurement for improving the *HAI* in smart homes, I used an enhanced task performance. However, this should not be at the cost of the subjective ratings of the agent and the interaction itself. To this end, I measured not only the task performance, but also considered side effects. Table 11.1.2 depicts again the measurements of each cycle.

Table 11.1.2: Measurements of task performance and side efects for each cycle.

		Task Perfm.			Side effects					
					Subjective Ratings				vFoA	
EC	**Autonomy**	Information	Prac in Interact.	Prac. post Interact.	Godspeed	MOS	Interview	Free Text	Look Away	Total Time
EC1	WoZ	✓			✓			✓	✓	✓
EC2	Autonomous	✓			✓			✓	✓	✓
EC3	WoZ		✓	✓	✓			✓		
EC4	Autonomous			✓	✓	✓	✓	✓	✓	✓
EC5	Semi-Autonom.	✓		✓	✓	✓	✓	✓	✓	✓

Next, I discuss the results for my hypothesis and the side effects in the interaction on more detail.

11.2 TASK PERFORMANCE HYPOTHESIS

My hypothesis is that the *AHM* leads to a higher task performance. As shown in the overview of my study measurements (table 11.1.2), I measured task performance differently in my cycles. My hypothesis is confirmed in three of the five evaluation cycles, as depicted in 11.2.1. This might be due to the following reasons.

Table 11.2.1: Results regarding the task performance hypothesis and the side effects in the visual attention and the subjective ratings.

		Side Effects	
EC	**Task Performance**	Visual Attention	Subjective Ratings
EC1	-	!	-
EC2	-	-	!
EC3	✓		!
EC4	✓	-	!
EC5	✓	-	-

In the pilot study in *EC1*, the results indicate that participants just guessed the answerers in the post interaction information recall. Here, one problem might be the generally high difficulty of the questionnaire. The second pilot study in *EC2* also shows no effect of the *AHM* condition on the task performance in general, even though there is some indication that embodied information could benefit from it.

In the third pilot study in *EC3* and the two interaction studies in *EC4* and *EC5*, the participants in *AHM* condition achieved higher results than participants in the baseline. In *EC3*, participants make fewer errors in the *AHM* condition by fetching ingredients. Interestingly, the positive effect on the task performance can only be measured in a higher finding rate in the sweets task in *EC4* and *EC5*. In the information part of *EC5*, this effect is not discernible. One explanation for this effect are the different way of measuring task performance. In *EC3*, the task was slightly different and allowed a measurement of task performance during the interaction. Therefore, it plays a special role and is not directly comparable with the other cycles, in which studies the task performance is measured after the interaction itself. To figure out why the task performance increased in some cases but not in others, we need to take a closer look at the study procedures. One main difference between the studies is the number of participants, while *EC1* and *EC2* are pilot studies with relatively few participants, the studies in *EC4* and *EC5* are conducted with 20 participants per condition. However, this does not explain the difference in task performance between the information and the sweets part in *EC5*. Whereas in *EC1* and *EC2* the post interaction information recall is measured through simple questionnaires at the end of the interaction, the information recall in the sweets parts in *EC4* and *EC5* is measured via the finding rate of candy in the apartment. In *EC5* participants had to perform both, the information and the sweets part. Interestingly, the participants achieve significantly higher values for the sweets than the information part in both conditions (*AHM*: $V = 3, p_{adj} < .001, r = -0.54$, *BASE*: $V = 20, p_{adj} < .046, r = -0.54$). One possible explanation could be that the sweet task was easier, even though attention was paid to use similar descriptions in both tasks. Another explanation could be that the motivation of the participants may have been higher in the sweets parts. This can be because the agent requested for high attention or the due to the fact that it promised them that they can keep all sweets they find. A last explanation could be that participants needed some time to accustom themselves to the situation. Interaction with an intelligent apartment via a virtual, embodied agent was, in general, new to participants and particularly its adaptive behavior, which was based on heir manner. The studies in *EC4* and *EC5* both had two phases of interaction (the info and the sweets part). Thus, participants had more time to familiarize themselves with the situation and may have overcome some novelty effects. Randomizing the order of these parts, would allow testing the impact of novelty and human adaption on their performance.

Regarding the characteristics of the *AHM*, the **when to react** differ between studies. The *AHM* in *EC1* and *EC4* uses the same attention concepts. However, only *EC4* supports my hypothesis. Analog, *EC2* and *EC5* evaluated similar attention concepts and only *EC5* supports my hypothesis. Concerning the **how to react**, the effective intervention strategies used an incremental hesitation strategy of unfilled pause introduce by lengthening in *EC4*, and additionally hesitation vowels and repetitions in *EC5*. However, in *EC3*, the hesitation strategy used only pauses without lengthening and additionally repetitions.

11.2.1 *HHI and Attention Hypotheses*

In this section, I compare the results of the task performance with observations known from related work. As depicted in section 2.2.3, researchers investigated the effect of disfluencies from the *HHI* perspective and found that hesitations can be beneficial for the listeners' comprehension. In fact, a beneficial effect on the recall of the entire discourse, not only on the manipulated plot points [FW11]. In contrast to these studies, which investigate the effect on recall in *HHI*, the disfluent system speech is not predefined in my studies. This means, that the behavior of the system is based on the behavior of the human. Furthermore, the studies from the *HHI* perspective are not embedded in a real interaction—which is of course difficult due to the not reproducible nature of *HHI*. However, the interaction with an artificial agent makes it possible to embed such experiments in an interaction and still carefully and reproducible manipulate the interaction strategy.

Although this is an interaction with an agent, it may be possible to draw conclusions for *HHI*.

The higher task performance in *EC3*-*EC5* is a contribution to the discussion of why disfluent speech can have a positive effect in comprehension or memory performance. Fraundorf and Watson [FW11] hypothesize that:

H1 *predictive processing hypothesis*: Listeners can use hesitations to predict what they will hear next.

H2 *attentional orienting hypothesis*: hesitations (re-) orient the listeners' attention towards upcoming speech.

H3 *processing-time hypothesis*: Listeners have more time to process the information.

Due to the fact, that the intervention strategy is started depending on the participants' behavior, the listener could not predict what they heard next. Therefore, my findings do not contribute to for the *predictive processing hypothesis*. Only the repetition part of *EC5* can be seen as an argument to support this hypothesis. However, that does not explain the positive effect of the study in *EC4*. As the human has more time to process the presented information, my findings are in line with the *processing-time hypothesis*. Particularly, in *EC3*, the agent only continues speaking when the sub-goal of the task is completed. Furthermore, in *EC4*, the agent only continues speaking when the human looks back at the agent. In this case, the listener can take as much time as needed to look at the current *FoD* and to understand the spoken words. In *EC5*, the agent only starts the intervention strategy, when the listener neither looks at the agent nor at the current *FoD*. This is in line with the *attentional orienting hypothesis*.

To sum up, the repetitions during *FoD* changes provide arguments to all the hypotheses, depending on the type of intervention strategy and the moment of the starting. However, the lengthening and (un-)filled pauses only support the *processing-time* and *attentional orienting hypothesis*.

11.2.2 *Comparison to other HAI Experiments*

As depicted in section 3.2, several works in the HAI literature focus on disfluent speech in *HAI*. Few of them evaluate the effect of disfluent system speech on the task performance of the human interaction partner. Kousidis et al. evaluated a situated in-car dialogue system with a non-adaptive strategy in an interaction study with 17 participants [Kou+14]. During a driving simulation, the dialogue system presented information about calendar entries. After the information was given, the participants had to answer a short false-true question about the presented information to access their memory recall. In the interrupting condition, the system paused its information presentation during lane changes, whereas in the non-adaptive condition the system did not react to these external events. Kousidis et al. found, that both tasks—lane change maneuver and memory task—benefit from the adaptive version. The participants had a better information recall in the adaptive version and performed a better lane change when the system is silent [Kou+14].

In my studies, I could not reproduce the positive effect on the information recall in all studies. Possible reasons are discussed at the beginning of section 11.2. However, there are some important differences between the in-car and my investigations. First, the interaction foci differ. In the in-car scenario, the main activity is the driving task, the system interaction is a distraction. In contrast, in the *AHM* studies, the interaction with the system is the main task. Nevertheless, both scenarios operate with the same concept of divided attention (see section 2.1.1). The consequences for attention models are discussed late in section 11.4. Another main difference is the chosen intervention strategy. Whereas in the in-car study the dialogue system always repeats the last utterance, in the *AHM* studies, repetitions are only used in during discourse changes (in *EC2* and *EC5*) or after explicit request from the human interaction partner (in *EC3*). More important, these repetitions are only contained subjective references (e.g., "to your left side") and do not contain further information about the locations or objects in *EC2* and *EC5*. In contrast, in the in-car study, content information, which is queried afterwards, is repeated as well. Thus, it remains unanswered if the better memory performance in the in-car study is rooted in the pauses, the repetitions, or both. In *EC4*, the intervention strategy do not include repetitions, thus the positive effect on the task performance must have other reasons. Therefore, my *AHM* studies provide further evidence for the positive effect of disfluent speech on the task performance.

Another work focussing on the effect of disfluent speech of an agent on the task performance is the study by Ohta et al. [OKN14]. They investigated the effects of filled and unfilled pauses in a tourist-guiding task. In their experiment, the participants listened to an agent which explained how to travel from a departure point to a destination. During the explanation, the participants solved simple calculation tasks as a distraction. The agent performed either a filled pause, an unfilled pause, or no pauses after an information chunk. The hesitations are performed independently of the listener's attentions at these predefined points of interaction. Furthermore, the condition changed in the second half of the interaction. It's important to mention that the speech is prerecorded with an actor and modified afterwards. The authors reported a positive effect on the user's comprehension (TP) only for the filled pause condition ($\mathsf{TP}_{\mathsf{WithoutPause}} = 36.8\%$; $\mathsf{TP}_{\mathsf{FilledPause}} = 43.4\%$; $\mathsf{TP}_{\mathsf{Silent}} = 34.0\%$. Unfortunately, they did not present any statistical analysis. Thus, a comparison remains difficult. Nevertheless, the tourist-guiding study and my *AHM* studies use the same concept of divided attention. In contrast to most of my studies (excerpt for *EC3*), the moments of the filled and unfilled pauses are predefined and do not depend on other—for the system external—conditions. Another important difference is the used language and respectively the participants' cultural background. Whereas in my studies, the agent speaks German, the agent in tourist-guiding study speaks Japanese. Whether (un-)filled pauses are perceived similar or have the same functions in different languages and cultures, is a fascinating research question, which is out of the scope of this thesis.

The work of Palinko et al. focuses on the effect of disfluent system speech on the listener in a slightly different way. In a dictating scenario, the robot produces an unfilled pause, until the human has finished the sentence of the dictate. Only when the participant gazes at the robot, it continues speaking. In contrast to the other studies, this dictation scenario does not involve a second or disturbing task. This is comparable with the scenario in my *EC3*. Therefore, it does not use the concept of divided attention. The authors found no effect on task performance (in this case, the number of errors in writing) [Pal+15]. In contrast, the participants in the *AHM* condition in *EC3* performed better in the task as the baseline.

As a summary of the results of the studies, the following conclusions for my *AHM* can be drawn. Regarding the effect on the task performance, evidence for higher task performances could be found for filled pauses in both the tourist-guiding scenario by Ohta et al. [OKN14] and my *EC5*. Additionally, the in-car study provides evidence that unfilled pauses, followed by repetitions, can also improve the task performance. In most of the studies that only use silence (*EC1*, *EC2*, and in the dictating scenario [Pal+15]) a positive effect on the task performance could not be observed. The study in *EC3* is an exception in my investigations, because of the different interaction scenarios. However, adding lengthening as an introduction to the unfilled pause does improve the task performance (*EC4*, *EC5*). The moment at which the intervention is started is rather different between the studies with improved task performance. In the tourist-guiding scenario, the system starts the intervention at predefined points of the speech, in the in-car study it started by a predefined external event (when a lane change need to be performed), and only my *AHM* reacts to the behavior of the user.

11.3 SIDE EFFECTS

In this section, I take a closer look at the side effects in the different studies.

11.3.1 *Subjective Ratings*

I investigated whether the *AHM* influenced the subjective rating of the agent. Therefore, I used the *GQS* to access the five key values *anthropomorphism*, *animacy*, *likability*, *perceived intelligence*, and *perceived safety* in all studies. Furthermore, I assessed the speech quality through the MOS value in *EC4* and *EC5*. In all studies, the participants had the opportunity to leave free form comments in the post-hoc questionnaire or in the interviews afterwards, to allow more qualitative insights.

GODSPEED QUESTIONNAIRE: Figure 11.1 gives an overview of the results of the five key concepts *anthropomorphism*, *animacy*, *likability*, *perceived intelligence*, and *perceived safety* for all *AHM* studies. In general, Flobi received rather high values for *likability*, *perceived intelligence* and *perceived safety* and low values for *anthropomorphism* and *animacy*. In the godspeed evaluation of the five *EC*, it is apparent that *EC2* is the only study with significant differences between the *BASE* and the *AHM* condition. Especially, the *likability* of the agent decreases significantly in this experiment. A comparison of these results with the other studies illustrates that even though the Flobi in the *AHM* receives low ratings, the Flobi in the *BASE* condition of the same study receives rather high likability values. Similarly, distributions can be observed for the key concept *intelligence*. Interestingly, the agent in the studies in *EC1* and *EC4* is rated less intelligent than the agent in *EC5*, regardless of the condition or the interaction part (info/sweet). In contrast to this, the agent in *EC5* is perceived as less safe.

SPEECH QUALITY: The speech quality was only assessed in the main studies in *EC4* and *EC5*. The results of the MOS values are depicted in fig. 11.2 (middle and right plot). Whereas in *EC4*, the participants rated the speech quality significantly lower in the *AHM* condition than in the baseline, no such effect could be observed in *EC5*. This could have different reasons. First, *EC4* was fully autonomous. Thus, errors in the attention detection could have lead to unintentional behavior of the agent, which may have influenced the ratings. Furthermore, the implementation was improved for better sound quality. This could also have influenced the speech quality ratings. Lastly, the *AHM* was incrementally improved between the studies. The moment the agent started its intervention strategy as well as the strategy itself were therefore different and could affect these synthesis ratings.

In parallel to the study performed in *EC4*, we conducted an online crowdsourcing study with similar stimuli to compare the results of the speech quality assessment [Bet+18]. Figure 11.2 on the left visualizes the corresponding MOS values. The baseline without hesitations was rated similar in all studies, whether taken online or in interaction. For the speech synthesis with hesitation—the *AHM* condition— the picture

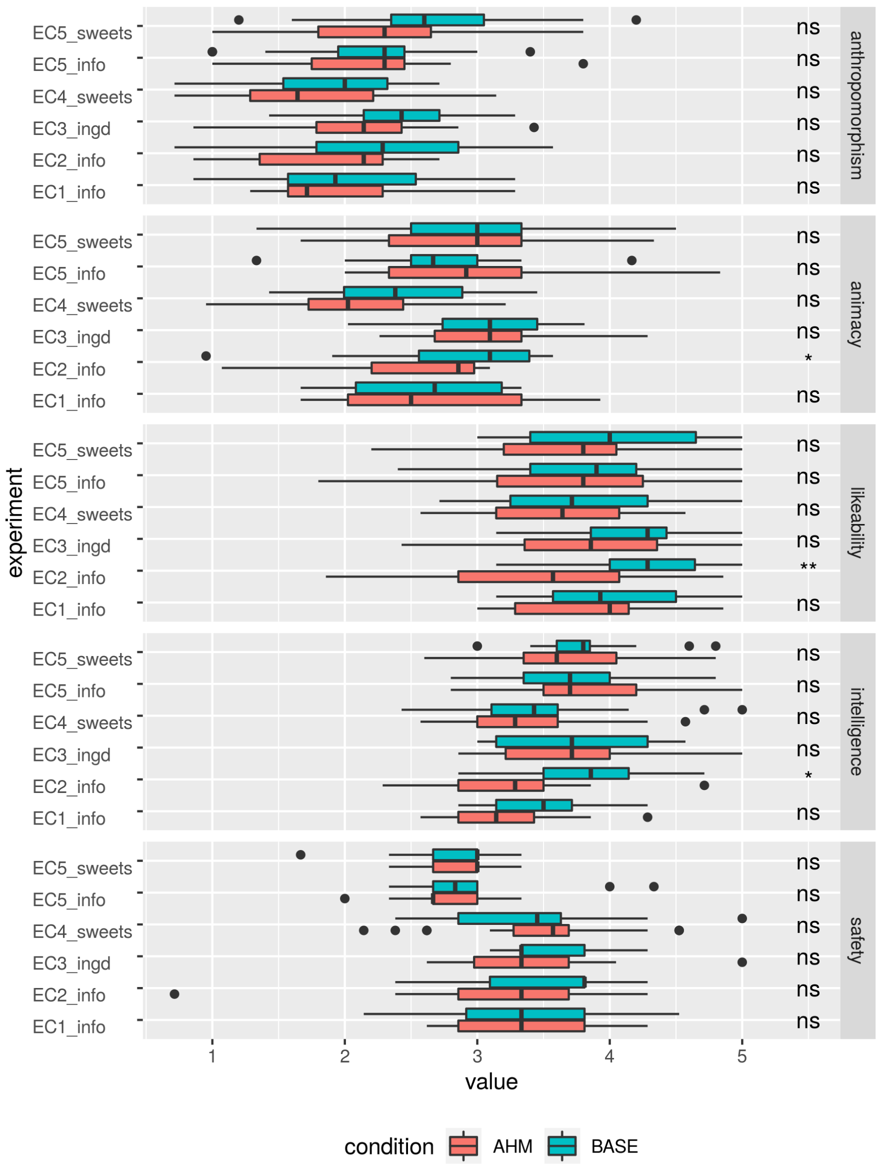


Figure 11.1: Subjective ratings of the agent for the five key concepts *anthropomorphism*, *animacy*, *likability*, *perceived intelligence*, and *perceived safety* over all studies.

looks different. A comparison of the results of the online and the interaction study (fig. 11.2 in the middle) shows that the speech quality was rated differently in the two settings. Whereas in the crowdsourcing study the hesitation and baseline synthesis are not rated significantly different, the *AHM* condition in the interaction study *EC4* performs significantly worse. This may have the following reasons. In the online evaluation, (1) the locations of the hesitations are predefined (always after information chunks), (2) stimuli are prerecorded with the used *Text-to-Speech Synthesis (TTS)* system and modified afterwards, and the most important difference is, that (3) the stimuli are not rated in

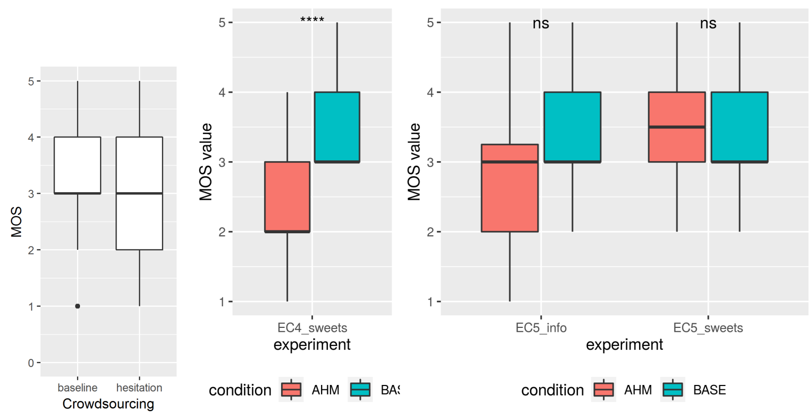


Figure 11.2: Subjective ratings of the voice quality on a Likert scale (5-very good; 1 - very bad). Left: MOS values of an online study [Bet+18]. Middle: *EC4*. Right: *EC5*

interaction. This means that participants in the online study could not experience the same interaction as participants in the interaction study. The results show that an evaluation of system speech without embedding it into a corresponding interaction leads to different results. Therefore, I argue again that effects on participants need measured in a real interaction, rather than rates based on video and audio recordings. A system speech which quality is rated positively in an online study can perform differently in interaction, because of influences of other system components, the agent's outward appearance, the interaction itself, or the participants' behavior.

POST-HOC STATEMENTS AND INTERVIEWS: All participants had the opportunity to leave general comments in the questionnaires. Additionally, in *EC4* and *EC5*, I performed semi-structured interviews to collect more information about the perception of the current *AHM*.

The *general comments in the questionnaires* are quite different between my *AHM* studies. In *EC1*, no participant commented regarding the adaptive behavior or the voice quality. One reason could be that the interaction in the wardrobe was quite short.

In *EC2*, approximately the same number of participants leaf negative feedback regarding the voice quality in the comments of the questionnaire. Whereas the participants in the baseline objected the monotonous speech without emphasis, in the *AHM* condition the speech was often described as chopped or staccato. Additionally, the repetitions were judged as very penetrating by one participant. In *EC3*, the incremental information presentation was perceived as less timely, but more suitable regarding length.

In *EC4*, several participants left negative feedback regarding the voice quality. The hesitations were sometimes misinterpreted as sound disruptions instead of hesitations. In the *interviews* of *EC4*, the answers were very more nuanced. Some participants noticed the adaptive behavior and found it useful. Other participants did not understand it as an intervention strategy of the agent and perceived it as sound disruptions. However, the judgement of such behavior was not clear.

The feedback in *EC5* was more positive. Pauses and hesitation vowels were perceived as natural and only one participant disliked the interventions, even though several participants did not notice them as such. The judgement of the repetitions was conclusive—several participants found it useful, whereas others found it rather unnecessary.

COMPARISON TO OTHER HAI EXPERIMENTS: In previous studies in the literature, disfluent system speech is also evaluated subjectively. There, I first consider the in-car [Kou+14] and the tourist-guiding [OKN14] experiments, which I mentioned above. Kousidis et al. asked participants which system behavior they would prefer. Interestingly, only three of their 17 participants preferred the adaptive behavior. This was mainly due to the missing ability to control the adaption strategy. Independently of the behavior of the user and the need for interruption, the systems interrupted itself [Kou+14]. This could be one reason for the difference in the ratings of the intervention strategies between the in-car and my scenarios. In my last studies, post-hoc interviews, most of the participants stated they would like to have an adaptive agent. In contrast to the proposed strategy of Kousidis et al., the *AHM* reacts to the behavior of the user and thus to their special needs and is therefore adaptable to the user. Thereby, the keep the (indirect) control of the initiation of the intervention strategy.

Ohta et al. also asked their 24 participants which system behavior they would prefer in the tourist-guiding scenario. The participants did not clearly prefer any system. Comparing the system *without pauses* with *with filled pauses* or *without pauses* with *with silence*, the preferences for the systems are divided. Only in the head-to-head comparison with the system which produces *silence*, the *filled pauses*' system speech was slightly preferred. In contrast, participants prefer the system with *filled pauses* over *silent pauses* and this over *without pauses*.[1] [OKN14]. Besides the mentioned weaknesses in the reporting of (statistical) results, the authors did not consider any order effects between the different conditions. (Un-) filled pauses may be perceived abnormal after the first part of the interaction was fluent. Due to the small number of participants, those effects could not have been eliminated. As the authors themselves state, the high values for naturalness might be because they used professional actors for the utterances and not a *TTS* system. Furthermore, they reported that participants had problems distinguishing unfilled pauses from turn ends, which may be a reason for the low listenability values.

1 It should be noted, that the paper does not provide information about the statistical significance of these results.

A comparison of their results with the results of the *in-car* and my *AHM* studies, create a uniform picture. A simple break-of of the system speech makes it difficult for the listener to distinguish between a pause, an error in the speech synthesis, and a turn-end. This was reported in the tourist-guiding and some *AHM* studies. However, lengthening before an unfilled pause reduces this problem. In the post-hoc interviews, participants did not agree regarding the preference of a system in any of these studies. Arguments against the disfluent intervention strategy were the loss of control and the distinguishability between unfilled pauses and a turn-ends. Both arguments can be counteracted with my *AHM*. On the one hand, with the interactive, indirect control possibilities based on human behavioral cues (e.g., gaze) and on the other hand with the use of lengthening before the unfilled pauses are applied.

The within-subject study design of the in-car, dictating, and tourist guiding scenario makes it difficult to deal with the novelty effect. Through the change of the agents' behavior within the interaction, the familiarization is hampered. The repeated measures of the subjective ratings in my last study *EC5* show an improvement in the ratings of the speech quality between the first (information) and the second (sweets) part. This can be attributed to the fact that the participants had time to accustom themselves to the situation and the hesitating behavior of the agent. This is also in line with the results of the task performance.

The previously mentioned fundamental differences between these studies (e.g., concepts of the main task and cultural background) are of course also restrictions in the comparability of the subjective ratings.

Based on the collected data, I cannot confirm, that all the *AHM* not influenced the subjective ratings of the agent as a side effect in interaction. In the study in *EC2*, the agent received lower values for some key concepts. In the study in *EC4*, the speech quality was rated low, which could be improved in the last study.

Nevertheless, in the post-hoc statements and interviews, I got a good first impression of the judgement of the different intervention strategies. These results are of course qualitative and not suitable for statistical analyses. However, they allow a better understanding of the participants' perception of the agent's behavior and the interaction itself.

11.3.2 *Visual Focus of Attention*

Furthermore, I investigated if the participants in the *AHM* condition are less inattentive than participants in the baseline, measured by the number of attention lost and the total time of being inattentive—meaning the *VFoA* moves away from the agent or the current *FoD*. This effect could only be observed in *EC1*. In the other studies no effect of the *AHM* on the overall time participants looked away or how often they looked away could be observed. A reason could be the more complex interaction scenarios. In contrast to *EC1*, the *FoD* changed within the interaction in the other studies. This raises the question, if the *AHM* had an effect during theses *FoD* changes.

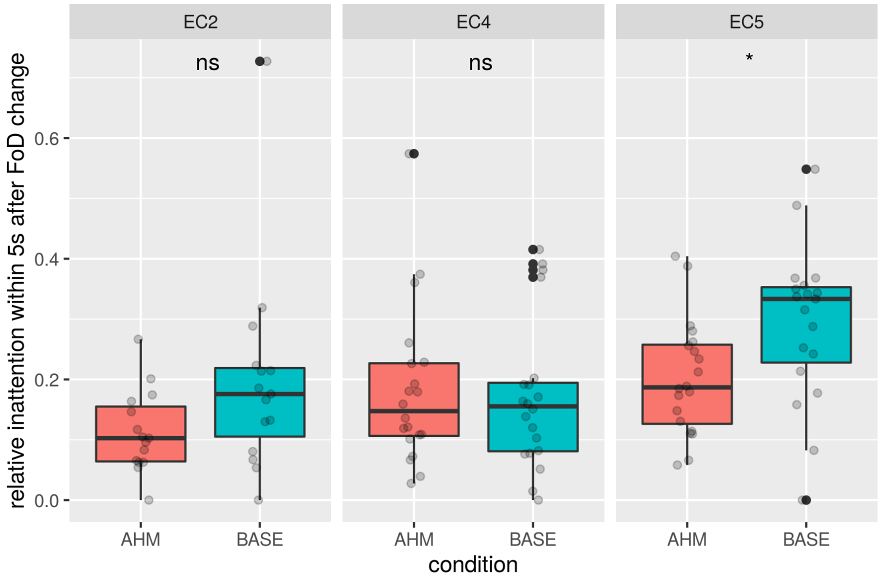


Figure 11.3: Percentage portion participants are inattentive after a *FoD* change (time window=5s) for the tree interactions studies with *FoD* changes.

Figure 11.3 depicts the amount of time participants are inattentive within a 5s time window after a *FoD* change. One can see, that the time participants are inattentive increases for the last study. This could be due to the fact, that the interaction is longer and have the most *FoD* changes. Interestingly, participants in the *AHM* condition are less likely to become distracted after changes of the *FoD*. The disruption draws a similar picture. Figure 11.4 visualizes that participants in the last study are less likely to become distracted during the disruptions if they are in the *AHM* condition. It is difficult to compare these results with other studies, since I could not find any research, which analyses the effect of disfluent system speech in *HAI* on the attention of participants.

During the analysis of my studies, I observed that multiple participants left the interaction space to continue the questionnaire, before Flobi finished speaking. In the last study, Flobi produces an "uhm" when this happens. This behavior made participants come back to Flobi and listen till the end of its explanation. This behavior suggests, that the system could attract attention. If the participant leave the interaction space while the system speaks, it could have two reasons (1) they do not care or (2) they think they can listen well from a distance. The production of the hesitation vowels may act as a social signal that Flobi expects full attention. Furthermore, Flobi indicates

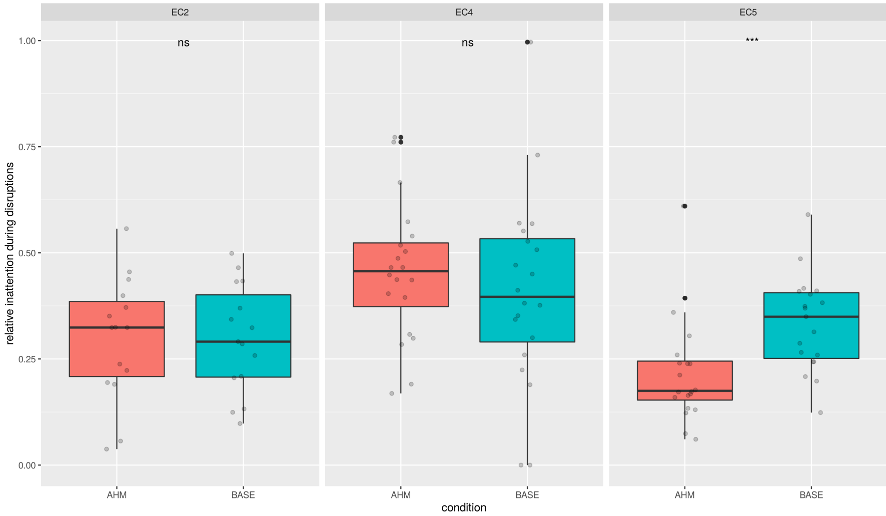


Figure 11.4: Percentage of participants time participants are inattentive during disruptions.

that it "noticed" their inattention. It would be interesting to further investigate this hypothesis in dedicate experiments.

Summarizing my results, I could show that in a single focus interaction, self-interruptions of the agent—already a simple pause—can affect the participant's *VFoA*. This can be measured in terms of significantly less inattention and less attention shifts. For more complex interactions with different foci of discourse, I could not measure such an effect.

11.4 THE FINAL AHM: BACK TO THE ROOTS OF ATTENTION THEORIES

In section 2.1.1, I presented different attention theories, with a focus on the *capacity theory of attention* by Kahneman [Kah73]. Based on this model, I draw tree considerations for the development of human-agent interactions:

1. The human attention capacity is limited. Therefore, the agent needs to be careful with the allocation of the human's attention.

2. Changes in the environment can change the attention policy of the human. The agent needs to be aware of this and consider environmental changes, to understand human responses.

3. The process of attention allocation can be top-down or bottom-up. So, there are various approaches how human attention can be attracted, and the agent needs to decide which one is suitable in the current situation.

I addressed theses three considerations in my *AHM* as follows. The general idea of the *AHM* is a non-intrusive allocation of the human's attention. Therefore, I investigated the unobtrusive speech phenomenon of hesitations as an implicit attention-grabbing intervention strategy. Furthermore, with the cascade of interventions, initiated through the behavior of the human, it is possible to "pause" the interaction whenever the attention of the human interaction partner is needed somewhere else. In fact, as the intervention strategy only ends when the attention of the user is back, the human can take as much time as necessary.

At the moment, environmental changes are only implicitly considered, through the notion of attention shifts. However, it is of course possible, that e.g., the agent uses the sensors of the intelligent apartment to recognize environmental changes. One example could be that the agent starts its intervention strategy whenever another human enters the interaction zone and starts speaking.

Whereas the unfilled pauses and the hesitation vowels serve as button-up attention allocations, the repetition of the location information serve more as a top-down attention allocation. Of course, both are relatively less intrusive. For more important or urgent situations, a more intrusive and explicit attention grabbing strategy has to be considered. Besides different verbal approaches, the agent has the possibility to use the actors within of the apartment, such as light and displays.

According to my studies, the two main capabilities of *DM* (1) when to (re-)act and (2) how to (re-)act of the *AHM* should be as follows. Even though participants perform better in the *AHM* condition in *EC3*, *EC4* as well in *EC5*, I would argue to choose the attention concept of *EC3* or *EC5*, depending on the task at hand. In the post-hoc interviews, this timing strategy in *EC5* received better feedback than in *EC4*. It seems more natural to look at the current *FoD* and listen to the information while observing relevant features of the environment. In addition, other task related information should be considered if the interaction consists of a practical task during interaction, as in the *EC3*. To this end, the attention concept should be a combination of the model in *EC3* and *EC5*, depicted in fig. 11.5. It uses mutual gaze and task related features to distinguish inattentiveness based on missing engagement or difficulties in understanding.

For the intervention strategy itself, the strategies in *EC5* also perform best, depicted in fig. 11.6. The combination of lengthening, hesitation vowels, and repetitions allows the agent to react better to different situations. The lengthening introduces the unfilled pause as a communicative signal, which allows participants to distinguish between a pause and a turn end. Furthermore, the hesitation vowels perform as an attention catcher. The repetitions act as a repair mechanism for not working attention shifts, that were not followed by the participant.

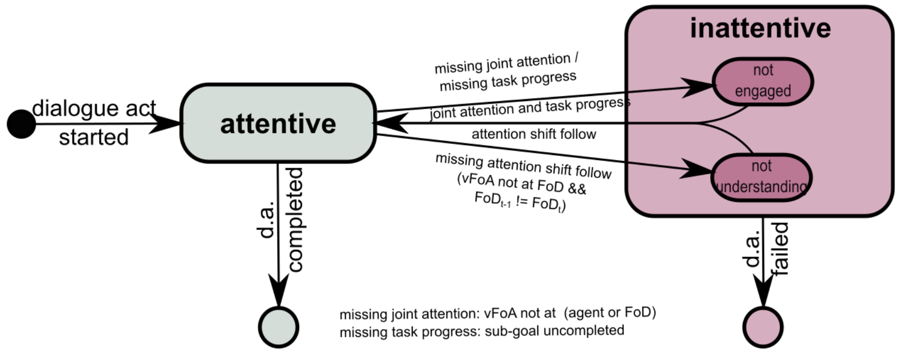


Figure 11.5: Final concept of attention: using mutual gaze and task related features to distinguish inattentiveness based on missing engagement or difficulties in understanding.

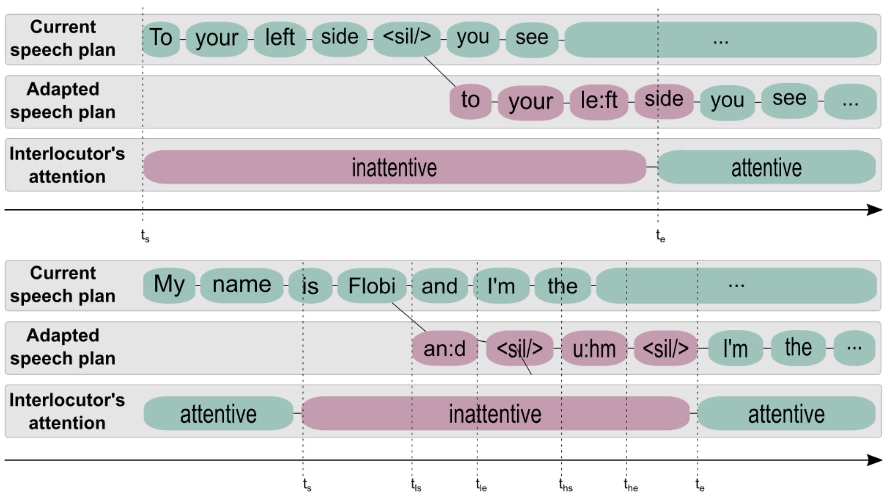


Figure 11.6: Final hesitation intervention strategies: (a) highlight (b) re-attention.

12 SUMMARY OF PART III

In this third part (*Learning from Experiments*), I evaluated and enhanced the *AHM* in five *ECs*, consisting of three pilot- and two HAI studies in a smart-home environment to investigate RQ 3. In section 9.1, I discussed benefits and disadvantages of different ways for the evaluation of dialogue systems. Based on this, I presented the chosen evaluation method of *ECs* in section 9.2, explained the need of interaction studies for the investigation of my hypothesis and gave an overview of the carried out *ECs*. During these cycles in chapter 10, I tested my hypothesis that the *AHM* increases the task performance of *HAI* in an information-giving smart home scenario. Thereby, I did not only look at the task performance, but also considered the side effects that the *AHM* can have on the interaction. Between the cycles, I iteratively improved the model and evaluation approach. Furthermore, I explored features for both the attention concept and the hesitation intervention strategy.

In section 10.1 (*EC1*), I performed the first pilot study investigating a simple attention-regain strategy. I showed that already in short interactions—without a change of the discourse—unfilled pauses based as a reaction on missing mutual gaze have a positive effect on the gazing behavior of the interlocutors.

Furthermore, I demonstrated in section 10.2 (*EC2*) that three things can be implemented and integrated into an autonomous system: (1) a multimodal attention guiding, (2) an attention monitoring system, and (3) different intervention strategies to react on an inattentive interaction partner. Furthermore, the result of the subjective ratings in the study indicated that users may struggle with the differentiation of unfilled pauses from turn-ends in more complex scenarios.

In section 10.3 (*EC3*), I explored a more practical task—the preparation phase of cooking—and investigated how such an interactive scenario can be formalized and modeled. Therefore, I improved an existing interaction scenario in the *CSRA* by applying the *AHM* on one dialogue act of the information exchange between the human and the agent. Furthermore, I showed in the evaluation of the *AHM* that the participants in the *AHM* condition perform significantly better in the task during the interaction, measured by a lower error rate.

Influenced by the results for the practical task, I changed the evaluation approach in the next cycle (*EC4*) in section 10.4. The task performance is now measured by a practical task after the interaction. In doing so, and through the integration of lengthening as a hesitation feature, I showed that participants in the *AHM* condition performance better in the task than participants in the baseline. This effect is, however, accompanied by lower subjective ratings of the agent's voice quality, although the use of lengthening counteract the lower subjective ratings of the agent itself.

In the last cycle (*EC5*) in section 10.5, I combined the differentiation of two states in the attention concepts (used in *EC2* and *EC3*) with the enhanced hesitation strategy of *EC4*. In addition, hesitation vowels are introduced as hesitation feature. I showed that participants in the *AHM* condition perform better in the task measured as post-interaction information recall than participants in the *BASE* condition, with no worse subjective ratings of the agent's voice quality or the agent itself. However, this effect is only observed in the *sweets* and not the *information* part of the interaction.

In chapter 11, I discussed the results of the five *ECs* further. In doings so, I compared results regarding the task performance hypothesis between my *ECs* in section 11.2. Furthermore, I discussed my outcome with *HHI* research results and my contribution to the different hypotheses regarding the reason why disfluent speech can have a positive effect on the listener (see section 11.2.1). In addition, I compared my results with other *HAI* experiments which showed a positive effect of hesitations on the task performance in section 11.3.1. Furthermore, in section 11.3, I discussed the results regarding the side effects on the interaction: the subjective ratings (see section 11.3.1) and the visual attention (see section 11.3.2). At the end of the chapter, I took a look back on the theories of attention and their considerations for the development of human-agent interactions and presented the resulting final *AHM* (see section 11.4).

13 CONCLUSION AND PERSPECTIVE

In this chapter, I present the insights gained throughout this thesis and elaborate on my contribution and the consequences for the dialogue coordination in smart homes. Furthermore, I discuss limitations of this work and present research questions that should be investigated in future work.

13.1 SUMMARY OF THIS THESIS

In this work, I investigated how human attention can be incorporated into dialogue management to improve *HAI* in smart homes.

In the first part of this thesis, I investigated how the coordination of human attention and system speech can be modeled. To this end, a **comprehensive literature review on the concepts of attention and hesitation** was presented, starting with research from Human-Human Interaction. Based on the *capacity theory of attention*, I identified considerations for the incorporation of human attention in *HAI* in section 2.1.1. Furthermore, a closer look at common concepts and definitions of the attention term—which are often defined only vaguely—was taken (see section 2.1). Similar terms were differentiated and an overview of the role of gaze in *HHI* was given. On this basis, a concept of attentional state was developed. The evidence from the literature shows, that the human gaze is a reliable indicator for their attention and higher cognitive processes in *HHI*. Then, disfluent speech and its use in *HHI* was discussed in section 2.2. Findings from the linguistic and psychological research on *HHI* was presented to investigate the claim that hesitations in speech can improve the listener's comprehension in *HHI* and are often produced as a reaction to listener's inattentiveness. It was shown that the human gaze is a reliable indicator for attention and higher cognitive processes, and evidence was presented that hesitations can improve the listener's comprehension in *HHI*. These findings from multiple research fields were summarized, and **the relationship between the concept of attention and hesitating speech was pointed out**.

In the third chapter, findings from research focusing on attention and hesitations in *HAI* were discussed. This showed that these concepts are also important in *HAI*. I gave an overview about the use of gaze in *HAI*, which revealed that the human gaze also plays an important role in *HAI*. Furthermore, existing systems and studies which deal with hesitations in *HAI* are presented. Thereby, a gap in the interaction research field became evident. Even though the effect of gaze in *HAI* is examined, only little research investigates how to deal with an inattentive interaction partner. Additionally, using hesitations intentionally to support the interaction—as a conversational act to deal with missing attention—has rarely been studied. Based on these insights, I developed a **model to react on inattentiveness of the human interaction partner, which uses hesitations as intervention strategies: the Attention-Hesitation Model (AHM)**. This model is the first one, which deals with missing attention during systems speech in smart-home interactions and thereby acknowledges that human attention is a resource and gives the interaction partner the necessary time.

In the second part, I investigated which requirements such a model does impose on the design of dialogue systems. To this end, from a system engineering perspective, the classical natural language processing pipeline of spoken dialogue systems and the coordination of dialogue in more detail was examined. Based on this, I illustrated the **requirements posed by the technical realization of my *AHM*** in section 6.2. In chapter 7, I described the **technical realization of a dialogue system, which allows further investigation of my research question**. The selected the research platform is presented (section 7.1), consisting of a dialogue system architecture (section 7.2) meeting the requirements, extracted from the general architecture often used in speech-based systems. Two main concepts for dialogue modeling are identified: (1) the use of interaction patterns with system task descriptions for a rapid prototyping of interaction scenarios and (2) the concept of the *IU* model to deal with the incremental nature of human dialogue. With the combination of the frameworks *Pamini* and *inprotk* both concepts are considered in my dialogue system, which serves as an appropriate basis for the investigation of my research hypothesis and is the fundamental work that allows autonomous *HAI* and the investigation of the effects of my *AHM* in interaction. In addition, multiple scenarios and research studies on diverse platforms are presented, which used the resulting dialogue system, including a first integration of human visual attention. In a multi-party human-robot interaction scenario, the benefit of the incorporation of human attention in the establishing phase of the interaction could be shown. The dialogue system benefits from the information of the interlocutor's eye-gaze by reducing false-positive reactions from the robot (section 7.5).

In the third part of this thesis, I investigated how the *AHM* affects the *HAI* in a smart home. I **evaluated and enhanced the model in five Evaluation Cycles (ECs)**, consisting of three pilot- and two *HAI* studies in the *CSRA*. Thereby, the model and the evaluation approach were iteratively improved. Different features of the attention concept and the hesitation strategies were explored to find an intervention strategy that can **improve the task performance without having negative side effects on the interaction**. In the first *EC*, I showed that already in short interactions without a change of the discourse, the participants interacting with an agent that uses my *AHM* are significantly **less inattentive** than participants in the baseline. This was measured by the overall time participants looked away from the agent and the number of attention shifts (*EC1*). However, this effect could not be measured in the following interaction scenarios with changing *FoDs*. In addition, I demonstrated that participants interacting with an agent that uses my *AHM* **perform significantly better in some practical tasks** than participants in the baseline (*EC3*, *EC4*, *EC5*). However, this effect is accompanied by lower subjective ratings of the agent or its voice quality(*EC2*, *EC3*, *EC4*). The ratings especially show that repetitions can be perceived as annoying (*EC2*) and unfilled pauses as rude (*EC2*, *EC3*). Lengthening of phonemes can enhance the hesitation strategy, but lower the voice quality ratings(*EC4*). With the final *AHM*, participants perform significantly better in the task, without negative effects on the subjective ratings (*EC5*). Furthermore, I showed that the *AHM* can work fully autonomously (*EC2*, *EC4*) and thereby improve the task performance (*EC4*).

To summarize, in this thesis, I presented a first model to deal with inattentive interaction partners with hesitations and thereby improving the *HAI* measured by task performance, which works autonomously, and is evaluated in a real *HAI* scenario in a smart home. Thereby, I made several interdisciplinary scientific contributions, which I summarize in the following.

13.2 CONTRIBUTIONS OF THIS THESIS

With this thesis, I contribute to the research field of cognitive interaction research and to dialogue modeling in the system engineering research. Therefore, I present my contributions to the different fields in the following two sections.

13.2.1 *Cognitive Interaction Research*

The contribution regarding the cognitive interaction research consist of the *AHM* and its evaluation to investigate the effect of hesitations.

Interaction Model to Handle Inattentive Interaction Partners

I presented the *AHM*, the first model that uses hesitations during a speech act as a conversational signal for inattentive interlocutors to improve the interaction in a smart home environment (chapter 4) In contrast to other models I incorporate all the following aspects in the *AHM*:

- I acknowledge that the human attention as a valuable resource and give the inattentive interlocutors the time they need. To this end, the AHM uses hesitations as a non-intrusive way to deal with missing attention.

- I differentiate between two possible reasons for inattention: (a) engagement problems or (b) difficulties in understanding. Therefore, the AHM can react with different hesitation intervention strategies on these states, either with a (a) re-attention strategy or a (b) highlight attention strategy.
- I make it possible to adapt the system behavior during the speech act. For this purpose, the model observes the person during the system speech and can react to inattentiveness during speech acts, in contrast to only adapting in between speech acts.

Furthermore, in contrast to other models, the AHM is evaluated in five ECs to measure the effect of the intervention strategy on the interaction.

Effect of Hesitations in Interaction

Through the evaluation of my model in five EC, I contribute the following insights to interaction research. The main contribution is that participants interacting with an agent that uses my AHM perform significantly better in some practical tasks than participants interacting with an agent that does not react to the interlocutor's inattention (EC3, EC4, EC5). Taking a closer look at the attention concept of my model, the following findings regarding the used features can be made:

- A hesitation intervention strategy initiated based on a simple attention concept—only relying on mutual gaze—can lead to higher task performance of the agent at the cost of less positive subjective ratings by the user (EC3).
- Using (1) mutual gaze and task related features to distinguish inattentiveness based on missing engagement or difficulties in understanding and (2) different strategies to deal with these improve the task performance without negative side effects on the interaction (EC5).

Taking a closer look at the hesitation intervention strategies of my model, the following findings regarding the used features can be made:

- Already in short interactions, without a change of the discourse, unfilled pauses based on missing mutual gaze have a positive effect on the gazing behavior of the interlocutors (EC1).
- Users may struggle with the differentiation of unfilled pauses from turn-ends in more complex scenarios (EC2, EC3).
- The use of lengthening may counteract this problem (EC4, EC5).

- Some users perceived repetitions as annoying (*EC2*, *EC5*).
- To deal with inattention based on missing engagement, a cascade of lengthening, followed by an unfilled pause, followed by a hesitation vowel, followed by another unfilled pause is used. For difficulties in understanding, the model uses repetitions with lengthening. This combination improves the task performance without negative side effects on the interaction (*EC5*).

In contrast to other models using hesitations, the *AHM* uses them, they are used as an explicit communicative act to reacquire attention. I showed that hesitations during a speech act can be used as a communicative act not only to signal the listener that the system needs more time (*to buy time for the system*), but rather to signal that the system wants the interlocutor's attention and thereby *to buy time for the listener*. Hesitating to buy time for the listener was already done before—e.g., in the in-car system [Kou+14]—but not based on the interlocutor's inattention and to signal that the system wants the interlocutor's attention back.

13.2.2 *System Engineering Research of Dialogue Modeling*

Besides these results in cognitive interaction research, I additionally contribute to the system engineering research of dialogue modeling of human-agent interactions by (1) presenting guidelines for the integration of human attention into dialogue management and (2) the technical realization of an incremental dialogue system.

Guidelines for Integrating the Attention in Dialogue Coordination

It is remaining an unsolved question how the coordination of natural *HAI* can be modeled. However, with my research, I can provide guidelines on how to incorporate the human attention into the dialogue management, based on the requirements stated from the system engineering perspective in section 6.2:

LEVELS OF DECISION-MAKING: The task of the dialogue management component is to decide *when* to (re-)act and *how* to (re-)act. In dialogue systems, this is usually done based on conversational acts: When A states a question, B answers. I argue that dialogue management needs more distinct levels of decision-making: (1. *high level)* the overall goal/plan of the agent, (2. *discourse level* the interaction management based on dialogue acts, and (3. *dialogue act level*) the adaptions or repair strategies during a dialogue act (section 7.3.2). The capability to adapt one's behavior is a crucial requirement for successful interaction. Intelligent systems are developed for the human. Therefore, they should be able to react to the continuous feedback of human interaction partners, not only once per dialogue act, but continuously. This has consequences for the coordination of human-agent dialogue.

REPRESENTATION OF HUMAN MENTAL STATES: To allow this, the system needs to be aware of humans and their feedback signals. This requires a representation of the interaction partners and their mental states. Based on observations of the human, the system needs to draw conclusions about these states—e.g., whether the human interaction partner pays attention, whether there are indicators for engagement in the interaction, or the understanding of the task. I propose to differentiate the mental states of inattentiveness based on (1) engagement errors or (2) difficulties in understanding (section 4.1). In doing so, features from the visual perception as well as the dialogue system is needed (see section 11.4). In case the system notices that the human interaction partner is inattentive, the dialogue coordination has the means to repair in different ways.

REPAIR STRATEGIES: One approach to regaining attention is to request it explicitly. For example, by stopping, directly addressing the problem, and actively requesting feedback. Another—less intrusive—way is to give the interaction partner the necessary time. I propose the dialogue system should have at least one of these less intrusive intervention strategy to acknowledge that human attention is a valuable resource (section 4.2.2). A prerequisite to this approach is that the human attention is not urgently required. Hesitating is a phenomenon which occurs regularly in *HHI* and can be used to react in such situations. Even though in *HHI* research, hesitations are often seen as a mechanism to buy time for the speaker, I show in this thesis that they can also serve for granting time to the listener. These hesitation strategies can be consisted of different features, based on the possibility of the system (see section 11.4).

EVALUATION: Humans and real-world interactions with them need to be integrated into the design process of dialogue modeling as early as possible. In my opinion, it is highly relevant that systems, which are meant for interaction, are actually tested *in* interaction. For dialogue coordination, this can be very difficult. Dialogue management has to rely on previous processing modules, like *TTS* and *Natural Language Understanding (NLU)*, so their results affect its performance. However, this makes testing in interaction even more necessary. Human communication is full of errors, inaccuracies, and misunderstandings. All processing modules need to deal with them. Although, it is good and necessary to evaluate components separately, such investigations cannot show how the components perform in combination and especially not in a real interaction with a human communication partner. Therefore, I propose the methodology of *ECs* (see section 9.2). To this end, the human is "in the loop"—the system is evaluated in interaction. Furthermore, both fully autonomous interactions and WoZ scenarios should be considered. In doing so, we have the benefit of controlled *WoZ* behavior without the drawback, that such controlled interactions may not be replicable with autonomous agents.

Dialogue Coordination between a human and an agent is still a difficult topic, and the integration of the concepts of attention and hesitations leads to several technical consequences for the dialogue coordination itself. First, the system must be able to work incrementally. The dialogue manager needs the capability to adapt its plan. Second, context information is necessary. Each interaction occurs in a particular context, which leads to different interpretations of the same information. Third, the system needs a representation of the interaction partner, and the corresponding mental states. These are fundamentals for the evaluation of autonomous *HAI*.

Incremental Dialogue System

Finally, I made a contribution to the technical realization of incremental dialogue systems by combining the two toolkits *inprotk* and *Pamini*. I thereby address the following aspects of dialogue systems:

INCREMENTALLY: Explicit consideration of the incremental nature of dialogue processing is achieved in two ways: (1) incremental processing capabilities are given in the speech recognition and speech synthesis modules of *inprotk* through the concept of *IUs* and (2) the concept of tasks allows the interruption of system actions. This is achieved through the combination of two toolkits for modeling dialogue: *inprotk* and *Pamini* (see section 7.3.2).

MODULARITY: Furthermore, by using the concept of *services* within the *CSRA* (explained in section 7.2), the dialogue system is modular (7.3). I defined interfaces to allow an exchange of single components.

MULTI-MODALITY: Multi-modality is achieved by using the services of the apartment (see section 7.2). On the input side, services such as the *Speech Recognition*, *Face Recognition*, or *Pointing Recognition* provide information which can be further processed, e.g., in the attention module (section 7.3.2). On the output side, the generation of multimodal dialogue acts can be configured, including, e.g., verbal output with facial expressions or head animations. In addition, further multimodal system actions can be triggered using the task state interface. Multi-modal attention guiding is realized through the highlight-service (see section 7.3.3).

TOPOLOGY: Based on the previous requirements, the topology of the dialogue system is organized in layers (see section 7.2). However, through the used middleware, each component can observe or request information from other components. In addition, through the concept of *IUs*, the dialogue system can assume various topologies and is a network rather a single pipeline (see section 6.1.4).

GENERALIZABILITY: The overview of various interactions—using the whole or parts of my dialogue system presented in chapter 7—shows that the dialogue system can deal with different scenarios and is platform-independent.

Furthermore, I could show the *AHM* can work fully autonomously in *HAI* (in *EC2* and *EC4*).

13.3 CONSEQUENCES FOR SMART-HOMES

My research has additional consequences for smart-home interactions. I am convinced, that if we want to change the *Smart Personal Assistants (SPAs)* in our smart-homes from sole tools to real *assistive systems*, it is mandatory to coordinate the system's talk with the human's attention. From *HHI* research, we knew that dialogue is highly incremental and humans adapt their behavior based on the feedback received from their interaction partners continuously. Interaction is a joint action, as Goodwin pointed out:

> "To engage successfully in conversation, participants are required not only to produce sentences but also to coordinate, in a meaningful fashion, their talk with the talk of others present." [Goo81]

To achieve a meaningful conversation, joint and shared attention is necessary. Not having or losing attention makes people adapt or repair an ongoing interaction. To be able to assist humans in more complex tasks—such as cooking—or to be able to interact longer with the human, e.g., for longer explanations, the *assistive agent* should ensure that the human is engaged and understanding. To detect problems with the human's attention, the *SPA* needs to use the sensors within smart-homes or needs additional sensors to detect the attention by itself. It is—of course—a very controversial subject, whether additional sensors for detecting attention at home are appropriate. On the contrary, the users may have the feeling of being better understood and having a more situated interaction with the *SPA*. The users may feel that they are being watched because of the cameras. In fact, we already have various camera around us, e.g., as surveillance in public places, on our mobile phones and laptops for communication with other people, or on game consoles to enable new game interaction possibilities. If the users see the advantages and purpose, they will be more comfortable with it. Nevertheless, data security must be ensured. The households are a sensitive area and data, e.g., from a camera, should not leave this sensitive area. The more data a system collects, the more important it is to protect this data. Especially when the system monitors the user and builds a representation of the human interaction partners and their mental states. On the other side, the smart-home itself needs to be more connected. Some smart isolated solutions do not allow a comprehensive monitoring of the user. To have the best advantage of the available sensors in the smart home, they need to be connected and share their information with the *SPA*. The other way around, the *SPA* should be able to use the actors. Already now, *SPA* can control some of these actors, e.g., switching the status of lights, when the user wants that. The *SPAs* should also be able to use such actors as a communicative act, as my attention highlight in section 7.3.3 demonstrates.

13.4 LIMITATIONS AND FUTURE RESEARCH QUESTIONS

Besides the contributions, this work also has some limitations, encourage future work and further research questions. In addition, throughout this thesis, I pointed out several further research questions, which are out of the scope of this thesis.

GENERALIZABILITY: The presented interaction studies have several limitations. Even though these studies are conducted in a smart apartment which is designed to look more like a living space than a laboratory, further research should investigate interactions in real-world scenarios. Furthermore, the study participants are mostly German university students with a WEIRD (western, educated, industrialized, rich and democratic) [HHN10] background. Other populations should be addressed as well. In addition, the presented interaction scenarios were short, and the participants did not repeatedly interact with the agents over a long period. It would be interesting to see what kind of effects these attention strategies have in long term, repeated interactions. Finally, investigating other, more cooperative interactions would be interesting to elaborate the generalizability of the presented model.

FEATURES OF THE AHM: Further potential for improvement lies in the attention module. I explored a very limited set of features. Based on my model, other modalities and feedback signals can be integrated to improve the current estimation of the human attention state. One example would be head nods and short verbal back-channels. In addition, the distinction between the two states of inattentiveness was very basic in my *ECs*. Comprehensive classifiers should be trained here, but that was outside the scope of this work. I am currently researching this more closely. Besides this, the intervention strategy itself provides a wide range of different behaviors. In chapter 4, I described hesitation features, their combinations and the possibilities for varying these behaviors. Even though the choice of the used hesitation strategies is well justified, this research area has more potential for further research questions and investigations. In particular, the use of fillers and their effect on the listener should be investigated more closely. Furthermore, the use of additional smart-home sensors and actuators for both the estimation of inattentiveness of the human interaction partner and the possibilities of expressions for the agent need to be further investigated. This offers, in my opinion, a great potential for new forms of smart-home interactions.

COMPARISON TO OTHER INTERVENTION STRATEGIES: Since we now have an effective hesitation intervention strategy, a comparison to more intrusive intervention strategies would be the next consequential step and is out of the scope of this thesis. To this end, an evaluation of which kind of strategies performs best in which situation is necessary. It can be investigated if a less intrusive system—such the *AHM*—performs better than a more intrusive system in every task. In their everyday life, people seldom devote their attention to just one thing. Divided attention is often the case because we have to constantly react to our environment and to other people in it. Depending on the needs of human attention in the current situation, a more or less intrusive intervention strategy is preferable to deal with the lack of attention and thereby recognize that human attention is a resource.

INFLUENCE OF THE TASK: In my research, I measure a better task performance in the *AHM* condition only for practical tasks measurement. It should be elaborated in more detail, what kind of influence the task itself has on the effect of the *AHM*. More specifically, it is unclear whether the task itself, its measurement, or awareness of the task itself for the participants influenced their performance.

INFLUENCE OF PERSONAL FACTORS: The data in *EC3*-*EC5* suggests that especially participants with low performance in the memory pretest benefit from the *AHM*. This raises the question, which characteristics of participants influence the effect of the *AHM*. It needs to be further investigated, if the memory performance of participants affects the improvement of the task performance or not. Due to the ceiling effect, a statement based on the current data is not possible. In addition, other personal factors should be considered. It is possible that different strategies are more suitable for different people.

NEW INTERACTION SCENARIOS: My *AHM* lays the foundation for further interaction scenarios. The possibility to react on inattentive interaction partners can be used in several ways. In addition to improving information systems in smart homes, it is conceivable to use them in other scenarios. Currently, I am investigating how such an agent can be used for the training of children with *Attention Deficit Hyperactivity Disorder (ADHD)*. People with *ADHD* may find it more difficult than others to focus on and complete tasks. Systems like the *AHM* may help to train being attentive.

With my research, I laid the foundation for further investigations in the research area of attentive *HAI* design, which uses hesitations as a conversational act for the attention coordination of dialogue and thereby sets new standards for smart-home interaction.

BIBLIOGRAPHY

OWN PUBLICATIONS

[Ber+16] Jasmin Bernotat et al. "Welcome to the future–How naive users intuitively address an intelligent robotics apartment". In: *International conference on social robotics*. Springer. 2016, pp. 982–992. DOI: https://doi.org/10.1007/978-3-319-47437-3_96 (cit. on pp. 7, 109).

[Bet+18] Simon Betz, Birte Carlmeyer, Petra Wagner, and Britta Wrede. "Interactive hesitation synthesis: modelling and evaluation". In: *Multimodal Technologies and Interaction* 2.1 (2018), p. 9. DOI: https://doi.org/10.3390/mti2010009 (cit. on pp. 176, 179, 180, 182, 216, 218).

[Car+18] Birte Carlmeyer, Simon Betz, Petra Wagner, Britta Wrede, and David Schlangen. "The Hesitating Robot-Implementation and First Impressions". In: *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction-HRI'18*. 2018. DOI: https://doi.org/10.1145/3173386.3176992 (cit. on p. 176).

[CCW17a] Birte Carlmeyer, Monika Chromik, and Britta Wrede. "Interaction Model for Incremental Information Presentation". In: *Proceedings of the 5th International Conference on Human Agent Interaction*. HAI '17. Bielefeld, Germany: ACM, 2017, pp. 335–339. ISBN: 978-1-4503-5113-3. DOI: 10.1145/3125739.3132582 (cit. on p. 153).

[CSW16a] Birte Carlmeyer, David Schlangen, and Britta Wrede. ""Look at Me!": Self-Interruptions as Attention Booster?" In: *Proceedings of the Fourth International Conference on Human Agent Interaction - HAI '16*. Singapore: Association for Computing Machinery (ACM), 2016. DOI: 10.1145/2974804.2980488 (cit. on p. 131).

[CSW16b] Birte Carlmeyer, David Schlangen, and Britta Wrede. "Exploring self-interruptions as a strategy for regaining the attention of distracted users". In: *Proceedings of the 1st Workshop on Embodied Interaction with Smart Environments - EISE '16*. Association for Computing Machinery (ACM), 2016. DOI: 10.1145/3008028.3008029 (cit. on p. 141).

[CCW17b] Monika Chromik, Birte Carlmeyer, and Britta Wrede. "Ready for the Next Step?: Investigating the Effect of Incremental Information Presentation in an Object Fetching Task". In: *Proceedings of the Companion of the 2017*

ACM/IEEE International Conference on Human-Robot Interaction - HRI '17. Association for Computing Machinery (ACM), 2017. DOI: 10.1145/3029798.3038352 (cit. on p. 165).

[Hol+16] Patrick Holthaus et al. "How to Address Smart Homes with a Social Robot? A Multi-modal Corpus of User Interactions with an Intelligent Environment". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016. ISBN: 978-2-9517408-9-1 (cit. on pp. 89, 109, 125).

[Hux+19] Tamino Huxohl, Marian Pohling, Birte Carlmeyer, Britta Wrede, and Thomas Hermann. "Interaction Guidelines for Personal Voice Assistants in Smart Homes". In: *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE. 2019, pp. 1–10. DOI: https://doi.org/10.1109/SPED.2019.8906642 (cit. on p. 9).

[Neu+17] Alexander Neumann et al. ""KogniChef": A Cognitive Cooking Assistant". In: *KI - Künstliche Intelligenz* 31.3 (Aug. 2017), pp. 273–281. ISSN: 1610-1987. DOI: 10.1007/s13218-017-0488-6 (cit. on p. 111).

[Ric+16] Viktor Richter, Birte Carlmeyer, Florian Lier, Sebastian Meyer zu Borgsen, Franz Kummert, Sven Wachsmuth, and Britta Wrede. "Are you talking to me? Improving the robustness of dialogue systems in a multi party HRI scenario by incorporating gaze direction and lip movement of attendees". In: *Proceedings of the Fourth International Conference on Human-agent Interaction*. Singapore: ACM Digital Library, 2016. DOI: 10.1145/2974804.2974823 (cit. on pp. 109, 110).

GENERAL REFERENCES

[ABE07] Jordi Adell, Antonio Bonafonte, and David Escudero. "Filled pauses in speech synthesis: towards conversational speech". In: *International Conference on Text, Speech and Dialogue*. Springer. 2007, pp. 358–365. DOI: 10.1007/978-3-540-74628-7_47 (cit. on p. 48).

[AEB12] Jordi Adell, David Escudero, and Antonio Bonafonte. "Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence". In: *Speech Communication* 54.3 (2012), pp. 459–476. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2011.10.010 (cit. on p. 48).

[AS17] Henny Admoni and Brian Scassellati. "Social eye gaze in human-robot interaction: a review". In: *Journal of Human-Robot Interaction* 6.1 (2017), pp. 25–63. DOI: https://doi.org/10.5898/JHRI.6.1.Admoni (cit. on pp. 9, 29, 30, 39–41).

[All11] Joseph J. Allaire. "RStudio: Integrated Development Environment for R". In: *The R User Conference, useR! 2011 August 16-18 2011 University of Warwick, Coventry, UK*. 2011, p. 14 (cit. on p. 127).

[ABR13] Gouzhen An, David-Guy Brizan, and Andrew Rosenberg. "Detecting laughter and filled pauses using syllable-based features." In: *INTERSPEECH*. 2013, pp. 178–181 (cit. on p. 48).

[And+13] Elisabeth André, Jean-Claude Martin, Florian Lingenfelser, and Johannes Wagner. "Multimodal fusion in human-agent dialogue". In: *Coverbal Synchrony in Human-Machine Interaction* (2013), pp. 387–410. DOI: 10.1201/b15477-16 (cit. on p. 74).

[And+16] Bejarano Andrés, Fernández Alejandra, Jimeno Miguel, Salazar Augusto, and Wightman Pedro. "Towards the evolution of smart home environments: A survey". In: *International Journal of Automation and Smart Technology* 6.3 (2016), pp. 105–136. DOI: 10.5875/ausmt.v6i3.1039 (cit. on p. 6).

[And+12a] Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher. "A Head-eye Coordination Model for Animating Gaze Shifts of Virtual Characters". In: *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*. Gaze-In '12. Santa Monica, California: ACM, 2012, 4:1–4:6. ISBN: 978-1-4503-1516-6. DOI: 10.1145/2401836.2401840 (cit. on p. 43).

[And+12b] Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher. "Designing Effective Gaze Mechanisms for Virtual Agents". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: ACM, 2012, pp. 705–714. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2207777 (cit. on p. 43).

[And+14] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. "Conversational gaze aversion for humanlike robots". In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM. 2014, pp. 25–32. DOI: 10.1145/2559636.2559666 (cit. on pp. 29, 30).

[ACC94] Michael Argyle, Mark Cook, and Duncan Cramer. "Gaze and mutual gaze". In: *The British Journal of Psychiatry* 165.6 (1994), pp. 848–850. DOI: doi : 10 . 1017 / S0007125000073980 (cit. on pp. 9, 29).

[ABS13] Suket Arora, Kamaljeet Batra, and Sarabjit Singh. "Dialogue System: A Brief Review". In: *CoRR* abs/1306.4134 (2013). arXiv: 1306.4134. URL: http://arxiv.org/abs/1306.4134 (cit. on p. 74).

[Bal95] Dare A Baldwin. "Understanding the link between joint attention and language". In: *Joint attention: Its origins and role in development* (1995), pp. 131–158 (cit. on p. 23).

[BS10] Dale J. Barr and Mandana Seyfeddinipur. "The role of fillers in listener attributions for speaker disfluency". In: *Language and Cognitive Processes* 25.4 (2010), pp. 441–455. DOI: 10.1080/01690960903047122 (cit. on pp. 34, 35).

[Bar+09] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots". In: *International journal of social robotics* 1.1 (2009), pp. 71–81 (cit. on pp. 118, 126, 287).

[BWB09] Andrea Bauer, Dirk Wollherr, and Martin Buss. "Information retrieval system for human-robot communication-asking for directions". In: *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE. 2009, pp. 4150–4155 (cit. on p. 78).

[Bau+17] Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. "Recognising conversational speech: What an incremental asr should do for a dialogue system and how to get there". In: *Dialogues with Social Robots*. Springer, 2017, pp. 421–432 (cit. on p. 96).

[BCJ02] Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. "Listener responses as a collaborative process: The role of gaze". In: *Journal of Communication* 52.3 (2002), pp. 566–580 (cit. on p. 30).

[Bax+16] Paul Baxter, James Kennedy, Emmanuel Senft, Severin Lemaignan, and Tony Belpaeme. "From characterising three years of HRI to methodology and reporting recommendations". In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2016, pp. 391–398 (cit. on pp. 121–123).

[Bea81] Geoffrey W Beattie. "A further investigation of the cognitive interference hypothesis of gaze patterns during conversation". In: *British Journal of Social Psychology* 20.4 (1981), pp. 243–248 (cit. on p. 29).

[BAT09] Nikolaus Bee, Elisabeth André, and Susanne Tober. "Breaking the ice in human-agent communication: Eye-gaze based initiation of contact with an embodied conversational agent". In: *International Workshop on Intelligent Virtual Agents*. Springer. 2009, pp. 229–242 (cit. on p. 43).

[Bel+18] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. "Social robots for education: A review". In: *Science Robotics* 3.21 (2018), eaat5954 (cit. on p. 5).

[BM10] Cindy L. Bethel and Robin R. Murphy. "Review of Human Studies Methods in HRI and Recommendations". In: *International Journal of Social Robotics* 2.4 (Dec. 2010), pp. 347–359. ISSN: 1875-4805. DOI: 10.1007/s12369-010-0064-9 (cit. on p. 121).

[BWS15] Simon Betz, Petra Wagner, and David Schlangen. "Micro-structure of disfluencies: Basics for conversational speech synthesis". In: *Interspeech 2015* (2015) (cit. on p. 48).

[BWV16] Simon Betz, Petra Wagner, and Jana Vosse. "Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data". In: *Phonetik und Phonologie 12*. München, 2016 (cit. on pp. 51, 176, 178).

[BCL17] Miriam Bilac, Marine Chamoux, and Angelica Lim. "Gaze and filled pause detection for smooth human-robot conversations". In: *Humanoid Robotics (Humanoids), 2017 IEEE-RAS 17th International Conference on*. IEEE. 2017, pp. 297–304 (cit. on pp. 48, 49).

[BH14] Dan Bohus and Eric Horvitz. "Managing Human-Robot Engagement with Forecasts and... um... Hesitations". In: *Proc. Int. Conf. Multi-modal Interfaces - ICMI*. 2014 (cit. on pp. 47, 51, 53, 58, 63–65).

[BH09] Dan Bohus and Eric Horvitz. "Models for Multiparty Engagement in Open-World Dialog". In: *Proc. SIGdial'09*. 2009 (cit. on pp. 47, 65).

[BR09] Dan Bohus and Alexander I. Rudnicky. "The RavenClaw dialog management framework: Architecture and systems". In: *Computer Speech and Language* 23.3 (2009), pp. 332–361. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2008.10.001 (cit. on p. 74).

[Bon36] Carlo Bonferroni. “Teoria statistica delle classi e calcolo delle probabilita”. In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8 (1936), pp. 3–62 (cit. on p. 127).

[Bor+01] Heather Bortfeld, Silvia D Leon, Jonathan E Bloom, Michael F Schober, and Susan E Brennan. “Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender”. In: *Language and speech* 44.2 (2001), pp. 123–147 (cit. on pp. 31, 33).

[Bou+12] Jean-David Boucher, Ugo Pattacini, Amelie Lelong, Gerard Bailly, Frederic Elisei, Sascha Fagel, Peter F Dominey, and Jocelyne Ventre-Dominey. “I reach faster when I see you look: gaze effects in human–human and human–robot face-to-face cooperation”. In: *Frontiers in neurorobotics* 6 (2012), p. 3 (cit. on p. 30).

[BS99] Cynthia Breazeal and Brian Scassellati. “A context-dependent attention system for a social robot”. In: *Proceedings of the 16th international joint conference on Artificial intelligence-Volume 2*. 1999, pp. 1146–1151 (cit. on p. 41).

[Bre92] Andrew P Breen. “Speech synthesis models: a review”. In: *Electronics & communication engineering journal* 4.1 (1992), pp. 19–31. DOI: 10.1049/ecej:19920006 (cit. on p. 74).

[BS01] Susan E Brennan and Michael F Schober. “How listeners compensate for disfluencies in spontaneous speech”. In: *Journal of Memory and Language* 44.2 (2001), pp. 274–296 (cit. on p. 34).

[Bro58] Donald Eric Broadbent. “Perception and communication.” In: *Pergamon Press* (1958) (cit. on pp. 20, 21).

[BSG07] Robert Brueckmann, Andrea Scheidig, and Horst-Michael Gross. “Adaptive Noise Reduction and Voice Activity Detection for improved Verbal Human-Robot Interaction using Binaural Data”. In: *Proceedings 2007 IEEE International Conference on Robotics and Automation*. Apr. 2007, pp. 1782–1787. DOI: 10.1109/ROBOT.2007.363580 (cit. on p. 76).

[Bug+04] Guido Bugmann, Ewan Klein, Stanislao Lauria, and Theocharis Kyriacou. “Corpus-based robotics: A route instruction example”. In: *Proceedings of Intelligent Autonomous Systems*. Citeseer. 2004, pp. 96–103 (cit. on p. 120).

[Bun90] Claus Bundesen. “A theory of visual attention.” In: *Psychological review* 97.4 (1990), p. 523 (cit. on p. 22).

[Bus18] Hendrik Buschmeier. "Attentive Speaking. From Listener Feedback to Interactive Adaptation". In: (2018). DOI: https://doi.org/10.4119/unibi/2918295 (cit. on pp. 26, 44).

[CW14] Erik Cambria and Bebo White. "Jumping NLP curves: A review of natural language processing research". In: *IEEE Computational intelligence magazine* 9.2 (2014), pp. 48–57 (cit. on p. 73).

[Can+10] Rehj Cantrell, Matthias Scheutz, Paul Schermerhorn, and Xuan Wu. "Robust spoken instruction understanding for HRI". In: *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press. 2010, pp. 275–282 (cit. on p. 48).

[Che53] Edward Colin Cherry. "Some experiments on the recognition of speech, with one and with two ears". In: *The Journal of the acoustical society of America* 25.5 (1953), pp. 975–979. DOI: https://doi.org/10.1121/1.1907229 (cit. on p. 20).

[Cho76] Noam Chomsky. "l965. Aspects of the Theory of Syntax". In: *Cambridge, Massachussetts: MIT Press. ChomskyAspects of the theory of syntax1965* (1976) (cit. on p. 33).

[Chr16] Monika Chromik. "Entwicklung eines Modells zur inkrementellen Informationsdarstellung in einer Mensch-Agent-Interaktion". MA thesis. Germany: Faculty of Technology, Bielefeld University, 2016 (cit. on p. 155).

[CK04] Herbert H Clark and Meredyth A Krych. "Speaking while monitoring addressees for understanding". In: *Journal of memory and language* 50.1 (2004), pp. 62–81 (cit. on p. 26).

[CS87] Herbert H Clark and Edward F Schaefer. "Collaborating on contributions to conversations". In: *Language and cognitive processes* 2.1 (1987), pp. 19–41 (cit. on pp. 25–27).

[Cla96] Herbert H. Clark. *Using language*. Vol. 1996. Cambridge University Press Cambridge, 1996. DOI: https://doi.org/10.1017/CBO9780511620539 (cit. on pp. 5, 9, 25, 30).

[CB+91] Herbert H. Clark, Susan E. Brennan, et al. "Grounding in communication". In: *Perspectives on socially shared cognition* 13.1991 (1991), pp. 127–149. DOI: https://doi.org/10.1037/10096-006 (cit. on pp. 6, 25).

[CF02] Herbert H. Clark and Jean E. Fox Tree. "Using uh and um in spontaneous speaking". In: *Cognition* 84.1 (2002), pp. 73–111. ISSN: 0010-0277. DOI: https://doi.org/10.1016/S0010-0277(02)00017-3 (cit. on pp. 12, 33).

[Coh77] Jacob Cohen. "Statistical power analysis for the behavioral sciences, Rev". In: (1977) (cit. on p. 127).

[Col09] Philip Collard. “Disfluency and listeners’ attention: An investigation of the immediate and lasting effects of hesitations in speech”. PhD thesis. University of Edinburgh, 2009 (cit. on p. 34).

[Coo77] Mark Cook. “Gaze and Mutual Gaze in Social Encounters: How long - and when - we look others in the eye is one of the main signals in nonverbal communication”. In: *American Scientist* 65.3 (1977), pp. 328–333 (cit. on p. 29).

[CS08] Martin Corley and Oliver W Stewart. “Hesitation disfluencies in spontaneous speech: The meaning of um”. In: *Language and Linguistics Compass* 2.4 (2008), pp. 589–602 (cit. on pp. 32–34).

[CKB18] Aykut Coskun, Gül Kaner, and İdil Bostan. “Is smart home a necessity or a fantasy for the mainstream user? A study on users’ expectations of smart household appliances”. In: *International Journal of Design* 12.1 (2018), pp. 7–20 (cit. on p. 7).

[Cum08] Geoff Cumming. “Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better”. In: *Perspectives on Psychological Science* 3.4 (2008), pp. 286–300 (cit. on p. 122).

[DTW16] Rasmus Dall, Marcus Tomalin, and Mirjam Wester. “Synthesising Filled Pauses: Representation and Datamixing”. In: *Proc. SSW9, Cupertino, CA, USA* (2016) (cit. on p. 48).

[Dan+16] Timo Dankert, Michael Goerlich, Sebastian Wrede, Raphaela Gehle, and Karola Pitsch. “Engagement Detection During Deictic References in Human-Robot Interaction”. In: *International conference on social robotics*. Springer. 2016, pp. 930–939 (cit. on pp. 45, 63, 65).

[Dau07] Kerstin Dautenhahn. “Socially intelligent robots: dimensions of human–robot interaction”. In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 362.1480 (2007), pp. 679–704 (cit. on pp. 6, 41).

[De +05] Boris De Ruyter, Privender Saini, Panos Markopoulos, and Albert Van Breemen. “Assessing the effects of building social intelligence in a robotic interface for the home”. In: *Interacting with computers* 17.5 (2005), pp. 522–541 (cit. on p. 6).

[DD95] Robert Desimone and John Duncan. “Neural mechanisms of selective visual attention”. In: *Annual review of neuroscience* 18.1 (1995), pp. 193–222 (cit. on p. 23).

[DD63] J Anthony Deutsch and Diana Deutsch. “Attention: Some theoretical considerations.” In: *Psychological review* 70.1 (1963), p. 80 (cit. on pp. 20, 21).

[DP05] Gwyneth Doherty-Sneddon and Fiona G. Phelps. "Gaze aversion: A response to cognitive or social difficulty?" In: *Memory & Cognition* 33.4 (June 2005), pp. 727–733. ISSN: 1532-5946. DOI: 10.3758/BF03195338 (cit. on p. 29).

[DLH14] Christian Dondrup, Christina Lichtenthäler, and Marc Hanheide. "Hesitation signals in human-robot head-on encounters: a pilot study". In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM. 2014, pp. 154–155 (cit. on p. 53).

[DR07] Finale Doshi and Nicholas Roy. "Efficient model learning for dialog management". In: *Human-Robot Interaction (HRI), 2007 2nd ACM/IEEE International Conference on*. Mar. 2007, pp. 65–72. DOI: 10.1145/1228716.1228726 (cit. on p. 79).

[EPK17] Eustace Ebhotemhen, Volha Petukhova, and Dietrich Klakow. "Incremental Dialogue Act Recognition: token- vs chunk-based classification". In: *inform* 26 (2017), pp. 2–2 (cit. on p. 74).

[Ech+11] Gilberto Echeverria, Nicolas Lassabe, Arnaud Degroote, and Séverin Lemaignan. "Modular OpenRobots Simulation Engine: MORSE". In: *Proceedings of the IEEE ICRA*. 2011 (cit. on p. 91).

[ELP97a] Wieland Eckert, Esther Levin, and Roberto Pieraccini. "User modeling for spoken dialogue system evaluation". In: *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. Dec. 1997, pp. 80–87. DOI: 10.1109/ASRU.1997.658991 (cit. on pp. 72, 73).

[ELP97b] Wieland Eckert, Esther Levin, and Roberto Pieraccini. "User modeling for spoken dialogue system evaluation". In: *Proc. of ASRU*. Vol. 97. 1997 (cit. on p. 120).

[Eic+07] Tobias Eichner, Helmut Prendinger, Elisabeth André, and Mitsuru Ishizuka. "Attentive presentation agents". In: *International Workshop on Intelligent Virtual Agents*. Springer. 2007, pp. 283–295 (cit. on pp. 43, 62, 64).

[Ekl04] Robert Eklund. "Disfluency in Swedish human–human and human–machine travel booking dialogues". PhD thesis. Linköping University Electronic Press, 2004 (cit. on pp. 31, 32).

[Ekm04] Paul Ekman. "Emotional and Conversational Nonverbal Signals". In: *Language, Knowledge, and Representation*. Dordrecht: Springer Netherlands, 2004, pp. 39–50 (cit. on p. 9).

[Fer00] Fernanda Ferreira. “Syntax in language production: An approach using tree-adjoining grammars”. In: *Aspects of language production* (2000), pp. 291–330 (cit. on p. 33).

[FD14] Joao Filipe Ferreira and Jorge Dias. “Attentional mechanisms for socially interactive robots–a survey”. In: *IEEE Transactions on Autonomous Mental Development* 6.2 (2014), pp. 110–125 (cit. on p. 48).

[FMF12] Andy P Field, Jeremy Miles, and Zoë Field. *Discovering statistics using R/Andy Field, Jeremy Miles, Zoë Field.* 2012 (cit. on p. 126).

[FC12] Ian R Finlayson and Martin Corley. “Disfluency in dialogue: An intentional signal from the speaker?” In: *Psychonomic bulletin & review* 19.5 (2012), pp. 921–928 (cit. on p. 33).

[FND03] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. “A survey of socially interactive robots”. In: *Robotics and autonomous systems* 42.3 (2003), pp. 143–166 (cit. on p. 6).

[Fox01] Jean E. Fox Tree. “Listeners’ uses of um and uh in speech comprehension”. In: *Memory & cognition* 29.2 (2001), pp. 320–326 (cit. on pp. 12, 34).

[Fox95] Jean E. Fox Tree. “The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech”. In: *Journal of Memory and Language* 34.6 (1995), pp. 709–738. ISSN: 0749-596X. DOI: https://doi.org/10.1006/jmla.1995.1032 (cit. on pp. 12, 31).

[FW11] Scott H. Fraundorf and Duane G. Watson. “The disfluent discourse: Effects of filled pauses on recall”. In: *Journal of Memory and Language* 65.2 (2011), pp. 161–175. ISSN: 0749-596X. DOI: https://doi.org/10.1016/j.jml.2011.03.004 (cit. on pp. 34, 35, 212, 213).

[FF05] Chris Frith and Uta Frith. “Theory of mind”. In: *Current Biology* 15.17 (2005), R644–R645. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2005.08.041 (cit. on p. 5).

[Gal+17] M. Galle, E. Kynev, N. Monet, and C. Legras. “Context-aware selection of multi-modal conversational fillers in human-robot dialogues”. In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Aug. 2017, pp. 317–322. DOI: 10.1109/ROMAN.2017.8172320 (cit. on pp. 52, 53, 64).

[Gar+03] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M Angela Sasse. "The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2003, pp. 529–536 (cit. on p. 43).

[GHL06] Kallirroi Georgila, James Henderson, and Oliver Lemon. "User simulation for spoken dialogue systems: Learning and evaluation". In: *Ninth International Conference on Spoken Language Processing*. 2006 (cit. on p. 119).

[GZF18] Omid Ghahabi, Wei Zhou, and Volker Fischer. "A robust voice activity detection for real-time automatic speech recognition". In: *Proc of ESSV* (2018) (cit. on p. 76).

[GP15] Nadine Glas and Catherine Pelachaud. "Definitions of engagement in human-agent interaction". In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. Sept. 2015, pp. 944–949. DOI: 10.1109/ACII.2015.7344688 (cit. on p. 24).

[Goe+18] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. "Explainable ai: the new 42?" In: *International cross-domain conference for machine learning and knowledge extraction*. Springer. 2018, pp. 295–303 (cit. on p. 79).

[Goo81] Charles Goodwin. "Conversational organization". In: *Interaction between speakers and hearers* (1981) (cit. on pp. 5, 8, 9, 12, 29, 35–37, 63, 132, 176, 190, 234).

[GB00] Zenzi M. Griffin and Kathryn Bock. "What the Eyes Say About Speaking". In: *Psychological Science* 11.4 (2000), pp. 274–279. DOI: 10.1111/1467-9280.00255 (cit. on p. 30).

[GP05] Hatice Gunes and Massimo Piccardi. "Affect recognition from face and body: early fusion vs. late fusion". In: *2005 IEEE International Conference on Systems, Man and Cybernetics*. Vol. 4. Oct. 2005, 3437–3443 Vol. 4. DOI: 10.1109/ICSMC.2005.1571679 (cit. on p. 74).

[HB07] Joy E Hanna and Susan E Brennan. "Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation". In: *Journal of Memory and Language* 57.4 (2007), pp. 596–615 (cit. on p. 30).

[Har+14] Justin W Hart, Brian Gleeson, Matthew Pan, AJung Moon, Karon MacLean, and Elizabeth Croft. "Gesture, gaze, touch, and hesitation: Timing cues for collaborative work". In: *HRI Workshop on Timing in Human-Robot Interaction, Bielefeld, Germany*. 2014, p. 21 (cit. on p. 53).

[HN09] Robert J Hartsuiker and Lies Notebaert. "Lexical access problems lead to disfluencies in speech". In: *Experimental psychology* (2009) (cit. on pp. 32, 33).

[Hee+08] Marcel Heerink, Ben Kröse, Vanessa Evers, BJ Wielinga, et al. "The influence of social presence on acceptance of a companion robot by older people". In: *Journal of Physical Agents* 2.2 (2008), pp. 33–40 (cit. on p. 6).

[HEW10] Frank Hegel, Friederike Eyssel, and Britta Wrede. "The social robot Flobi: Key concepts of industrial design." In: *RO-MAN*. 2010, pp. 107–112 (cit. on p. 90).

[HHN10] Joseph Henrich, Steven J Heine, and Ara Norenzayan. "Most people are not WEIRD". In: *Nature* 466.7302 (2010), p. 29 (cit. on p. 235).

[Hey05] Dirk K.J. Heylen. "A closer look at gaze". In: *Creating Bonds with Embodied Conversational Agents*. University of Utrecht, 2005, pp. 3–9 (cit. on p. 30).

[HPW11] Patrick Holthaus, Karola Pitsch, and Sven Wachsmuth. "How Can I Help?" In: *International Journal of Social Robotics* 3.4 (Nov. 2011), pp. 383–393. DOI: 10.1007/s12369-011-0108-9 (cit. on p. 7).

[HSC05] Jin-Hyuk Hong, Youn-Suk Song, and Sung-Bae Cho. "A Hierarchical Bayesian Network for Mixed-Initiative Human-Robot Interaction". In: *Robotics and Automation, 2005. ICRA 2005. Proc. of the 2005 IEEE International Conference on*. Apr. 2005, pp. 3808–3813. DOI: 10.1109/ROBOT.2005.1570701 (cit. on p. 79).

[Hun+09] Victor Hung, Miguel Elvir, Avelino Gonzalez, and Ronald DeMara. "Towards a method for evaluating naturalness in conversational dialog systems". In: *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*. IEEE. 2009, pp. 1236–1241 (cit. on pp. 117, 118).

[IK01] Laurent Itti and Christof Koch. "Computational modelling of visual attention". In: *Nature reviews neuroscience* 2.3 (2001), pp. 194–203 (cit. on p. 23).

[Jam90] William James. *The principles of psychology*. Henry Holt, 1890 (cit. on p. 19).

[JO13] Dinesh Babu Jayagopi and Jean-Marc Odobez. "Given that, should I respond? contextual addressee estimation in multi-party human-robot interactions". In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2013, pp. 147–148 (cit. on p. 45).

[JLE15] Michiel Joosse, Manja Lohse, and Vanessa Evers. "Crowdsourcing Culture in HRI: Lessons Learned From Quantitative and Qualitative Data Collections". In: *3rd International Workshop on Culture Aware Robotics at ICSR*. Vol. 15. 2015 (cit. on p. 120).

[Jur+11] Filip Jurcicek, Simon Keizer, Milica Gasic, Francois Mairesse, Blaise Thomson, Kai Yu, and Steve Young. "Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk". In: *Twelfth Annual Conference of the International Speech Communication Association*. 2011 (cit. on p. 120).

[Kah73] Daniel Kahneman. *Attention and effort*. Vol. 1063. Citeseer, 1973 (cit. on pp. 20, 21, 222).

[Kaj05] Annabel Bhamani Kajornboon. "Using interviews as research instruments". In: *E-journal for Research Teachers* 2.1 (2005), pp. 1–9 (cit. on p. 196).

[KC16] S. Karpagavalli and Evania Haris Chandra. "A review on automatic speech recognition architecture and approaches". In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 9.4 (2016), pp. 393–404. DOI: 10.14257/IJSIP.2016.9.4.34 (cit. on p. 72).

[KTH10] Mayank Kaushik, Matthew Trinkle, and Ahmad Hashemi-Sakhtsari. "Automatic detection and removal of disfluencies from spontaneous speech". In: *Proceedings of the 13-th Australasian International Conference on Speech Science and Technology (SST). Melbourne, Australia*. 2010, pp. 98–101 (cit. on p. 48).

[Ken67] Adam Kendon. "Some functions of gaze-direction in social interaction". In: *Acta psychologica* 26 (1967), pp. 22–63 (cit. on pp. 9, 29).

[KKS13] Casey Kennington, Spyros Kousidis, and David Schlangen. "Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information". In: *Proceedings of the SIGDIAL 2013 Conference*. Metz, France: Association for Computational Linguistics, Aug. 2013, pp. 173–182 (cit. on p. 73).

[Kip15] Andreas Kipp. "Integration and evaluation of a gaming situation for long-term human-robot interaction: playing a game of pairs with Flobi using contextual knowledge". PhD thesis. Bielefeld University, 2015 (cit. on p. 77).

[Klo+11] 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 Conference*. Association for Computational Linguistics. 2011, pp. 341–343 (cit. on pp. 45, 47).

[Kno+18] Robin Knote, Andreas Janson, Laura Eigenbrod, and Matthias Söllner. "The What and How of Smart Personal Assistants: Principles and Application Domains for IS Research". In: *Multikonferenz Wirtschaftsinformatik (MKWI)*. 2018 (cit. on p. 6).

[Kno+19] Robin Knote, Andreas Janson, Matthias Söllner, and Jan Marco Leimeister. "Classifying Smart Personal Assistants: An Empirical Cluster Analysis". In: *Proceedings of the 52nd Hawaii International Conference on System Sciences*. 2019 (cit. on p. 6).

[KU87] Christof Koch and Shimon Ullman. "Shifts in selective visual attention: towards the underlying neural circuitry". In: *Matters of intelligence*. Springer, 1987, pp. 115–141 (cit. on p. 23).

[KG17] Ioannis Kostavelis and Antonios Gasteratos. "Robots in crisis management: A survey". In: *International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries*. Springer. 2017, pp. 43–56 (cit. on p. 5).

[Kou+14] Spyros Kousidis, Casey Kennington, Timo Baumann, Hendrik Buschmeier, Stefan Kopp, and David Schlangen. "Situationally aware in-car information presentation using incremental speech generation: Safer, and more effective". In: *Proceedings of the EACL 2014 Workshop on Dialogue in Motion*. 2014, pp. 68–72 (cit. on pp. 46, 49, 50, 53, 64, 213, 214, 219, 231).

[KT13] Teun F Krikke and Khiet P Truong. "Detection of nonverbal vocalizations using Gaussian mixture models: looking for fillers and laughter in conversational speech." In: *Interspeech*. 2013, pp. 163–167 (cit. on p. 48).

[Kul04] Vladimir Kulyukin. "Human-Robot Interaction Through Gesture-Free Spoken Dialogue". In: *Autonomous Robots* 16.3 (May 2004), pp. 239–257. ISSN: 1573-7527. DOI: 10.1023/B:AURO.0000025789.33843.6d (cit. on p. 74).

[KHD18] Minae Kwon, Sandy H Huang, and Anca D Dragan. "Expressing Robot Incapability". In: *Proceedings of the 2018 ACM IEEE International Conference on Human-Robot Interaction*. ACM. 2018, pp. 87–95 (cit. on p. 53).

[LH01] Michael F Land and Mary Hayhoe. "In what ways do eye movements contribute to everyday activities?" In: *Vision research* 41.25-26 (2001), pp. 3559–3565 (cit. on p. 30).

[Lan+03] Sebastian Lang, Marcus Kleinehagenbrock, Sascha Hohenner, Jannik Fritsch, Gernot A. Fink, and Gerhard Sagerer. "Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot". In: *Proc. Int. Conf. on Multimodal Interfaces*. ACM, 2003 (cit. on p. 41).

[LMD12] Binh H. Le, Xiaohan Ma, and Zhigang Deng. "Live Speech Driven Head-and-Eye Motion Generators". In: *IEEE Transactions on Visualization and Computer Graphics* 18.11 (Nov. 2012), pp. 1902–1914. ISSN: 1077-2626. DOI: 10.1109/TVCG.2012.74 (cit. on p. 43).

[Lee+07] Jina Lee, Stacy C. Marsella, David Traum, Jonathan Gratch, and Brent Lance. "The Rickel Gaze Model: A Window on the Mind of a Virtual Human". In: *Lecture Notes in Artificial Intelligence; Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA)*. Vol. 4722. Paris, France, 2007, pp. 296–303 (cit. on p. 43).

[Lem+16] Séverin Lemaignan, Fernando Garcia, Alexis Jacq, and Pierre Dillenbourg. "From real-time attention assessment to "with-me-ness" in human-robot interaction". In: *ACM/IEEE International Conference on Human-Robot Interaction*. Vol. 2016-April. 2016, pp. 157–164. ISBN: 9781467383707. DOI: 10.1109/HRI.2016.7451747 (cit. on pp. 46, 47, 65).

[Lev89] William J. M. Levelt. *Speaking: From Intention to Articulation*. MIT Press, 1989. ISBN: 978-0-262-12137-8 (cit. on p. 33).

[Lev60] Howard Levene. *Robust tests for equality of variance*. 1960 (cit. on p. 127).

[Li+16] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. "Deep Reinforcement Learning for Dialogue Generation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (2016). DOI: 10.18653/v1/d16-1127 (cit. on p. 120).

[LLM15] Ruijiao Li, Bowen Lu, and Klaus D. McDonald-Maier. "Cognitive assisted living ambient system: a survey". In: *Digital Communications and Networks* 1.4 (2015), pp. 229–252. ISSN: 2352-8648. DOI: https://doi.org/10.1016/j.dcan.2015.10.003 (cit. on pp. 6, 10).

[Lin+15] Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen M Meng, and Li Deng. "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends". In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 35–52 (cit. on p. 74).

[Lis13] Pierre Lison. "Structured Probabilistic Modelling for Dialogue Management". PhD thesis. University of Oslo, 2013 (cit. on p. 79).

[LK16] Pierre Lison and Casey Kennington. "OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules". In: *Proceedings of ACL-2016 System Demonstrations*. 2016, pp. 67–72 (cit. on p. 77).

[Liu+16] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, pp. 2122–2132 (cit. on p. 120).

[LT00] Luís Seabra Lopes and Antonio Teixeira. "Human-robot interaction through spoken language dialogue". In: *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000) (Cat. No.00CH37113)*. Vol. 1. Oct. 2000, 528–534 vol.1. DOI: 10.1109/IROS.2000.894658 (cit. on p. 74).

[Lue+10] Ingo Luetkebohle, Frank Hegel, Simon Schulz, Matthias Hackel, Britta Wrede, Sven Wachsmuth, and Gerhard Sagerer. "The Bielefeld Anthropomorphic Robot Head Flobi". In: *2010 IEEE International Conference on Robotics and Automation*. accepted. IEEE. Anchorage, Alaska: IEEE, Mar. 2010 (cit. on pp. 41, 90).

[MBM15] Suhas R Mache, Manasi R Baheti, and C Namrata Mahender. "Review on text-to-speech synthesizer". In: *International Journal of Advanced Research in Computer and Communication Engineering* 4.8 (2015), pp. 54–59 (cit. on p. 74).

[McN00] David McNeill. *Language and gesture*. Vol. 2. Cambridge University Press, 2000 (cit. on p. 9).

[McT04] Michael F. McTear. "Components of a Spoken Dialogue System — Speech Input and Output". In: *Spoken Dialogue Technology: Toward the Conversational User Interface*. London: Springer London, 2004, pp. 79–105. ISBN: 978-0-85729-414-2. DOI: 10.1007/978-0-85729-414-2_5 (cit. on p. 74).

[Mey20] Sebastian Meyer zu Borgsen. "Nonverbal Communication During Human-Robot Object Handover. Improving Predictability of Humanoid Robots by Gaze and Gestures in Close Interaction". PhD thesis. Bielefeld University, 2020. DOI: https://doi.org/10.4119/unibi/2943731 (cit. on p. 42).

[Mic20] Thilo Michael. "Retico: An incremental framework for spoken dialogue systems". In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, July 2020, pp. 49–52 (cit. on pp. 76, 80).

[MSS06] Marek P. Michalowski, Selma Sabanovic, and Reid Simmons. "A spatial model of engagement for a social robot". In: *9th IEEE International Workshop on Advanced Motion Control, 2006*. Mar. 2006, pp. 762–767. DOI: 10.1109/AMC.2006.1631755 (cit. on p. 47).

[Mil68] Robert B. Miller. "Response Time in Man-computer Conversational Transactions". In: *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*. AFIPS '68 (Fall, part I). San Francisco, California: ACM, 1968, pp. 267–277. DOI: 10.1145/1476589.1476628 (cit. on p. 52).

[Moo+13] AJung Moon, Chris A. C. Parker, Elizabeth A. Croft, and H. F. Machiel Van der Loos. "Design and Impact of Hesitation Gestures During Human-robot Resource Conflicts". In: *J. Hum.-Robot Interact.* 2.3 (Sept. 2013), pp. 18–40. ISSN: 2163-0364. DOI: 10.5898/JHRI.2.3.Moon (cit. on p. 53).

[Moo+14] AJung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K.X.J. Pan, Minhua Zheng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A. Croft. "Meet Me Where I'm Gazing: How Shared Attention Gaze Affects Human-robot Handover Timing". In: *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*. HRI '14. Bielefeld, Germany: ACM, 2014, pp. 334–341. ISBN: 978-1-4503-2658-2. DOI: 10.1145/2559636.2559656 (cit. on pp. 41, 42).

[MDD97] Chris Moore, Philip J. Dunham, and Phil Dunham. "Joint Attention: Its Origins and Role in Development". In: *Contemporary Psychology* 42.1 (1997), p. 286 (cit. on p. 23).

[Mor70] Masahiro Mori. "The uncanny valley". In: *Energy* 7.4 (1970), pp. 33–35 (cit. on p. 87).

[Mur+07] Norman Murray, Dave Roberts, Anthony Steed, Paul Sharkey, Paul Dickerson, and John Rae. "An assessment of eye-gaze potential within immersive virtual environments". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 3.4 (2007), p. 8 (cit. on p. 43).

[MFH06] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. "A storytelling robot: Modeling and evaluation of human-like gaze behavior". In: *Humanoid robots, 2006 6th IEEE-RAS international conference on*. Citeseer. 2006, pp. 518–523 (cit. on pp. 42, 43).

[Mut+09] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. "Footing in human-robot conversations: how robots might shape participant roles using gaze cues". In: *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM. 2009, pp. 61–68 (cit. on p. 42).

[Nom+08] Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kensuke Kato. "Prediction of Human Behavior in Human–Robot Interaction Using Psychological Scales for Anxiety and Negative Attitudes Toward Robots". In: *IEEE Transactions on Robotics* 24.2 (Apr. 2008), pp. 442–451. ISSN: 1552-3098. DOI: 10.1109/TRO.2007.914004 (cit. on p. 118).

[Nor+18] Nahal Norouzi, Kangsoo Kim, Jason Hochreiter, Myungho Lee, Salam Daher, Gerd Bruder, and Greg Welch. "A Systematic survey of 15 years of user studies published in the intelligent virtual agents conference". In: *Proceedings of the 18th international conference on intelligent virtual agents*. ACM. 2018, pp. 17–22 (cit. on p. 7).

[OKJ06] Keith Oatley, Dacher Keltner, and Jennifer M Jenkins. *Understanding emotions*. Blackwell publishing, 2006 (cit. on p. 9).

[OKN14] Kengo Ohta, Norihide Kitaoka, and Seiichi Nakagawa. "Modeling filled pauses and silences for responses of a spoken dialogue system". In: *International Journal of Computers* 8 (2014), pp. 136–142 (cit. on pp. 52, 53, 214, 215, 219).

[Pal+15] Oskar Palinko, Alessandra Sciutti, Lars Schillingmann, Francesco Rea, Yukie Nagai, and Giulio Sandini. "Gaze contingency in turn-taking for human robot interaction: Advantages and drawbacks". In: *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*. IEEE. 2015, pp. 369–374 (cit. on pp. 43, 44, 50, 53, 64, 215).

[PE11] Gabriel Parent and Maxine Eskenazi. "Speaking to the crowd: looking at past achievements in using crowd-sourcing for speech and predicting future challenges". In: *Twelfth Annual Conference of the International Speech Communication Association*. 2011 (cit. on p. 120).

[PSM13] Cécile L Paris, William R Swartout, and William C Mann. *Natural language generation in artificial intelligence and computational linguistics*. Vol. 119. Springer Science & Business Media, 2013 (cit. on p. 73).

[Pel+12] J. Peltason, N. Riether, B. Wrede, and I. Lütkebohle. "Talking with robots about objects: A system-level evaluation in HRI". In: *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Mar. 2012, pp. 479–486. DOI: 10.1145/2157689.2157841 (cit. on p. 74).

[Pel14] Julia Peltason. "Modeling Human-Robot-Interaction based on generic Interaction Patterns". PhD thesis. Bielefeld University, 2014 (cit. on p. 78).

[PW10a] Julia Peltason and Britta Wrede. "Modeling human-robot interaction based on generic interaction patterns". In: *2010 AAAI Fall Symposium Series*. 2010 (cit. on pp. 78, 102).

[Pet+05] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. "A Model of Attention and Interest Using Gaze Behavior". In: *Lecture Notes in Computer Science* (2005), pp. 229–240. ISSN: 0954-3945. DOI: 10.1007/11550617_20 (cit. on pp. 24, 25, 47, 57).

[PH05] Roberto Pieraccini and Juan Huerta. "Where do we go from here? Research and commercial spoken dialog systems". In: *6th SIGdial Workshop on Discourse and Dialogue*. 2005 (cit. on pp. 74, 76).

[Pit+16] Karola Pitsch, Timo Dankert, Raphaela Gehle, and Sebastian Wrede. "Referential practices. Effects of a museum guide robot suggesting a deictic 'repair' action to visitors attempting to orient to an exhibit". In: *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Aug. 2016, pp. 225–231. DOI: 10.1109/ROMAN.2016.7745135 (cit. on pp. 45, 46, 63).

[Pit+09] Karola Pitsch, Hideaki Kuzuoka, Yuya Suzuki, Luise Sussenbach, Paul Luff, and Christian Heath. "The first five seconds: Contingent stepwise entry into an interaction as a means to secure sustained engagement in HRI". In: *Robot and Human Interactive Communication, 2009. RO-*

MAN 2009. The 18th IEEE International Symposium on. IEEE. 2009, pp. 985–991 (cit. on pp. 47, 48).

[Pot+04] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. "Audio-visual automatic speech recognition: An overview". In: *Issues in visual and audio-visual speech processing* 22 (2004), p. 23 (cit. on p. 74).

[Rei+18] Arsénio Reis, Dennis Paulino, Hugo Paredes, Isabel Barroso, Maria João Monteiro, Vitor Rodrigues, and João Barroso. "Using intelligent personal assistants to assist the elderlies An evaluation of Amazon Alexa, Google Assistant, Microsoft Cortana, and Apple Siri". In: *2018 2nd International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW)*. IEEE. 2018, pp. 1–5 (cit. on p. 7).

[Ric+10] Charles Rich, Brett Ponsleur, Aaron Holroyd, and Candace L. Sidner. "Recognizing Engagement in Human-robot Interaction". In: *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*. HRI '10. Osaka, Japan: IEEE Press, 2010, pp. 375–382. ISBN: 978-1-4244-4893-7 (cit. on p. 47).

[Ric20] Viktor Richter. "Addressing in Smart Environments. An Investigation of Human Conversational Behaviours Towards Devices and Autonomous Agents in a Smart Environment". PhD thesis. Bielefeld University, 2020. DOI: https://doi.org/10.4119/unibi/2943368 (cit. on p. 110).

[RK16] Viktor Richter and Franz Kummert. "Towards Addressee Recognition in Smart Robotic Environments: An Evidence Based Approach". In: *Proceedings of the 1st Workshop on Embodied Interaction with Smart Environments*. EISE '16. Tokyo, Japan: ACM, 2016, 2:1–2:6. ISBN: 978-1-4503-4555-2. DOI: 10.1145/3008028.3008030 (cit. on p. 76).

[Rie12] Laurel D Riek. "Wizard of oz studies in hri: a systematic review and new reporting guidelines". In: *Journal of Human-Robot Interaction* 1.1 (2012), pp. 119–136 (cit. on pp. 121, 123).

[Roh+20] Katharina J Rohlfing et al. "Explanation as a social practice: Toward a conceptual framework for the social design of AI systems". In: *IEEE Transactions on Cognitive and Developmental Systems* (2020) (cit. on p. 26).

[RGL06] Bernard Rosner, Robert J Glynn, and Mei-Ling T Lee. "The Wilcoxon signed rank test for paired comparisons of clustered data". In: *Biometrics* 62.1 (2006), pp. 185–192 (cit. on p. 127).

[Ruh+15] Kerstin Ruhland, Christopher E Peters, Sean Andrist, Jeremy B Badler, Norman I Badler, Michael Gleicher, Bilge Mutlu, and Rachel McDonnell. "A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception". In: *Computer Graphics Forum*. Vol. 34. Wiley Online Library. 2015, pp. 299–326 (cit. on pp. 39, 40).

[Sal+17] Hanan Salam, Oya Celiktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. "Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions". In: *IEEE Access* 5 (2017), pp. 705–721. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2016.2614525 (cit. on p. 48).

[SC15] Hanan Salam and Mohamed Chetouani. "A multi-level context-based modeling of engagement in Human-Robot Interaction". In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 03. May 2015, pp. 1–6. DOI: 10.1109/FG.2015.7284845 (cit. on p. 47).

[SPV13] Hugues Salamin, Anna Polychroniou, and Alessandro Vinciarelli. "Automatic detection of laughter and fillers in spontaneous mobile phone conversations". In: *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*. IEEE. 2013, pp. 4282–4287 (cit. on p. 48).

[SGY05] Jost Schatzmann, Kallirroi Georgila, and Steve Young. "Quantitative evaluation of user simulation techniques for spoken dialogue systems". In: *6th SIGdial Workshop on DISCOURSE and DIALOGUE*. 2005 (cit. on p. 119).

[Sch10] Matthias Scheutz. "Robust natural language dialogues for instruction tasks". In: *Unmanned Systems Technology XII*. Vol. 7692. International Society for Optics and Photonics. 2010, p. 769205 (cit. on p. 48).

[Sch05] David Schlangen. "Modelling dialogue: Challenges and Approaches". English. In: *Künstliche Intelligenz* 3/05 (2005), pp. 23–28 (cit. on pp. 77, 78).

[SS11] David Schlangen and Gabriel Skantze. "A General, Abstract Model of Incremental Dialogue Processing". English. In: *Dialogue and Discourse* 2.1 (2011), pp. 83–111. ISSN: 2152-9620 (cit. on pp. 76, 77, 80, 81).

[Sch20] Simon Schulz. "Design, Control, and Evaluation of a Human-Inspired Robotic Eye". PhD thesis. Bielefeld University, 2020. DOI: https://doi.org/10.4119/unibi/2943734 (cit. on p. 91).

[Sch+13] Simon Schulz, Florian Lier, Ingo Luetkebohle, and Sven Wachsmuth. "Robot reality - A motion capture system that makes robots become human and vice versa". In: *2013 IEEE International Conference on Robotics and Automation*. May 2013, pp. 2126–2133. DOI: 10.1109/ICRA.2013.6630862 (cit. on p. 41).

[SW65] Samuel S. SHAPIRO and Martin B. WILK. "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3-4 (Dec. 1965), pp. 591–611. DOI: 10.1093/biomet/52.3-4.591 (cit. on p. 127).

[SJD14] Kshitij Sharma, Patrick Jermann, and Pierre Dillenbourg. "With-me-ness: A gaze-measure for students' attention in MOOCs". In: *Proceedings of International Conference of the Learning Sciences 2014*. ISLS. 2014, pp. 1017–1022 (cit. on pp. 23, 24).

[She+13] Samira Sheikhi, Vasil Khalidov, David Klotz, Britta Wrede, and Jean-Marc Odobez. "Leveraging the Robot Dialog State for Visual Focus of Attention Recognition". In: *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*. ICMI '13. Sydney, Australia: ACM, 2013, pp. 107–110. ISBN: 978-1-4503-2129-7. DOI: 10.1145/2522848.2522881 (cit. on p. 45).

[Shi+08] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. "How Quickly Should Communication Robots Respond?" In: *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*. HRI '08. Amsterdam, The Netherlands: ACM, 2008, pp. 153–160. ISBN: 978-1-60558-017-3. DOI: 10.1145/1349822.1349843 (cit. on pp. 52, 53, 64).

[Shr99] Elizabeth E Shriberg. *Phonetic consequences of speech disfluency*. Tech. rep. SRI INTERNATIONAL MENLO PARK CA, 1999 (cit. on p. 48).

[Shr94] Elizabeth Ellen Shriberg. "Preliminaries to a theory of speech disfluencies". PhD thesis. Citeseer, 1994 (cit. on p. 31).

[Sid+04] Candace L. Sidner, Cory David Kidd, Christopher Lee, and Neal Lesh. "Where to look: a study of human-robot engagement". In: *Proceedings of the 9th international conference on Intelligent user interfaces*. 2004, pp. 78–84. ISBN: 1581138156. DOI: 10.1145/964442.964458 (cit. on p. 24).

[Sid+05] Candace L. Sidner, Christopher Lee, Cory D. Kidd, Neal Lesh, and Charles Rich. "Explorations in Engagement for Humans and Robots". In: *Artif. Intell.* 166.1-2 (Aug. 2005), pp. 140–164. ISSN: 0004-3702. DOI: 10.1016/j.artint.2005.03.005 (cit. on pp. 25, 42).

[SLL03] Candace L. Sidner, Christopher Lee, and Neal Lesh. "Engagement by Looking: Behaviors for Robots when Collaborating with People". In: *In Kruiff-Korbayova and Kosny (EDS.), Diabruck: The proceedings of the seventh workshop on the semantics and pragmatics of dialogue. University of Saarland*. 2003, pp. 123–130 (cit. on pp. 24, 25, 57).

[SG09] Gabriel Skantze and Joakim Gustafson. "Attention and Interaction Control in a Human-human-computer Dialogue Setting". In: *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. SIGDIAL '09. London, United Kingdom: Association for Computational Linguistics, 2009, pp. 310–313. ISBN: 978-1-932432-64-0 (cit. on pp. 44–46, 60, 65).

[SH13] Gabriel Skantze and Anna Hjalmarsson. "Towards incremental speech generation in conversational systems". In: *Computer Speech and Language* 27.1 (2013), pp. 243–262 (cit. on pp. 52, 53, 64).

[SH10] Gabriel Skantze and Anna Hjalmarsson. "Towards incremental speech generation in dialogue systems". In: *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics. 2010, pp. 1–8 (cit. on pp. 76, 80).

[SJB14] Gabriel Skantze, Martin Johansson, and Jonas Beskow. "Exploring Turn-taking Cues in Multi-party Human-robot Discussions about Objects". In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 2014, pp. 67–74 (cit. on p. 42).

[SBH95] Ronnie W Smith, Alan W Biermann, and D Richard Hipp. "An architecture for voice dialog systems based on prolog-style theorem proving". In: *Computational Linguistics* 21.3 (1995), pp. 281–320 (cit. on p. 77).

[Str+12] Kyle Strabala, Min Kyung Lee, Anca Dragan, Jodi Forlizzi, and Siddhartha S Srinivasa. "Learning the communication of intent prior to physical collaboration". In: *RO-MAN, 2012 IEEE*. IEEE. 2012, pp. 968–973 (cit. on pp. 30, 31).

[Sty06] Elizabeth Styles. *The psychology of attention*. Psychology Press, 2006 (cit. on pp. 19, 20).

[TF11] Barbara G. Tabachnick and Linda S. Fidell. "Multivariate Analysis of Variance (MANOVA)". In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 902–904. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_394 (cit. on p. 126).

[TF86] Michael Tomasello and Michael Jeffrey Farrar. "Joint Attention and Early Language". In: *Child Development* 57.6 (1986), pp. 1454–1463. ISSN: 00093920, 14678624. DOI: https://doi.org/10.2307/1130423 (cit. on p. 23).

[TL03] David R. Traum and Staffan Larsson. "The Information State Approach to Dialogue Management". English. In: *Current and New Directions in Discourse and Dialogue*. Ed. by Jan van Kuppevelt and RonnieW. Smith. Vol. 22. Text, Speech and Language Technology. Springer Netherlands, 2003, pp. 325–353. ISBN: 978-1-4020-1615-8. DOI: 10.1007/978-94-010-0019-2_15 (cit. on p. 78).

[Tre64] Anne Treisman. "Monitoring and storage of irrelevant messages in selective attention". In: *Journal of Verbal Learning and Verbal Behavior* 3.6 (1964), pp. 449–459 (cit. on pp. 20, 21).

[VYK14] Herwin Van Welbergen, Ramin Yaghoubzadeh, and Stefan Kopp. "AsapRealizer 2.0: The next steps in fluent behavior realization for ECAs". In: *International Conference on Intelligent Virtual Agents*. Springer. 2014, pp. 449–462 (cit. on p. 75).

[Van16] Aimee Van Wynsberghe. *Healthcare robots: Ethics, design and implementation*. Routledge, 2016 (cit. on p. 5).

[VJC16] Dominique Vaufreydaz, Wafa Johal, and Claudine Combe. "Starting engagement detection towards a companion robot using multimodal features". In: *Robotics and Autonomous Systems* 75 (2016). Assistance and Service Robotics in a Human Environment, pp. 4–16. ISSN: 0921-8890. DOI: https://doi.org/10.1016/j.robot.2015.01.004 (cit. on p. 47).

[Ver+01] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. "Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '01. Seattle, Washington, USA: ACM, 2001, pp. 301–308. ISBN: 1-58113-327-8. DOI: 10.1145/365024.365119 (cit. on pp. 29, 43).

[Vil+18] Valeria Villani, Fabio Pini, Francesco Leali, and Cristian Secchi. "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications". In: *Mechatronics* 55 (2018), pp. 248–266. ISSN: 0957-4158. DOI: https://doi.org/10.1016/j.mechatronics.2018.02.009 (cit. on p. 5).

[Vol+14] Anna-Lisa Vollmer, Manuel Mühlig, Jochen J Steil, Karola Pitsch, Jannik Fritsch, Katharina J Rohlfing, and Britta Wrede. "Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning". In: *PloS one* 9.3 (2014), e91349. DOI: https://doi.org/10.1371/journal.pone.0091349 (cit. on p. 41).

[Wal+97] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. "PARADISE: A framework for evaluating spoken dialogue agents". In: *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 1997, pp. 271–280 (cit. on pp. 118, 119).

[Wei10] Astrid Weiss. *Validation of an evaluation framework for human-robot interaction: the impact of usability, social acceptance, user experience, and societal impact on collaboration with humanoid robots*. na, 2010 (cit. on p. 121).

[WB15] Astrid Weiss and Christoph Bartneck. "Meta Analysis Of The Usage Of The Godspeed Questionnaire Series". In: *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN2015)*. IEEE, 2015, pp. 381–388. DOI: 10.1109/ROMAN.2015.7333568 (cit. on p. 118).

[Wel47] Bernard L. Welch. "The generalization of "Student's" problem when several different population variances are involved". In: *Biometrika* 34.1-2 (Jan. 1947), pp. 28–35. ISSN: 0006-3444. DOI: 10.1093/biomet/34.1-2.28 (cit. on p. 126).

[Wre+17] Sebastian Wrede, Christian Leichsenring, Patrick Holthaus, Thomas Hermann, Sven Wachsmuth, and The CSRA Team. "The Cognitive Service Robotics Apartment". In: *KI - Künstliche Intelligenz* 31.3 (Aug. 2017), pp. 299–304. ISSN: 1610-1987. DOI: 10.1007/s13218-017-0492-x. URL: https://doi.org/10.1007/s13218-017-0492-x (cit. on p. 92).

[Yan+10] Zhaojun Yang, Baichuan Li, Yi Zhu, Irwin King, Gina Levow, and Helen Meng. "Collection of user judgments on spoken dialog system with crowdsourcing". In: *2010 IEEE Spoken Language Technology Workshop*. 2010, pp. 277–282. DOI: 10.1109/SLT.2010.5700864 (cit. on p. 120).

[YAW04] Chen Yu, Paul M Aoki, and Allison Woodruff. "Detecting user engagement in everyday conversations". In: *arXiv preprint cs/0410027* (2004) (cit. on p. 48).

[YD16] Dong Yu and Li Deng. *AUTOMATIC SPEECH RECOGNITION*. Springer, 2016 (cit. on p. 72).

[YBH15] Zhou Yu, Dan Bohus, and Eric Horvitz. "Incremental coordination: Attention-centric speech production in a physically situated conversational agent". In: *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2015, p. 402 (cit. on pp. 45, 46, 50, 53, 62, 64, 65).

[Zai+18] Aws Alaa Zaidan et al. "A survey on communication components for IoT-based technologies in smart homes". In: *Telecommunication Systems* 69.1 (Sept. 2018), pp. 1–25. ISSN: 1572-9451. DOI: 10.1007/s11235-018-0430-8 (cit. on p. 6).

[Zar+14] Abolfazl Zaraki, Daniele Mazzei, Manuel Giuliani, and Danilo De Rossi. "Designing and Evaluating a Social Gaze-Control System for a Humanoid Robot". In: *IEEE Transactions on Human-Machine Systems* 44.2 (Apr. 2014), pp. 157–168. ISSN: 2168-2291. DOI: 10.1109/THMS.2014.2303083 (cit. on pp. 40, 41).

[ZE16] Tiancheng Zhao and Maxine Eskenazi. "Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning". In: *arXiv preprint arXiv:1606.02560* (2016) (cit. on p. 77).

SOFTWARE

[BS12] Timo Baumann and David Schlangen. "The InproTK 2012 Release". In: *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*. 2012, pp. 29–32 (cit. on pp. 76, 80, 83).

[Bie19a] CoR-Lab Universität Bielefeld. *Robotics Systems Types Repository (RST)—Documentation*. 2019. URL: http://docs.cor-lab.de//rst-manual/0.16/html/index.html (visited on 04/12/2019) (cit. on p. 93).

[Bie19b] CoR-Lab Universität Bielefeld. *RSB Documentation—Tools*. 2019. URL: http://docs.cor-lab.de//rsb-manual/0.16/html/tools.html (visited on 04/11/2019) (cit. on p. 105).

[Clo19] Google Cloud. *Cloud Speech-to-Text*. 2019. URL: https://cloud.google.com/speech-to-text/ (visited on 04/03/2019) (cit. on p. 96).

[Fac19] Central Lab Facilities. *Simple Robot Gaze Based On HLRC*. 2019. URL: https://github.com/CentralLabFacilities/simple_robot_gaze (visited on 04/12/2019) (cit. on p. 109).

[Kin09] Davis E. King. "Dlib-ml: A Machine Learning Toolkit". In: *Journal of Machine Learning Research* 10 (2009), pp. 1755–1758 (cit. on p. 96).

[Lam+03] Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. "The CMU SPHINX-4 speech recognition system". In: *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*. Vol. 1. Citeseer. 2003, pp. 2–5 (cit. on p. 96).

[LWW14] Florian Lier, Sven Wachsmuth, and Sebastian Wrede. "Modeling Software Systems in Experimental Robotics for Improved Reproducibility—A Case Study with the iCub Humanoid Robot". In: *2014 IEEE-RAS 14th International Conference on Humanoid Robots*. Madrid, Spain, 2014 (cit. on p. 93).

[Lim19] LimeSurvey Project Team / Carsten Schmitz. *LimeSurvey: An Open Source survey tool*. LimeSurvey Project. Hamburg, Germany, 2019. URL: http://www.limesurvey.org (visited on 03/11/2020) (cit. on p. 125).

[Lüt+11] Ingo Lütkebohle, Roland Philippsen, Vijay Pradeep, Eitan Marder Eppstein, and Sven Wachsmuth. "Generic middleware support for coordinating robot software components: The Task-State-Pattern". In: *Journal of Software Engineering in Robotics* 1.2 (2011), pp. 20–39. ISSN: 2035–3928 (cit. on pp. 78, 79).

[NLU19] RASA NLU. *Open source machine learning framework to automate text- and voice-based conversations*. 2019. URL: https://github.com/RasaHQ/rasa (visited on 04/12/2019) (cit. on p. 98).

[PW10b] Julia Peltason and Britta Wrede. "Pamini: A framework for assembling mixed-initiative human-robot interaction from generic interaction patterns". English. In: *Proceedings of the SIGDIAL 2010 Conference*. Tokyo, Japan: Association for Computational Linguistics (ACL), 2010, pp. 229–232 (cit. on pp. 75, 78–80).

[PLH19] Marian Pohling, Christian Leichsenring, and Thomas Hermann. "Base Cube One: A location-addressable service-oriented smart environment framework". In: *Journal of Ambient Intelligence and Smart Environments* 11.5 (2019), pp. 373–401 (cit. on p. 93).

[Pov+11] Daniel Povey et al. "The Kaldi speech recognition toolkit". In: *IEEE 2011 workshop on automatic speech recognition and understanding*. CONF. IEEE Signal Processing Society. 2011 (cit. on p. 96).

[R C19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. 2019. URL: https://www.R-project.org (visited on 03/11/2020) (cit. on p. 127).

[Rev19] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.9.12. Northwestern University. Evanston, Illinois, 2019. URL: https://CRAN.R-project.org/package=psych (cit. on p. 127).

[RSt15] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc. 2015. URL: http://www.rstudio.com/ (visited on 03/11/2020) (cit. on p. 127).

[SN15] Lars Schillingmann and Yukie Nagai. "Yet another gaze detector: An embodied calibration free system for the iCub robot". In: *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, Nov. 2015, pp. 8–13. DOI: 10.1109/HUMANOIDS.2015.7363515 (cit. on pp. 96, 145).

[ST03] Marc Schröder and Jürgen Trouvain. "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching". In: *International Journal of Speech Technology* 6.4 (Oct. 2003), pp. 365–377. DOI: 10.1023/A:1025708916924 (cit. on p. 104).

[Sch+16] Simon Schulz, Florian Lier, Andreas Kipp, and Sven Wachsmuth. "Humotion: A human inspired gaze control framework for anthropomorphic robot heads". In: *Proceedings of the Fourth International Conference on Human Agent Interaction*. 2016, pp. 207–214 (cit. on p. 104).

[SA12] Gabriel Skantze and Samer Al Moubayed. "IrisTK: A Statechart-based Toolkit for Multi-party Face-to-face Interaction". In: *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. ICMI '12. Santa Monica, California, USA: ACM, 2012, pp. 69–76. ISBN: 978-1-4503-1467-1. DOI: 10.1145/2388676.2388698 (cit. on p. 78).

[Sun10] Inc. Sun Microsystems. *JSpeech Grammar Format Specification*. 2010. URL: https://www.w3.org/TR/2000/NOTE-jsgf-20000605/ (visited on 04/12/2019) (cit. on p. 96).

[Tor20] Marco Torchiano. *effsize: Efficient Effect Size Computation*. R package version 0.7.8. 2020. DOI: 10.5281/zenodo.1480624. URL: https://CRAN.R-project.org/package=effsize (cit. on p. 127).

[Wic16] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org (cit. on p. 127).

[WW11] Johannes Wienke and Sebastian Wrede. "A Middleware for Collaborative Research in Experimental Robotics". In: *2011 IEEE/SICE International Symposium on System Integration*. SII. IEEE, 2011, pp. 1183–1190. DOI: 10.1109/SII.2011.6147617 (cit. on p. 93).

[Wit+06] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. "Elan: a professional framework for multimodality research". In: *Proceedings of LREC*. Vol. 2006. 2006, 5th (cit. on p. 125).

ONLINE RESOURCES

[Ama] Inc. Amazon.com. *Amazon Alexa*. URL: https://developer.amazon.com/de-DE/alexa (visited on 01/10/2021) (cit. on p. 6).

[Bie19a] CoR-Lab Universität Bielefeld. *Robotics Systems Types Repository (RST)—Documentation*. 2019. URL: http://docs.cor-lab.de//rst-manual/0.16/html/index.html (visited on 04/12/2019) (cit. on p. 93).

[Bie19b] CoR-Lab Universität Bielefeld. *RSB Documentation—Tools*. 2019. URL: http://docs.cor-lab.de//rsb-manual/0.16/html/tools.html (visited on 04/11/2019) (cit. on p. 105).

[Clo19] Google Cloud. *Cloud Speech-to-Text*. 2019. URL: https://cloud.google.com/speech-to-text/ (visited on 04/03/2019) (cit. on p. 96).

[Cor19a] 2019 Sony Corporation. *Entertainment Robot "aibo" Announced*. 2019. URL: https://www.sony.net/SonyInfo/News/Press/201711/17-105E/index.html (visited on 03/19/2019) (cit. on p. 5).

[Cor19b] iRobot Corporation. *Vorstellung der neuen Roomba®-Robotermodelle*. 2019. URL: https://www.irobot.de/haushaltsroboter/staubsaugen (visited on 03/19/2019) (cit. on p. 5).

[Fac19] Central Lab Facilities. *Simple Robot Gaze Based On HLRC*. 2019. URL: https://github.com/CentralLabFacilities/simple_robot_gaze (visited on 04/12/2019) (cit. on p. 109).

[Int19] Human-Agent Interaction. *What is HAI?* 2019. URL: http://hai-conference.net/what-is-hai/ (visited on 03/19/2019) (cit. on p. 10).

[Lim19] LimeSurvey Project Team / Carsten Schmitz. *LimeSurvey: An Open Source survey tool*. LimeSurvey Project. Hamburg, Germany, 2019. URL: http://www.limesurvey.org (visited on 03/11/2020) (cit. on p. 125).

[Lim20] Google Ireland Limited. *Google Assistant*. 2020. URL: https://assistant.google.com/intl/de_de/ (visited on 01/10/2021) (cit. on p. 5).

[Mer19a] Meriam-Webster. *attention 1b*. 2019. URL: https://www.merriam-webster.com/dictionary/attention (visited on 03/11/2019) (cit. on p. 19).

[Mer19b] Meriam-Webster. *dialogue 2*. 2019. URL: https://www.merriam-webster.com/dictionary/dialog (visited on 03/12/2019) (cit. on p. 71).

[Mer19c] Meriam-Webster. *interaction*. 2019. URL: https://www.merriam-webster.com/dictionary/interaction (visited on 03/12/2019) (cit. on p. 10).

[Mic21] Microsoft. *"Your personal productivity assistant in Microsoft 365"*. 2021. URL: https://www.microsoft.com/en-us/cortana (visited on 01/10/2021) (cit. on p. 6).

[NLU19] RASA NLU. *Open source machine learning framework to automate text- and voice-based conversations*. 2019. URL: https://github.com/RasaHQ/rasa (visited on 04/12/2019) (cit. on p. 98).

[Pre19a] Cambridge University Press. *interaction*. In: *Cambridge Dictionary*. Cambridge University Press, 2019. URL: https://dictionary.cambridge.org/dictionary/english/interaction (visited on 03/12/2019) (cit. on p. 10).

[Pre19b] Oxford University Press. *interaction*. In: *Oxford Living Dictionaries*. Oxford University Press, 2019. URL: https://en.oxforddictionaries.com/definition/interaction (visited on 03/12/2019) (cit. on pp. 10, 11).

[R C19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. 2019. URL: https://www.R-project.org (visited on 03/11/2020) (cit. on p. 127).

[Rev19] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.9.12. Northwestern University. Evanston, Illinois, 2019. URL: https://CRAN.R-project.org/package=psych (cit. on p. 127).

[RSt15] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc. 2015. URL: http://www.rstudio.com/ (visited on 03/11/2020) (cit. on p. 127).

[Sun10] Inc. Sun Microsystems. *JSpeech Grammar Format Specification*. 2010. URL: https://www.w3.org/TR/2000/NOTE-jsgf-20000605/ (visited on 04/12/2019) (cit. on p. 96).

[Tor20] Marco Torchiano. *effsize: Efficient Effect Size Computation*. R package version 0.7.8. 2020. DOI: 10.5281/zenodo.1480624. URL: https://CRAN.R-project.org/package=effsize (cit. on p. 127).

[Uni19] Bielefeld University. *The Cognitive Service Robotics Apartment as an Ambient Host*. 2019. URL: https://cit-ec.de/en/csra (visited on 03/19/2019) (cit. on pp. 89, 100).

[Wic16] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org (cit. on p. 127).

A

INTERFACES OF THE DIALOGUE SYSTEM

In this chapter, I present some selected interfaces of the dialogue system. These serve to better describe the tasks of the single dialogue module. In addition, they allow individual subsystems to be exchanged.

```
syntax = "proto2";

package rst.audition;

option java˙outer˙classname = "SoundChunkType";

/**
 * Objects of this represent a chunk of an audio stream.
 *
 * The audio information for one or more @ref .channels is stored in
 * @ref .data as a sequence of @ref .sample˙count encoded samples, the
 * encoding of which is described by @ref .endianness and @ref
 * .sample˙type.
 *
 * Depending on the sample rate (@ref .rate), such a chunk of audio
 * corresponds to a certain amount of time during which its samples
 * have been recorded.
 *
 * Interpretation of RSB timestamps:
 *
 * create:
 *         Capture time of the audio buffer. More precisely, the
 *         timestamp should correspond to the first sample contained
 *         in the buffer.
 *
 * @author David Klotz ¡dklotz@techfak.uni-bielefeld.de¿
 */
// @constraint(len(.data) == 8 * .channels * .sample˙count * TODO(.sample˙type))
// @create˙collection
message SoundChunk –

    /**
     * The possible data types for representing individual samples.
     */
    enum SampleType –

        /**
         * Signed 8-bit samples.
         */
        SAMPLE˙S8 = 0;

        /**
         * Unsigned 8-bit samples.
         */
        SAMPLE˙U8 = 1;

        /**
         * Signed 16-bit samples.
         */
        SAMPLE˙S16 = 2;

        /**
```

```
         * Unsigned 16-bit samples.
         */
        SAMPLE˙U16 = 4;

        /**
         * Signed 24-bit samples.
         */
        SAMPLE˙S24 = 8;

        /**
         * Unsigned 24-bit samples.
         */
        SAMPLE˙U24 = 16;

    ”

    /**
     * The possible byte-orders for representing samples.
     */
    enum EndianNess –

        /**
         * Samples are represented with little Endian byte-order.
         */
        ENDIAN˙LITTLE = 0;

        /**
         * Samples are represented with big Endian byte-order.
         */
        ENDIAN˙BIG = 1;
    ”

    /**
     * The sequences of bytes representing the samples of this sound
     * chunk.
     *
     * The value of this field must be interpreted according to the
     * values of the @ref .sample˙count, @ref .channels, @ref
     * .sample˙type and @ref .endianness fields.
     */
    required bytes data = 1;

    /**
     * The number of samples contained in @ref .data.
     */
    // @unit(number)
    required uint32 sample˙count = 2;

    /**
     * The number of channels for which samples are stored in @ref
     * .data.
     */
    // @unit(number)
    optional uint32 channels = 3 [default = 1];

    /**
     * The rate with which the samples stored in @ref .data haven been
     * recorded or should be played.
     */
    // @unit(hz)
    optional uint32 rate = 4 [default = 44100];

    /**
     * The data type used for the representation of samples in @ref
     * .data.
     */
```

```
    optional SampleType sample˙type = 5 [default = SAMPLE˙S16];

    /**
     * The Endianness used for the representation of samples in @ref
     * .data.
     */
    optional EndianNess endianness = 6 [default = ENDIAN˙LITTLE];
˝
```

Listing A.1: Inteface for a sound chunk for the communication with the *audio server*. This allow the system to be independent from a microphone as well as a speaker.

```
ssyntax = "proto2"

package rst.dialog;

import "rst/dialog/SpeechHypothesis.proto";

option java˙outer˙classname = "SpeechHypothesesType";

/**
 * A set of potential speech hypotheses for a single utterance
 * representing different interpretations.
 *
 * @author Birte Carlmeyer ¡bcarlmey@techfak.uni-bielefeld.de¿
 * @author Soeren Klett ¡sklett@techfak.uni-bielefeld.de¿
 */
message SpeechHypotheses –

    /**
     * The best speech recognition result.
     */
    required SpeechHypothesis best˙result = 1;

    /**
     * A list of potential other interpretations of the speech signal
     * ordered by confidence. The first entry represents the hypothesis
     * with the highest confidence. The @ref .best˙result is not
     * included in this list.
     */
    repeated SpeechHypothesis further˙results = 2;

    /**
     * Indicates whether the current result might be superseded with
     * further results due to incremental processing or not. A value of
     * true indicates that no further hypotheses for the represented
     * speech utterance will be produced in the future.
     */
    required bool final = 3;

˝
```

Listing A.2: Inteface for a list of speech hypotheses to allow n-best results of the ASR.

```
syntax = "proto2";

package rst.dialog;

import "rst/timing/Interval.proto";

```

```
option java˙outer˙classname = ”SpeechHypothesisType”;

/**
 * A hypothesis about a speech recognition result.
 *
 * @author Birte Carlmeyer ¡bcarlmey@techfak.uni-bielefeld.de¿
 * @author Soeren Klett ¡sklett@techfak.uni-bielefeld.de¿
 */
message SpeechHypothesis –

    /**
     * Speech recognition result for a single word.
     */
    message Word –

        /**
         * Speech recognition result for a single word.
         */
        required string word = 1;

        /**
         * Start and end time for this word.
         *
         * If specified, this must be within the bounds of
         * @ref .rst.dialog.SpeechHypothesis.timestamp
         */
        optional timing.Interval timestamps = 2;

        /**
         * Part-of-speech tags for German using a modified version of
         * the Stuttgart-Tuebingen-Tagset (STTS).
         *
         * Differences w.r.t. STTS are:
         *
         * * KOMM instead of $,
         * * END instead of $.
         * * IPNCT instead of $(
         *
         * @see http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/stts.a
         *      ”Description of the STTS (in German)”
         */
        enum PartOfSpeechTag –...”

        /**
         * Stuttgart-Tuebingen-Tagset (STTS) Part-of-speech tag for
         * this word (assumes German).
         */
        optional PartOfSpeechTag part˙of˙speech˙tag = 3;

    ”

    /**
     * List of word speech recognition results.
     */
    repeated Word words = 1;

    /**
     * Confidence for this speech hypothesis.
     */
    // @constraint(0 ¡= value ¡= 1)
    optional float confidence = 2;

    /**
     * Start and end time for this speech hypothesis. Since some speech
     * recognizers may not provide detailed results for each word, this
     * field may be used to indicate the time of the overall hypothesis.
```

```
     */
    optional timing.Interval timestamp = 3;

    /**
     * The grammar tree of this speech hypothesis.
     */
    optional string grammar˙tree = 4;

”
```

Listing A.3: Inteface for a single speech hypothesis of the ASR.

```
syntax = "proto2";

package rst.dialog;

option java˙outer˙classname = "DialogActType";

import "rst/dialog/IncrementalUnit.proto";
import "rst/dialog/SpeechHypotheses.proto";

/**
 * A description of a dialog act as a specialization of an incremental
 * unit containing a representation of the underlying speech hypotheses.
 *
 * @author Birte Carlmeyer ¡bcarlmey@techfak.uni-bielefeld.de¿
 */
// @create˙collection
message DialogueAct –

    /**
     * Possible kinds of dialog acts.
     */
    enum Type –

        /**
         * Conversation opening.
         */
        GREET = 0;

        /**
         * Affirmative answer.
         */
        ACCEPT = 1;

        /**
         * Negative answer.
         */
        REJECT = 2;

        /**
         * Confirmation.
         */
        CONFIRM = 3;

        /**
         * Negation.
         */
        NEGATE = 4;

        /**
         * Speaker wants an information from addressee.
         */
        INFO˙REQUEST = 5;

```

```
        /**
         * Speaker offers to perform an action.
         */
        ACTION˙REQUEST = 6;

        /**
         * A statement.
         */
        STATEMENT = 7;

        /**
         * An answer (not a yes/no answer).
         */
        ANSWER = 8;

        /**
         * Conversation closing.
         */
        GOODBYE = 9;

        /**
         * The current dialog act doesn't match to any of the previous
         * types.
         */
        OTHER = 100;

    ”


    /**
     * The type of the current dialog act.
     */
    required Type type = 1;

    /**
     * The basic information of the dialog act IU.
     */
    required IncrementalUnit incremental˙unit = 2;

    /**
     * The underlying speech recognition result causing this dialog act.
     */
    optional SpeechHypotheses speech˙hypotheses = 3;

”
```

Listing A.4: Inteface for a dialogue act. This is the result of the of the NLU and is the basis for the DM decisions.

```
syntax = "proto2";

package rst.dialog;

option java˙outer˙classname = "IncrementalUnitType";

import "rst/timing/Interval.proto";

/**
 * A description of the basic information of an incremental unit (IU) of
 * the IU-model of incremental dialogue processing.
 *
 * @see http://wwwhomes.uni-bielefeld.de/dschlangen/inpro/abstractModel.html
 *      "A General, Abstract Model of Incremental Dialogue Processing"
 *
 * @author Birte Carlmeyer ¡bcarlmey@techfak.uni-bielefeld.de¿
```

```
 */
// @create˙collection
message IncrementalUnit –

    /**
     * The incremental unit can be in different states.
     */
    enum EditType –

        /**
         * Indicates a new incremental unit.
         */
        ADD = 0;

        /**
         * Replaces the last incremental unit.
         */
        UPDATED = 1;

        /**
         * Revokes a previously added unit.
         */
        REVOKE = 2;

        /**
         * The incremental unit has been finally committed and will not
         * be changed any more.
         */
        COMMIT = 3;

    ”

    /**
     * Start and end time of the incremental unit.
     */
    optional timing.Interval timestamps = 2;

    /**
     * A unique id for the incremental unit.
     */
    required bytes id = 3;

    /**
     * Optional id list of IU which have a same-level link.
     * Same-level links (sll) connect IUs, which are produced by the
     * same module and reflect their temporal order.
     */
    repeated bytes sll˙id = 4;

    /**
     * Optional id list of IU which have a grounded-in link.
     * Grounded-in links (gil) represent on which IUs they depend, thus
     * representing the possibility to build a hierarchical structure.
     * Entries are sorted by time and nesting level of the created
     * graph.
     */
    repeated bytes gil˙id = 5;

    /**
     * Edit type of the incremental unit.
     */
    required EditType state = 6;

”
```

Listing A.5: Inteface for a incremental unit to allow incremental processing. This is the basis for the communication with inprotk.

```
syntax = "proto2";

package rst.communicationpatterns;

option java˙outer˙classname = "TaskStateType";

/**
 * Represents the initiation or update of a potentially long-running task.
 *
 * The task is represented as a current state (@ref .state field) and
 * a datum or "specification" (represented by the @ref .wire˙schema
 * and @ref .payload fields).
 *
 * @see http://opensource.cit-ec.de/projects/xtt/wiki/TaskStateProtocol
 *      "A detailed description of the task state protocol"
 *
 * @author Jan Moringen ¡jmoringe@techfak.uni-bielefeld.de¿
 */
// @create˙collection
message TaskState –

    /**
     * Possible states of the task an update of which the @ref
     * .TaskState object represents.
     *
     * Initial task submission:
     *
     *   Applicable in states: none (since initial submission)
     *
     *   Possible values of @ref .state:
     *
     *   * @ref .INITIATED
     *   * @ref .ACCEPTED
     *   * @ref .REJECTED
     *
     * Updated of an accepted task:
     *
     *   Applicable in states: @ref .ACCEPTED
     *
     *   Possible values of @ref .state:
     *
     *   * @ref .UPDATE
     *   * @ref .UPDATE˙REJECTED
     *
     * Aborting an accepted task:
     *
     *   Applicable in states: @ref .ACCEPTED
     *
     *   Possible values of @ref .state:
     *
     *   * @ref .ABORT
     *   * @ref .ABORTED
     *   * @ref .ABORT˙FAILED
     *
     * Final states:
     *
     *   * @ref .RESULT˙AVAILABLE
     *   * @ref .COMPLETED
     *   * @ref .FAILED
     *
```

```
     */
    enum State –

        /**
         * Client submits specification with @ref .INITIATED state.
         *
         * Server accepts or rejects task and publishes specification
         * with @ref .ACCEPTED or @ref .REJECTED state accordingly.
         */
        INITIATED                = 0;

        /**
         * See @ref .INITIATED.
         */
        ACCEPTED                 = 1;

        /**
         * See @ref .INITIATED.
         */
        REJECTED                 = 2;

        /**
         * Client publishes (modified) specification with @ref .UPDATE
         * state.
         *
         * Server accepts or rejects the update and publishes
         * specification with @ref .ACCEPTED or @ref .UPDATE˙REJECTED
         * state accordingly.
         */
        UPDATE            = 3;

        /**
         * See @ref .UPDATE.
         */
        UPDATE˙REJECTED   = 4;

        /**
         * Client publishes specification with @ref .ABORT state to
         * request execution of the task to be aborted.
         *
         * Server accept or rejects this and publishes specification
         * with @ref .ABORTED or @ref .ABORT˙FAILED state accordingly.
         */
        ABORT               = 5;

        /**
         * See @ref .ABORT.
         */
        ABORTED             = 6;

        /**
         * See @ref .ABORT.
         */
        ABORT˙FAILED        = 7;

        /**
         * @todo document
         */
        RESULT˙AVAILABLE = 8;

        /**
         * See @ref .RESULT˙AVAILABLE.
         */
        COMPLETED        = 9;

        /**
```

```
           * See @ref .RESULT˙AVAILABLE.
           */
          FAILED            = 10;
    ”

    /**
     * Describes the role of the component doing the update.
     */
    enum Origin –

        /**
         * The task state update is performed by the submitter.
         */
        SUBMITTER = 0;

        /**
         * The task state update is performed by the handler.
         */
        HANDLER   = 1;
    ”

    /**
     * Describes the origin of the update.
     *
     * Has to be @ref .Origin.SUBMITTER when the task is being
     * initiated.
     */
    required Origin origin         = 1;

    /**
     * State to which the task should be updated.
     *
     * Has to be @ref .State.INITIATED when the task is being
     * initiated.
     */
    required State    state        = 2;

    /**
     * TODO
     */
    required uint32 serial         = 3;

    /**
     * Type of the payload blob.
     *
     * This field and the @ref .payload field are intended to be
     * processed by a (de)serialization mechanism that decodes/encodes
     * the payload blob according to the type information in
     * wireSchema.
     *
     * @todo "type should be ascii-string"
     */
    required bytes    wire˙schema = 4;

    /**
     * See @ref .wire˙schema.
     */
    required bytes    payload     = 5;

”
```

Listing A.6: Inteface for a system task: this allow the communication with other back-end systems and is a key concept of Pamini.

```
syntax = "proto2";
```

```

package rst.tts;

import "rst/tts/Prosody.proto";

option java˙outer˙classname = "TextToSpeechInstructionType";

/**
 * Instructions to a Text-to-Speech module regarding the production of
 * text.
 *
 * @author Soeren Klett ¡sklett@techfak.uni-bielefeld.de¿
 * @author Birte Carlmeyer ¡bcarlmey@techfak.uni-bielefeld.de¿
 */
message TextToSpeechInstruction –

    /**
     * The text to produce in case of @ref .PlaybackOption.PLAY. In all
     * other cases this needs to be empty.
     */
    optional string text = 1;

    /**
     * Prosody to be applied to everything contained in @ref .text.
     */
    optional Prosody prosody = 2;

    /**
     * Possible actions the TTS engine has to perform.
     */
    enum PlaybackOption –

        /**
         * Produce the text given in @ref .text.
         * If TTS is already playing a text message, this command
         * should be ignored.
         */
        PLAY = 0;

        /**
         * Stop the current production and discard it.
         */
        STOP = 1;

        /**
         * Pause the current production. This allows to resume it using
         * @ref .RESUME.
         */
        PAUSE = 2;

        /**
         * Resume a previously pause text production.
         * If nothing has been paused before, this should be ignored.
         */
        RESUME = 3;

    ”

    /**
     * Action to be executed by the Text-to-Speech engine.
     */
    optional PlaybackOption playback˙option = 3 [default = PLAY];

”
```

Listing A.7: Inteface for a TTS task: it allows the sythesis to stop or pause while speaking.

```
syntax = "proto2";

package rst.vision;

import "rst/vision/Face.proto";

option java˙outer˙classname = "FaceWithGazeType";

/**
 * An object of this type represents a human face detected in an image
 * including gaze informations.
 *
 * @author "Birte Carlmeyer" ¡bcarlmey@techfak.uni-bielefeld.de¿
 */
// @create˙collection
message FaceWithGaze –

    /**
     * The region of the image which corresponds to the face.
     */
    required Face region = 1;

    /**
     * If this is true, the eyelids are closed.
     */
    optional bool lid˙closed = 2;

    /**
     * Horizontal gaze estimation angle. A relative rotation from the
     * default gaze when the person looks straight into the camera.
     * Positive values means that the person is looking upwards and
     * negative value means that the person is looking downwards.
     */
    // @unit(radian)
    optional double horizontal˙gaze˙estimation = 3;

    /**
     * Vertical gaze estimation angle. A relative rotation from the
     * default gaze when the person looks straight into the camera.
     * Positive values means that the person is looking to the right and
     * negative values means that the person is looking to the left
     * (from the persons point of view).
     */
    // @unit(radian)
    optional double vertical˙gaze˙estimation = 4;

”
```

Listing A.8: Interfaces of the gaze estimation. Based on this, further features can be estimated, such as mutual gaze.

B QUESTIONNAIRES

B.1 TASK PERFORMANCE ASSESSMENT IN EC1

Hat Flobi folgendes wörtlich gesagt? *Did Flobi say the following literally?*		
Ich habe insgesamt neun Aktoren. *I have nine actuators in total.*	◯ ja ◯ yes	◯ nein ◯ no
Davon sind vier in meinen Lippen. *Four of them are in my lips.*	◯ ja ◯ *yes*	◯ nein ◯ *no*
Ich kann sechs verschiedene Emotionsausdrücke darstellen. *I can express six different expressions of emotions.*	◯ ja ◯ *yes*	◯ nein ◯ *no*
Freundlich gefällt mir natürlich am Besten *Of course, I like friendly the most.*	◯ ja ◯ *yes*	◯ nein ◯ *no*
Insgesamt habe ich fünf verschiedene Haarfarben. *I have five different hair colors in total.*	◯ ja ◯ *yes*	◯ nein ◯ *no*
Und drei verschiedenfarbige Augenbrauen. *And three different colored eyebrows.*	◯ ja ◯ *yes*	◯ nein ◯ *no*

Table B.1: Questionnaire to examine the information recall in EC1.

B.2 TASK PERFORMANCE ASSESSMENT IN EC2

Flobi hat Ihnen einiges über sich und das Apartment erzählt. Was hat Flobi gesagt? *Flobi told you a lot about himself and the apartment. What did Flobi say?*					
Welche Emotion kann ich darstellen? *Which emotions can I express?*	○ Traurigkeit ○ *sadness*	○ Scham ○ *shame*	○ Wut ○ *anger*	○ Angst. ○ *anxiety*	○ Das weiß ich nicht ○ *I don't know*
In welcher Farbe leuchten die Türgriffe? *What color do the door handles light up?*	○ Blau ○ *blue*	○ Rot ○ *red*	○ Grün ○ *green*	○ Gelb ○ *yellow*	○ Das weiß ich nicht ○ *I don't know*
Wofür werden Bildschirm und Tisch im Wohnzimmer genutzt? *What are the screen and table used for in the living room?*	○ Filme anschauen ○ *watch movies*	○ Demos vorführen ○ *show demos*	○ Spielen ○ *playing*	○ Arbeiten ○ *working*	○ Das weiß ich nicht ○ *I don't know*
Wie viel Emotionen kann ich darstellen? *How many emotions can I represent?*	○ Drei ○ *three*	○ Vier ○ *four*	○ Fünf ○ *five*	○ Sechs ○ *six*	○ Das weiß ich nicht ○ *I don't know*
Was ist an der Decke angebracht? *What is attached to the ceiling?*	○ Bewegungssensoren ○ *motion sensors*	○ Lampen ○ *lamps*	○ Bildschirme ○ *screens*	○ Feuermelder ○ *fire detectors*	○ Das weiß ich nicht ○ *I don't know*
Was ist meine Lieblingshaarfarbe? *What is my favorite hair color?*	○ Gelb ○ *yellow*	○ Rot ○ *red*	○ Blau ○ *blue*	○ Grün ○ *green*	○ Das weiß ich nicht ○ *I don't know*
Wann leuchten die Türgriffe blau auf? *When do the handles light up blue?*	○ Sobald die Küchentür auf geht. ○ *As soon as the door opens.*	○ Wenn die Musik läuft. ○ *When the music plays.*	○ Wenn dort Aufmerksamkeit hingelenkt werden soll. ○ *When attention is to be drawn there.*	○ Wenn jemand an der Tür klingelt. ○ *When someone rings the doorbell.*	○ Das weiß ich nicht ○ *I don't know*
Bitte wählen Sie eine der folgenden Antworten: *Please choose one of the following answers:*	○ Gesichtsfarbe ○ *complexion*	○ Haarfarbe ○ *hair color*	○ Gesichtsform ○ *face shape*	○ Frisur ○ *hairstyle*	○ Das weiß ich nicht ○ *I don't know*
Was kann man sich auf dem Tisch anschauen? *What can you see on the table?*	○ Videos ○ *videos*	○ Fotos von Flobi ○ *Photos of Flobi*	○ Nichts ○ *nothing*	○ Die Karte der Wohnung ○ *the apartment map*	○ Das weiß ich nicht ○ *I don't know.*
Die Küche ist... *The kitchen is...*	○ ...voll funktionsfähig. ○ *...fully functional.*	○ ...zur Zeit im Umbau. ○ *...currently under renovation.*	○ ...noch nicht ganz fertig. ○ *...not quite finished yet.*	○ ...teilweise benutzbar. ○ *...partially usable.*	○ Das weiß ich nicht ○ *I don't know.*

Table B.2: Questionnaire to examine the information recall in EC2.

B.3 MOS-BASED SYNTHESIS EVALUATION

Participants rate their overall impression of speech synthesis quality on a 5-point MOS scale. This scale was chosen for maximum comparability with traditional MOS-based synthesis evaluation:

Wie beurteilen Sie die Qualität der Stimme des Agenten:		
sehr schlecht	○○○○○	sehr gut

Table B.3: Traditional MOS-based synthesis evaluation.

B.4 THE GODSPEED QUESTIONNAIRE SERIES

Items translated from the GQS based on [Bar+09]:

Bitte bewerten Sie Flobi auf der folgenden Skala:

unecht	○○○○○	natürlich
wie eine Maschine	○○○○○	wie ein Mensch
hat kein Bewusstsein	○○○○○	hat ein Bewusstsein
künstlich	○○○○○	realistisch
bewegt sich steif	○○○○○	bewegt sich flüssig
tot	○○○○○	lebendig
unbewegt	○○○○○	lebendig
mechanisch	○○○○○	organisch
träge	○○○○○	interaktiv
apathisch	○○○○○	reagierend
nicht mögen	○○○○○	mögen
unfreundlich	○○○○○	freundlich
unhöflich	○○○○○	höflich
unangenehm	○○○○○	angenehm
furchtbar	○○○○○	nett
inkompetent	○○○○○	kompetent
ungebildet	○○○○○	unterrichtet
verantwortungslos	○○○○○	verantwortungsbewusst
unintelligent	○○○○○	intelligent
unvernünftig	○○○○○	vernünftig
ängstlich	○○○○○	entspannt
ruhig	○○○○○	aufgewühlt
still	○○○○○	überrascht

Table B.4: Questionnaire used in interaction studies to evaluate the subjective ratings of the five key concepts anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of the agent.

B.5 PREVIOUS EXPERIENCE ASSESSMENT

Bitte geben Sie an, wieviel Erfahrung Sie haben mit:
(Please specify how much experience you have with:)

Nutzung von Computern
(Using of computers)
keine ○○○○○○ sehr viel
(no experience) *(lots of experience)*

Nutzung von Systemen mit Spracheingabe
(Using systems with voice input)
keine ○○○○○○ sehr viel
(no experience) *(lots of experience)*

Nutzung von Robotersystemen
(Using of robotic systems)
keine ○○○○○○ sehr viel
(no experience) *(lots of experience)*

Programmierung von Computern
(Programming of computers)
keine ○○○○○○ sehr viel
(no experience) *(lots of experience)*

Dem Roboter Flobi oder seiner Simulation
(The Flobi robot or its simulation)
keine ○○○○○○ sehr viel
(no experience) *(lots of experience)*

Table B.5: Questionnaire to examine previous experience of the participants with technical systems.

B.6 ASSESSMENT OF APPROPRIATENESS OF AGENT'S STATEMENTS

Allgemein waren die Informationen von Flobi:
(In general, the information from Flobi was:)

zeitlich unangemessen *(inappropriately timed)*	○ ○ ○ ○ ○ ○	zeitlich angemessen *(timely)*
verzögert *(delayed)*	○ ○ ○ ○ ○ ○	passend *(suitable)*
zu schnell *(too fast)*	○ ○ ○ ○ ○ ○	passend *(suitable)*
zu langsam *(too slow)*	○ ○ ○ ○ ○ ○	passend *(suitable)*
unverständlich *(incomprehensible)*	○ ○ ○ ○ ○ ○	gut verständlich *(easy to understand)*
zu lang *(too long)*	○ ○ ○ ○ ○ ○	passend *(suitable)*
zu kurz *(too short)*	○ ○ ○ ○ ○ ○	passend *(suitable)*

Table B.6: Questionnaire to examine the appropriateness of Flobi's statements regarding their length and timing.

C

STUDY STIMULI

C.1 EXAMPLE STIMULI EC1

C.1.1 *Greeting of the Agent*

Example C.1.1: Introduction:

"Hallo, schön, dass du an dieser Studie teilnimmst. Mein Name ist Flobi und ich bin dein virtueller Ansprechpartner in dieser Studie."

"Hello, thank you for participating in this study. My name is Flobi and I am your virtual contact person in this study."

C.1.2 *Information about the Agent*

Example C.1.2: Inforamtion apout the agent:

"Mich gibt es auch als richtigen Roboter Kopf.
*Ich habe insgesamt 9 Aktoren. [**Disruption**]*
Davon sind 4 in meinen Lippen.
Ich kann 5 verschiedene Emotionsausdrücke darstellen.
Fröhlich gefällt mir natürlich am Besten.
Außerdem kann ich meine Haarfarbe wechseln.
Insgesamt habe ich 5 verschiedene Haarfarben.
Und 4 verschieden farbige Augenbrauen."

"I also exist as a real robot head.
I have 9 actuators in total. [**Disruption**]
4 of them are in my lips.
I can represent 5 different expressions of emotions.
Of course, I like happy the most.
I can also change my hair color.
I have a total of 5 different hair colors.
And 4 different colored eyebrows."

C.1.3 *Farewell of the Agent*

Example C.1.3: Farewell:

"So, genug von mir. Jetzt gehe bitte ins Es-zimmer und setz dich an den Rechner. Dort kannst du mit dem Fragebogen starten."

"So enough of me. Now please go to the dining room and sit down at the computer. There you can start with the questionnaire."

C.2 EXAMPLE STIMULI EC2

Example C.2.1: Information about the agend and the CSRA:

*"Ich bin in der Lage fünf verschiedene Emotionsausdrücke darstellen, davon fröhlich und [**Disruption**] auch wütend.*
Außerdem kann ich meine Haarfarbe und meine Augenbrauen wechseln.
Meine Lieblinghaarsfarbe ist blau.
Du bist hier in einem intelligenten Apartment, das mit einer Menge Technik ausgestattet ist.
***Zu deiner linken Seite** (—) siehst du dich Küche.*
Die Küche ist voll funktionsfähig.
An den Schränken sind an manchen Türgriffen Leuchten angebracht.
*Diese können blau [**Disruption**] aufleuchten, wenn ich deine Aufmerksamkeit dahin lenken möchte.*
Sobald der Schrank geöffnet wird, leuchten sie dann grün und wenn sie wieder geschlossen werden hören sie auf zu leuchten.
***Rechts von dir** (—) ist das Wohnzimmer.*
Wie du siehst hängt dort ein großer Bildschirm rechts an der Wand.
Der Tisch, der dort steht, ist interaktiv.
*Der Tisch und auch der Bildschirm können genutzt werden, [**Disruption**] um bei einer Besprechung Präsentationen und Demos zu zeigen.*
Das Apartment ist auch mit einer Reihe von Kameras und Bewegungsdetektoren ausgestattet, die größtenteils an der Decke hänge"

"I am able to represent five different expressions of emotion, including happy and [**Disruption**] angry.
I can also change my hair color and eyebrows.
My favorite hair color is blue.
You're here in an intelligent apartment that is equipped with a lot of technology.

To your left (—) you see kitchen.
The kitchen is fully functional.
There are lights on some door handles on the cupboards.
These can light up blue [**Disruption**] when I want to draw your attention to them.
When the cabinet is opened, they will turn green and when they are closed, they will stop glowing.
To your right (—) is the living room.
As you can see, there is a large screen hanging on the wall to the right.
The table standing there is interactive.
The table and the screen can be used [**Disruption**] to show presentations and demos during a meeting.
The apartment is also equipped with a number of cameras and motion detectors, most of which hang from the ceiling."

C.3 EXAMPLE STIMULI EC4

Example C.3.1: Introduction and Coverstory:

"Hallo, schön, dass du an dieser Studie teilnimmst.
Ich werde dir heute ein wenig über dieses Apartment erzählen, und dann habe ich eine kleine Aufgabe für dich.
Du könntest mir nämlich beim Suchen helfen.
Hier sind eben ein paar Sachen verloren gegangen.
Einige Handwerker waren hier im Apartment und haben die Küche umgebaut.
Ich konnte wegen des Staubs leider nicht genau erkennen, wo die Sachen versteckt wurden."

"Hello, nice of you to participate in this study.
I'm going to tell you a little bit about this apartment today, and then I have small task for you.
Because you could help me look.
Here are just a few things has been lost.
Some craftsmen were here in the apartment and rebuilt the kitchen.
Because of the dust I could not see exactly where the things were hidden."

Example C.3.2: Sweet part:

"Jemand hat die Waschmaschine bedient und das Waschpulverfach geöffnet.
Und ich habe gesehen, wie jemand zur Pflanze im Wohnzimmer gegan-

gen ist, und etwas am Blumentopf gemacht hat.
Danach hat jemand die Besteckschublade geöffnet und hat dort rumgewühlt.
Und dann habe ich beobachtet dass jemand den Schrank über der Mikrowelle aufgemacht hat.
Dann wurde einer der Stühle im Wohnzimmer bewegt.
Irgend etwas ist mit den Kaffeetassen auf dem Tisch im Wohnzimmer passiert.
Zu guter Letzt war noch jemand am Besteckfach der Spülmaschine."

"Someone started the washing machine and opened the washing powder compartment.
And I saw someone go to the plant in the living room and do something on the flower pot.
Then someone opened the cutlery drawer and rummaged around.
And then I noticed that somebody opened the cupboard above the microwave has opened.
Then one of the chairs in the living room was moved.
Something happened to the coffee cups on the table in the living room.
(**Disruption**) Last but not least, there was someone at the cutlery tray of the dishwasher."

Example C.3.3: Conclusion:

"Schau in beliebiger Reihenfolge an den Orten nach, die ich dir genannt habe."
"Look in any order at the places I told you to look."

C.4 EXAMPLE STIMULI EC5

Example C.4.1: Information part:

" Hallo, schön, dass du an dieser Studie teilnimmst.
Mein Name ist Flobi und ich bin dein virtueller Ansprechpartner hier im Apartment.
Ich werde dir heute ein wenig über dieses Apartment erzählen.
Danach habe ich eine kleine Aufgabe für dich.
Du befindest dich in der Küche.
Sobald die Bauarbeiten abgeschlossen sind, funktioniert sie auch wieder vollständig.
***Unter dir** (—) befindet sich ein kapazitiver Bodenbelag.*
Dieser hilft mir zu wissen wo du gerade bist.
*(**Disruption**) Damit wir uns auch gut verstehen, siehst du zum*

*Beispiel dass **über dir]** (—) ein Mikrophon in der Decke ist.*
***Links in der Küche** (—), über der Spüle,*
ist ein Fenster durch das ich raus in das CITEC gucken kann.
An den Schränken sind an manchen Türgriffen LEDs angebracht.
Diese werden blau, wenn ich dir dort etwas zeigen möchte.
*(**Disruption**) Das Viereck neben mir auf der Arbeitsplatte ist eine digitale Waage.*
So hast du immer eine griffbereit und du kannst sie nie verlegen.
Übrigens ist im Schrank über dem Herd eine Kamera enthalten.
Diese kann die Temperatur messen, so dass deine Milch nie mehr überläuft.
***Rechts von dir** (—), ist das Wohnzimmer.*
Der Tisch, der dort steht, ist ein interaktives Samsung Surface.
Man kann sich auf ihm einen Plan von dieser Wohnung anschauen.
Zusätzlich dazu haben wir einen Schwenk-Neige-Scheinwerfer oben an der Decke im Wohnzimmer befestigt.
Dieser kann bestimmte Bereiche im Wohnzimmer besonders hervorheben.
*(**Disruption**) Siehst du auf der Fensterbank meinen realen Kopf?*
Sei froh, dass deiner angewachsen ist.
***Hinter dir** (—), beim Flur gegenüber,*
befindet sich ein intelligenter Spiegel.
Dort kannst du dich von hinten angucken
Gleich kannst du dich noch weiter umsehen.
Bitte geh nun erstmal zum Tisch und fülle den ersten Teil des Fragebogens aus.
Somit weiß ich später, ob du genug Informationen über das Apartment hattest.
Komm danach wieder zu mir zurück und sag mir mit dem Wort FERTIG Bescheid, sodass wir weiter machen können."

"Hello, it's nice that you're taking part in this study.
My name is Flobi and I am your virtual contact person here in the apartment.
Today, I am going to tell you a little bit about this apartment.
After that I have a little task for you.
You are in the kitchen.
As soon as the construction work is done, it'll be fully functional again.
Underneath you (—) is a capacitive flooring.
It helps me to know where you are right now.
(**Disruption**) To understand each other well, for example, you see that **above you** (—) there is a microphone in the ceiling.
On the left in the kitchen (—), above the sink, is a window through which I can look out into the CITEC.
On the cabinets, some door handles have LEDs attached to

them.
These will turn blue if I want to show you something there.
(**Disruption**) The square next to me on the worktop is a digital scale.
So you always have one handy and you can never misplace it.
By the way, there is a camera in the cabinet above the stove.
This can measure the temperature, so that your milk never spills again.
To your right (—) is the living room.
The table that is standing there is an interactive Samsung Surface.
On it, you can take a look at a map of this apartment.
In addition, we have mounted a pan-tilt headlight onto the ceiling in the living room.
This can highlight certain areas in the living room.
(**Disruption**) Do you see my real head on the window-sill?
You can be glad that your one is attached.
Behind you (—), across the hallway, is an intelligent mirror.
There you can look at yourself from behind.
Shortly, you can explore everything further.
But for now, please go to the table and fill out the first part of the questionnaire.
So that I'll know later if you have received enough information about the apartment.
Then come back to me and tell me the word DONE so we can continue."

Example C.4.2: Sweet part:

"Vielen Dank für das Ausfüllen.
Ich möchte dir jetzt von der Aufgabe erzählen.
Du sollst mir helfen, ein paar Dinge wieder zu finden.
Vor etwa einer Stunde ist hier folgendes passiert:
Einige Handwerker waren hier im Apartment und haben die Küche umgebaut.
Durch den vielen Staub haben meine Sensoren nicht richtig funktioniert.
Währenddessen haben andere Leute Sachen hier im Apartment versteckt.
Ich vermute, es handelt sich dabei um die Süßwaren, die ich vorher auf dem Tisch gesehen habe.
Ich konnte wegen des Staubs leider nicht genau erkennen, wo die Sachen versteckt wurden,
aber ich werde dir alles erzählen, was ich weiß.
Dann kannst du versuchen, soviel wiederzufinden wie möglich und darfst sie am Ende auch behalten.

Pass jetzt gut auf. Ich sage dir was ich gesehen habe.
Versuch, dir alles zu merken!
Wenn ich dir alles erklärt habe, kannst du dich auf die Suche begeben.
Jedoch bist du ab dann auf dich allein gestellt.
Sprich bitte trotzdem jedes mal, bevor du weiter machst, laut aus wo du dich hinbegeben wirst.
***Links in der Küche** (—) hat jemand die Waschmaschine bedient und das Waschpulwerfach geöffnet.*
Da würde ich später auf jeden Fall mal nachsehen!
*(**Disruption**) Danach hat jemand die Beschteckschublade geöffnet und hat dort rumgewühlt.*
Vielleicht ist da etwas versteckt!
Und dann habe ich beobachtet dass jemand den Schrank über der Mikrowelle aufgemacht hat.
Schau doch da mal rein!
Es war noch jemand am Beschteckfach der Spülmaschine.
Ich weiß nicht, ob es mit der Sache zu tun hat, aber schau gleich mal nach.
*(**Disruption**) Das nächste was ich mitbekommen habe war, dass jemand den Dampfgarer **hinter dir** (—) benutzt hat.*
Vielleicht ist in dem Wasserbehälter etwas?
*Und ich habe gesehen, wie jemand **rechts im Wohnzimmer** (—) zur Pflanze gegangen ist, und etwas am Blumentopf gemacht hat*
Da solltest du auch mal nachsehen!
Dann wurde einer der Stühle im Wohnzimmer bewegt.
Die solltest du auch mal untersuchen.
Irgend etwas ist mit den Kaffeetassen auf dem Tisch im Wohnzimmer passiert.
Da könnte auch etwas versteckt sein.
*(**Disruption**) Es war wohl jemandem kalt auf den roten Sesseln.*
Ob sich etwas unter der Decke befindet?
Jemand hatte sich es auf dem Sofa bequem gemacht.
Nachgucken solltest du da auf jeden Fall.
Jetzt kommt dein Part, ich zähle auf dich.
Schau in beliebiger Reihenfolge an den Orten nach, die ich dir genannt habe.
Bevor du an einem Ort nachsiehst, sag mir bitte einmal den Namen des Ortes.
Alles Süße, was du findest, darfst du behalten.
Wenn du alles gefunden hast, fülle bitte den zweiten Teil des Fragebogens am Laptop im Wohnzimmer aus.
Dann wünsche ich dir viel Erfolg und Spaß bei der Suche!"

" Thank you for filling it in.
Now I want to tell you about the task.
I want you to help me to find some things again.

About an hour ago the following happened here:
Some craftsmen were here in the apartment remodeling the kitchen.
Due to all the dust, my sensors did not work properly.
Meanwhile, other people have hidden things here in the apartment.
I think it's the candy I saw on the table earlier.
Unfortunately, because of the dust, I couldn't see exactly where the things were hidden,
but I will tell you everything I know.
Then you can try to find as many as you can and keep them in the end.
Pay attention now. I'll tell you what I saw.
Try to remember everything!
When I've explained everything to you, you can start the search.
However, you are on your own from then on.
Please, each time before you continue, say out loud where you are going to go.
On the left in the kitchen (—), someone was running the washing machine and opening the washing powder compartment.
I would definitely check it later!
(**Disruption**) Then somebody opened the cutlery drawer and rummaged through it.
Maybe there is something hidden in there!
And then I noticed that someone opened the cupboard above the microwave.
Take a look inside!
There was someone else at the cutlery drawer of the dishwasher.
I do not know if it has anything to do with it, but check it out.
(**Disruption**) The next thing I noticed was that the steam cooker **behind you** (—) was used by someone.
Maybe there is something in the water tank?
And I saw someone go to the plant in **the living room on the right** (—) and do something to the flower pot.
You should have a look over there, too!
Then one of the chairs in the living room was moved.
You should also examine them.
Something happened to the coffee cups on the table in the living room.
Something could be hidden there, too.
(**Disruption**) Somebody must have been cold on those red armchairs.
Maybe something is under the blanket?
Someone had made himself comfortable on the sofa.
You should definitely check it out.
Now it's your turn, I count on you.

In no particular order, take a look at the places I've mentioned to you.
Before you look at a place, please tell me the name of the place.
Any candy you find you may keep.
Once you've found everything, please fill out the second part of the questionnaire on the laptop in the living room
I wish you much success and fun with the search!"

D INSTRUCTIONS FOR THE PARTICIPANTS

Liebe Teilnehmerin, lieber Teilnehmer, Herzlich Willkommen zur dieser Interaktionsstudie. Vielen Dank, dass Du dich bereit erklärt hast, an dieser Studie teilzunehmen. Du wirst gleich in das intelligente Apartment geführt, das mit einigen Sensoren, Kameras und Mikrofonen ausgestattet ist, die dich während der Interaktion aufzeichnen. Deine Daten werden anonymisiert und vertraulich behandelt (siehe Datenschutzerklärung). Wenn du das Apartment betreten hast, begebe dich in die Küche. Dort wirst du von Flobi begrüßt und führst mit ihm eine kurze Interaktion durch. Flobi wird dich durch diese Interaktion leiten. Ist sie zu Ende, gehe bitte in das Wohnzimmer und fülle dort am Computer einen kurzen Fragebogen aus. Neben der Tastatur befindet sich ein Zettel auf dem eine Nummer steht. Trage diese im Fragebogen als deine Probandennummer ein. Falls du noch Fragen hast, kannst du diese jetzt der Versuchsleitung stellen. Du hast jeder Zeit die Möglichkeit die Studie abzubrechen und das Apartment verlassen. Außerdem kannst du dich jederzeit an die Versuchsleitung wenden, die die ganze Zeit vor der Tür wartet.

E

DATA AGREEMENT OF THE CSRA

Einwilligungserklärung für Video- und Tonaufnahmen
Experimente und Demonstrationen im Rahmen des CSRA-Projekts (Cognitive Service Robotics Apartment)

Universität Bielefeld / CITEC

Ich (Name des Teilnehmers /der Teilnehmerin in Blockschrift)
____
bin mündlich und schriftlich von Herrn/Frau ______________________ darüber informiert worden, dass im Rahmen der folgenden Studie bzw. dieser Demonstration Video- und Tonaufnahmen von mir gemacht sowie meine Bewegungsmuster aufgezeichnet werden.
Die Aufnahmen dienen dazu, die Funktionsweisen dieses intelligenten Apartments zu testen, im Rahmen dieses Projektes wissenschaftlich auszuwerten und weiterzuentwickeln sowie darüber hinaus die Verhaltensweisen von StudienteilnehmerInnen und BesucherInnen in dieser Umgebung zu evaluieren.
Ich bin darüber informiert, dass die Aufzeichnung und Auswertung aller Daten zu keinem Zeitpunkt zusammen mit meinem Namen gespeichert werden. Diese Einwilligungserklärung wird in einem verschlossenen Schrank und getrennt von den aufgezeichneten Video- und Tondaten gelagert. Die aufgezeichneten Daten werden nur von Personen ausgewertet, die auf das Datengeheimnis verpflichtet wurden und die keine vertraulichen und personenbezogenen Informationen an Dritte weitergeben. Sie sind ausschließlich MitarbeiterInnen des Forschungsprojekts zugänglich und werden bis 10 Jahre nach Beendigung des Forschungsprojekts aufbewahrt. Das Projekt endet im Dezember 2018. Nähere Informationen zum Datenschutzprozedere können bei Bedarf ausgehändigt werden.
Mir ist bekannt, dass ich meine freiwillige Einwilligung zur Erhebung und Verarbeitung dieser Daten auch ohne Angabe von Gründen jederzeit widerrufen kann, ohne dass mir daraus Nachteile entstehen. Im Falle eines Widerrufs werden meine personenbezogenen Daten gelöscht. Hierzu wende ich mich mit einer formlosen E-Mail an die Versuchsleitung (Kontaktdaten siehe unten).

Ich hatte genügend Zeit für eine Entscheidung. Ich habe die vorliegende Information gelesen und verstanden und erkläre mich mit der oben beschriebenen Erhebung und Verarbeitung meiner Daten einverstanden.

Ich habe eine Ausfertigung dieser Einwilligungserklärung erhalten.

Bitte wenden!

Bitte machen Sie einige Angaben zur Verwendung Ihrer Daten:

Sind Sie damit einverstanden, dass die Aufzeichnungen, die von Ihnen innerhalb dieser Studie / Demonstration gemacht wurden, als Fallbeispiele in wissenschaftlichen Vorträgen / Konferenzen vorgestellt werden?

□ Ja, ich bin damit einverstanden.

□ Ja, ich bin damit einverstanden, aber nur, wenn die Daten vollständig anonymisiert sind, d.h. sofern die Video- oder Tonaufnahmen so entfremdet sind, dass kein Rückschluss auf meine Person möglich ist.

□ Nein, ich bin nicht damit einverstanden.

Dürfen wir Sie für eventuelle Folgeprojekte im Rahmen dieser Forschungsreihe kontaktieren?

□ Ja, Sie können mir unter folgender Mailadresse schreiben: ____________________

□ Nein, ich möchte nicht kontaktiert werden.

Bitte füllen Sie alle Felder dieser Einwilligungserklärung aus:

____
Name, Vorname (in Druckschrift)

________________________ ________________________
Ort, Datum Uhrzeit der Teilnahme

____
Unterschrift des Teilnehmers / der Teilnehmerin

Bei Fragen oder anderen Anliegen kann ich mich an die Projektleitung wenden:
KoordinatorInnen: Britta Wrede, Thomas Hermann, Sven Wachsmuth
E-Mail: csra@cit-ec.uni-bielefeld.de

F

ADDITIONAL ANALYSIS EC5

TASK EFFICIENCY Next, I will evaluate the task efficiency, the difference of task performance and the pretest:

$$\mathsf{EFF}_{total} = \underbrace{(\mathsf{TP}_{info} - \mathsf{TP}_{pretest})}_{\mathsf{EFF}_{info}} + \underbrace{(\mathsf{TP}_{sweet} - \mathsf{TP}_{pretest})}_{\mathsf{EFF}_{sweet}}$$

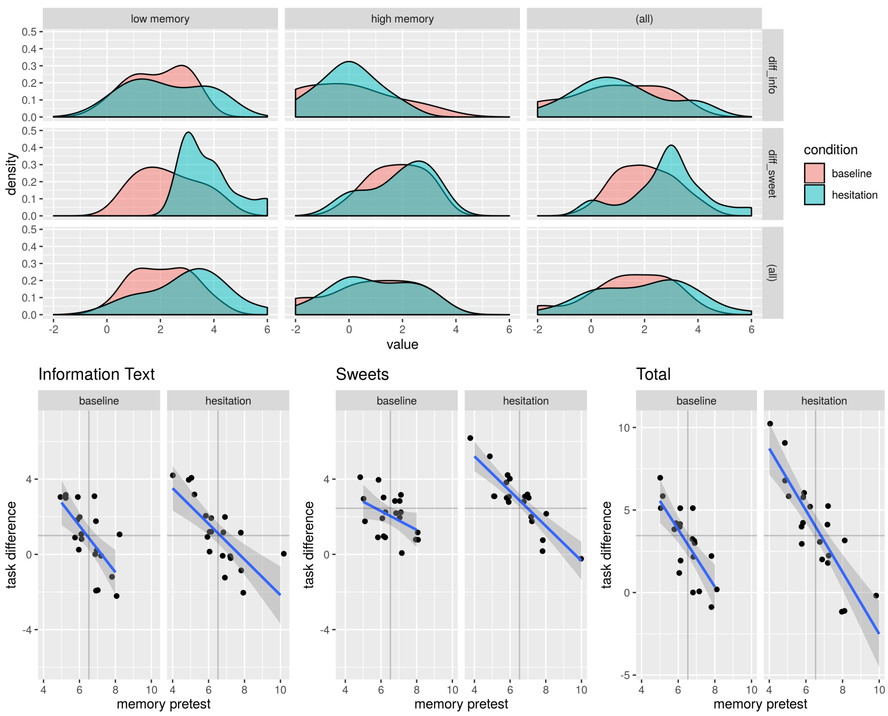


Figure F.1: Task efficiency during the information phase, the finding rate of of sweets and the performance in total for the baseline and hesitation condition as density plot (upper) and over the memory performance in the pretest (lower).

Figure F.1 visualizes the density of efficiency for each subtask. In addition, the difference is plottet over the performance in the pretest. It can be seen, that a linear dependency from the pretest exists. This make sense—in terms of the possible difference at all. A participant who reach 10 points in the pretest cannot reach a higher score in one of task, whereas a participant with four points can reach improve their points by a maximum of six. To analyses the influence of the hesitation condition on this difference, I carry out an ANCOVA with the performance in the pretest as covariate and the condition as independent variable. As expected, I found a statistically significant main effect of participants pretest performance, $F(1, 36) = 76.057, p < .001$.

Furthermore, the condition had a significant effect on the total task performance, $F(1, 36) = 4.38, p = .043$.

G ADDITIONAL STUDY RESULTS

G.1 RESULTS OF EC1

G.1.1 *Subjective Ratings*

key concept	condition	N	Mean	SD
anthropomorphism	AHM	13	1.98	0.65
anthropomorphism	BASE	14	1.98	0.77
animacy	AHM	13	2.69	0.78
animacy	BASE	14	2.62	0.60
likeability	AHM	13	3.81	0.57
likeability	BASE	14	4.02	0.55
intelligence	AHM	13	3.19	0.48
intelligence	BASE	14	3.47	0.46
safety	AHM	13	3.35	0.52
safety	BASE	14	3.37	0.75

Table G.1: Mean and SD values for the subjective ratings of *EC1*.

key concept	t(df)	p	CI	Sig.
anthropomorphism	t(24.74)=-0.01	.996	[-0.57, 0.57]	
animacy	t(22.54)= 0.79	.789	[-0.48, 0.62]	
likeability	t(24.68)=-0.96	.346	[-0.65, 0.24]	
intelligence	t(24.67)=-1.56	.131	[-0.66, 0.09]	
safety	t(23.23)=-0.52	.950	[-0.52, 0.49]	

Table G.2: Independent *T-Test* for comparing the mean values of each key concept between the *AHM* and *BASE* condition.

G.1.2 *Visual Attention*

variable	condition	N	Mean	Median	SD
NA_{number}	AHM	13	1.31	1.0	0.85
NA_{number}	BASE	14	1.93	2.0	0.83
NA_{total}	AHM	13	2.14	1.5	1.50
NA_{total}	BASE	14	4.48	4	3.56
TP	AHM	13	2.92	3	0.86
TP	BASE	14	3.36	3	0.84

Table G.3: Mean median, and SD values for the number of look away (NA_{number}), and the total time of being inattentive (NA_{total}), and the task performance (TP).

G.2 RESULTS OF EC2

G.2.1 *Subjective Ratings*

key concept	condition	N	Mean	Median	SD
anthropomorphism	AHM	15	1.90	2.14	0.65
anthropomorphism	BASE	15	2.22	2.29	0.82
animacy	AHM	15	2.47	2.86	0.69
animacy	BASE	15	2.86	3.10	0.74
likeability	AHM	15	3.44	3.57	0.87
likeability	BASE	15	4.29	4.29	0.58
intelligence	AHM	15	3.27	3.29	0.59
intelligence	BASE	15	3.82	3.86	0.57
safety	AHM	15	3.17	3.33	0.86
safety	BASE	15	3.48	3.81	0.62

Table G.4: Mean and SD values for the subjective ratings of EC2.

key concept	t(df) or *W*	p	CI	sig.
anthropomorphism	W=89.5	.348		
animacy	W=64.5	.048		*
likeability	t(24.46)=-3.14	.004	[-1.40, -0.29]	**
intelligence	t(27.96)=-2.59	.015	[-0.99, -0.11]	*
safety	W=84	.239		

Table G.5: Independent *T-Test* or Wilcoxon rank sum tets.

G.2.2 *Visual Attention*

variable	condition	N	Mean	Median	SD
NA_{number}	AHM	15	5.13	5	0.85
NA_{number}	BASE	15	6.33	6	2.99
NA_{total}	AHM	15	8.56	8	3.36
NA_{total}	BASE	15	10.19	9	6.09
$TP_{non-embodied}$	AHM	15	0.78		0.16
$TP_{non-embodied}$	BASE	15	0.78		0.16
$TP_{embodied}$	AHM	15	0.93		0.15
$TP_{embodied}$	BASE	15	0.83		0.24
TP_{All}	AHM	15	8.40	8	1.12
TP_{All}	BASE	15	8.00	8	1.51

Table G.6: Mean median, and SD values for the number of look away (NA_{number}), and the total time of being inattentive (NA_{total}), and the task performance (TP).

G.3 RESULTS OF EC3

G.3.1 *Subjective Ratings*

key concept	condition	N	Mean	SD
anthropomorphism	AHM	15	2.08	0.71
anthropomorphism	BASE	13	2.40	0.48
animacy	AHM	15	3.04	0.51
animacy	BASE	13	3.05	0.55
likeability	AHM	15	3.86	0.68
likeability	BASE	13	4.12	0.59
intelligence	AHM	15	3.66	0.58
intelligence	BASE	13	3.71	0.56
safety	AHM	15	3.43	0.62
safety	BASE	13	3.53	0.39

Table G.7: Mean and standard deviation for the *GQS* of *EC3*.

key concept	t(df) or *W*	p	CI	Sig.
anthropomorphism	t(24.74)=-3.14	.167	[-0.79, 0.15]	
animacy	t(24.75)=-0.05	.962	[-0.43, 0.40]	
likeability	t(25.99)=-1.09	.283	[-0.76, 0.29]	
intelligence	t(25.89)=-0.27	.793	[-0.50, 0.38]	
safety	t(23.76)=-0.51	.588	[-0.51, 0.29]	

Table G.8: Independent *T-Test* or Wilcoxon rank sum tets.

Subjective ratings	Condition	N	Mean	Median	SD
inappropriately timed-timely	AHM	15	3.80	4	1.15
inappropriately timed-timely	BASE	13	4.69	5	1.03
delayed-suitable	AHM	15	4.73	5	1.22
delayed-suitable	BASE	13	4.62	5	1.04
suitable-too fast	AHM	15	4.60	4	1.68
suitable-too fast	BASE	13	4.54	4	1.33
too slow-suitable	AHM	15	4.80	5	1.37
too slow-suitable	BASE	13	4.77	5	1.24
incomprehensible-easy to understand	AHM	15	5.27	6	1.44
incomprehensible-easy to understand	BASE	13	5.77	6	1.24
suitable-too	AHM	15	5.33	5	0.72
suitable-too	BASE	13	2.92	3	1.61
to short-suitable	AHM	15	5.80	6	1.26
to short-suitable	BASE	13	5.77	6	0.83

Table G.9: Ratings regarding the appropriateness of the agent's statements.

G.4 RESULTS OF EC4

G.4.1 *Task Performance*

variable	condition	N	Mean	Median	SD
$TP_{pretest}$	AHM	20	6.70	6	1.45
$TP_{pretest}$	BASE	20	6.85	6.5	1.39
TP_{sweets}	AHM	20	6.30	7	0.86
TP_{sweets}	BASE	20	5.50	6	1.28

Table G.10: Mean median, and SD values for the task performance in the sweets and info task.

G.4.2 *Visual Attention*

variable	condition	N	Mean	Median	SD
NA_{number} AHM	20	20.00	17	13.44	
NA_{number} BASE	20	15.15	14.5	5.01	
NA_{total} AHM	20	58.48	55	41.76	
NA_{total} BASE	20	39.89	40	16.37	

Table G.11: Mean median, and SD values for the number of look away (NA_{number}) and the total time of being inattentive (NA_{total}).

G.4.3 *Subjective Ratings*

condition	key concepts	N	Mean	SD
AHM	anthropomorphism	20	1.76	0.68
AHM	animacy	20	2.10	0.64
AHM	likeability	20	3.69	0.58
AHM	intelligence	20	3.38	0.54
AHM	safety	20	3.46	0.67
BASE	anthropomorphism	20	1.90	0.56
BASE	animacy	20	2.41	0.60
BASE	likeability	20	3.81	0.64
BASE	intelligence	20	3.44	0.65
BASE	safety	20	3.44	0.66
AHM	voice quality	20	2.30	0.80
BASE	voice quality	20	3.55	0.69

Table G.12: Mean and SD values for the subjective ratings.

key concept	t(df) or *W*	p	CI	Sig.
anthropomorphism	$t(36.57) = -0.69$	.497	[-0.54, 0.27]	
animacy	t(37.77)=-1.61	.116	[-0.71, 0.08]	
likeability	t(37.57)=-0.66	.511	[-0.52, 0.26]	
intelligence	t(36.86)=-0.34	.736	[-0.44, 0.32]	
safety	t(37.98)=0.11	.910	[-0.40, 0.45]	
voice quality	W=53	$> .001$	***	

Table G.13: Independent *T-Test* or *WR-Test*.

G.5 RESULTS OF EC5

G.5.1 *Subjective Ratings*

condition	scenario	value	N	Mean	SD
BASE	info	anthropomorphism	20	2.20	0.61
BASE	info	animacy	20	2.71	0.57
BASE	info	likeability	20	3.81	0.69
BASE	info	intelligence	20	3.70	0.49
BASE	info	safety	20	2.88	0.51
BASE	sweet	anthropomorphism	20	2.69	0.76
BASE	sweet	animacy	20	3.01	0.73
BASE	sweet	likeability	20	4.03	0.70
BASE	sweet	intelligence	20	3.79	0.42
BASE	sweet	safety	20	2.85	0.43
AHM	info	anthropomorphism	20	2.15	0.69
AHM	info	animacy	20	2.95	0.67
AHM	info	likeability	20	3.71	0.85
AHM	info	intelligence	20	3.80	0.56
AHM	info	safety	20	2.77	0.31
AHM	sweet	anthropomorphism	20	2.27	0.67
AHM	sweet	animacy	20	2.92	0.66
AHM	sweet	likeability	20	3.80	0.73
AHM	sweet	intelligence	20	3.69	0.63
AHM	sweet	safety	20	2.90	0.24
BASE	info	voice quality	20	3.35	0.75
BASE	sweet	voice quality	20	3.35	0.81
AHM	info	voice quality	20	2.90	0.97
AHM	sweet	voice quality	20	3.35	0.88

Table G.14: Mean and SD values for the subjective ratings.

key concept	$t(df)$ or W	p	CI	Sig.
$anthropomorphism_{info}$	t(38)=0.25	.809	[-0.37, 0.47]	
$animacy_{info}$	t(38)=-1.23	.228	[-0.64, 0.15]	
$likeability_{info}$	t(38)=0.41	.686	[-0.40, 0.60]	
$intelligence_{info}$	t(38)=-0.60	.553	[-0.44, 0.28]	
$safty_{info}$	W=210	.785		
$anthropomorphism_{sweets}$	t(38)=1.85	.072	[-0.03, 0.88]	
$animacy_{sweets}$	t(38)=0.42	.686	[-0.35, 0.53]	
$likeability_{sweets}$	t(38)=1.01	.317	[-0.23, 0.69]	
$intelligence_{sweets}$	t(38)=0.59	.558	[-0.24, 0.44]	
$safty_{sweets}$	W=202	.965		

Table G.15: Independent Two Sample t-tests or respectively *WR-Test* for Godspeed values of *AHM* and *BASE* condition

Key value	$t(df)$ or V	p	CI	Sig.
$anthropomorphism_{base}$	t(19)=-4.31	$>$ .001	[-0.73, -0.25]	$***$
$animacy_{base}$	t(19)=-2.61	.017	[-0.54, -0.05]	$*$
$likeability_{base}$	t(19)=-2.41	.024	[-0.41, -0.03]	$*$
$intelligence_{base}$	t(19)=-1.28	.216	[-0.24, 0.06]	
$safty_{base}$	V=84.5	.716		
$anthropomorphism_{hes}$	t(19)=-1.5	.15	[-0.29, 0.05]	
$animacy_{hes}$	t(19)=0.42	.678	[-0.13, 0.20]	
$likeability_{hes}$	t(19)=-0.94	.359	[-0.29, 0.11]	
$intelligence_{hes}$	t(19)=1.24	.231	[-0.08, 0.30]	
$safty_{hes}$	V=20	.142		

Table G.16: Dependent Two Sample t-tests or respectively *WSR-Test* with continuity correction for Godspeed values of *information* and *sweets* part.

variable	condition	N	Mean	Median	SD
NA_{number}	AHM	20	27.05	27	10.17
NA_{number}	BASE	20	29.26	28	9.59
NA_{total}	AHM	20	58.97	50	37.93
NA_{total}	BASE	20	61.95	56	29.47
TP_{info}	AHM	20	7.65	7.5	1.14
TP_{info}	BASE	20	7.40	7	1.31
TP_{sweets}	AHM	20	9.40	10	0.75
TP_{sweets}	BASE	20	8.55	9	1.10
TP_{All}	AHM	20	17.00	16.5	1.52
TP_{All}	BASE	20	15.95	16	0.47

Table G.17: Mean median, and SD values for the number of look away (NA_{number}), and the total time of being inattentive (NA_{total}), and the task performance (TP).

DECLARATION OF AUTHORSHIP

According to the Bielefeld University's doctoral degree regulations §8(1)g: I hereby declare to acknowledge the current doctoral degree regulations of the Faculty of Technology at Bielefeld University. Furthermore, I certify that this thesis has been composed by me and is based on my own work, unless stated otherwise. Third parties have neither directly nor indirectly received any monetary advantages in relation to mediation advises or activities regarding the content of this thesis. Also, no other person's work has been used without due acknowledgment. All references and verbatim extracts have been quoted, and all sources of information, including graphs and data sets, have been specifically acknowledged. This thesis or parts of it have neither been submitted for any other degree at this university nor elsewhere.

Birte Richter

Place, Date